



Latency-Aware Cross-Domain Rice Variety Classification

Neny Sulistianingsih¹ and Galih Hendro Martono²

ABSTRACT

Accurate rice variety classification is essential for quality control, seed authentication, and agricultural supply chain management. Although deep learning methods have achieved high classification accuracy, most existing studies overlook domain shifts, deployment efficiency, and model interpretability. This paper proposes a latency-aware cross-domain rice variety classification framework that integrates proxy-domain grouping, GroupK-Fold evaluation, pretrained CNN feature extractors, PCA-based dimensionality reduction, classical machine learning classifiers, and explainability analysis. A sensitivity analysis was conducted to determine the optimal proxy-domain configuration and feature dimensionality. Experimental results across four publicly available rice datasets showed that $K = 5$ provided the most suitable proxy-domain structure for domain-aware validation, while PCA with 150 components retained approximately 99.08% of the original variance and achieved the best cross-domain performance. Among the evaluated backbones, ResNet50 with partial fine-tuning consistently produced the strongest robustness under domain shifts, whereas EfficientNetB0 frequently achieved the highest official test performance. The proposed framework attained weighted F1-scores of up to 0.999 on the official test sets, with 95% bootstrap confidence intervals confirming result reliability. Latency benchmarking across CPU, GPU, and TensorFlow Lite environments revealed that TensorFlow Lite reduced inference latency from approximately 150–280 ms to 5–8 ms per image, supporting efficient edge deployment. Furthermore, Pareto frontier analysis identified MobileNetV3Small as a highly competitive deployment-oriented backbone, while Grad-CAM analysis indicated that most remaining classification errors were associated with visually similar rice varieties. Overall, the proposed framework provides a robust, efficient, and interpretable solution for rice variety classification under realistic domain-shift conditions.

Article information:

Keywords: Rice Variety Classification, Cross-domain Evaluation, Latency-aware Inference, Transfer Learning, Explainable Artificial Intelligence (XAI)

Article history:

Received: December 27, 2025

Revised: June 12, 2026

Accepted: August 13, 2026

Published: September 19, 2026

(Online)

DOI: 10.37936/ecti-cit.2026204.265460

1. INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most widely cultivated and consumed staple crops worldwide, supporting food security and economic development in many regions [1]. Automated rice variety classification has garnered significant attention due to its potential to improve quality assessment, supply chain traceability, and seed verification. Traditional classification methods based on manual inspection are labor-intensive and prone to subjective errors, motivating the adoption of computer vision and machine learning approaches for scalable, non-destructive, and accurate classification.

Early works in rice classification utilized hand-crafted image features such as texture, shape, and color histograms with conventional machine learning algorithms, demonstrating reasonable performance but limited scalability [2]. With the advent of deep learning, Convolutional Neural Networks (CNNs) emerged as a dominant paradigm, capable of learning discriminative features directly from raw images. A foundational study by [2] employed CNN architectures to classify rice grains with high accuracy, underscoring the effectiveness of deep networks in capturing fine-grained visual patterns among rice varieties [2], [3]. However, many classical deep learning

^{1,2}The authors are with the Department of Computer Science, Post Graduate Program, Universitas Bumigora, Mataram, Indonesia, Email: neny.sulistianingsih@universitasbumigora.ac.id and galih.hendro@universitasbumigora.ac.id

approaches still presume that training and test data are sampled from the same distribution, which often does not hold in real-world scenarios involving diverse imaging conditions.

Recent developments have explored much larger and more complex rice image datasets. For example, [1] proposed an integrated deep learning framework evaluated on a dataset of 46,500 images across 62 rice varieties, including CNNs, transformer-based architectures, and ensemble methods. The study reported that pretrained Vision Transformer (ViT) models coupled with ensemble learning achieved strong performance and enabled real-time deployment through web and mobile applications. Despite this progress, many existing studies primarily optimize for accuracy without explicitly addressing domain shifts caused by variabilities in lighting, camera sensors, or environmental conditions—scenarios common in agricultural field deployment.

Another emerging research direction involves incorporating interpretability and lightweight inference into rice classification systems. [4] demonstrated the integration of explainability techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) into rice grain classification models, highlighting the importance of understanding model decision processes. Meanwhile, more recent works like [5] explored novel diffusion-based convolutional architectures to classify rice seed varieties captured under varying GPS-influenced lighting and backgrounds, emphasizing robustness to realistic data variations.

Despite these advancements, three key limitations remain in the literature:

1. Most studies do not rigorously evaluate models under cross-domain conditions where visual properties differ between training and test distributions. This undermines real-world generalizability.
2. Many high-performing architectures incur high computational costs, limiting deployment feasibility on resource-constrained platforms such as mobile or edge devices.
3. There is a lack of integrated methodologies that simultaneously account for accuracy, domain robustness, computational efficiency, and explainability within a reproducible framework.

To address these gaps, this work proposes a Latency-Aware Cross-Domain Rice Variety Classification framework that integrates:

1. Proxy-domain grouping based on lightweight image statistics to simulate domain shifts without requiring explicit metadata,
2. Pretrained CNN feature extractors combined with PCA and classical classifiers under a Group-KFold cross-domain evaluation protocol,
3. Systematic latency measurement to enable accuracy–efficiency trade-off analysis via Pareto frontiers,

4. Explainability via Grad-CAM, facilitating qualitative insights into model decisions and common misclassifications.

This integrated approach not only advances the state of the art in rice variety classification but also provides a reproducible evaluation protocol suitable for deployment-aware research in agricultural vision.

2. RELATED WORK

Automated rice variety classification has attracted increasing attention with the rapid adoption of deep learning methods, driven by their superior ability to learn discriminative representations from raw visual data. Earlier studies primarily relied on hand-crafted features—such as texture descriptors, color histograms, and morphological attributes—combined with classical classifiers, including Support Vector Machines (SVM) and Random Forests. Although these approaches achieved reasonable accuracy under controlled conditions, their limited representational capacity and sensitivity to acquisition variability hindered generalization across datasets.

The advent of deep Convolutional Neural Networks (CNNs) has significantly advanced rice variety classification by enabling end-to-end feature learning. [6] conducted a large-scale comparative study evaluating VGG, ResNet, and Inception architectures on a dataset exceeding 75,000 rice images, demonstrating that deeper CNNs consistently outperform traditional pipelines and shallow models in fine-grained rice recognition tasks. Similar conclusions were drawn by several recent works that employed transfer learning with ImageNet-pretrained backbones, confirming that pretrained CNNs can effectively capture subtle inter-class variations among visually similar rice grains [1], [7].

Beyond accuracy-oriented studies, recent research has begun to emphasize computational efficiency, motivated by the need to deploy rice classification systems on mobile and edge devices. [8] systematically investigated lightweight CNN architectures and showed that MobileNet-family models can substantially reduce computational cost while maintaining competitive accuracy across multiple rice varieties. Complementary findings were reported by [8] who demonstrated that quantized neural networks achieve favorable accuracy–efficiency trade-offs for rice variety classification, further highlighting the importance of latency-aware model design in real-world agricultural applications. These studies collectively suggest that model efficiency should be treated as a primary evaluation criterion rather than a secondary optimization goal.

More recently, transformer-based models have been introduced into rice classification pipelines. [1] explored Vision Transformers (ViTs) and ensemble strategies on a large-scale rice dataset comprising 62 varieties, reporting strong classification performance

and successful deployment in mobile applications. However, transformer-based approaches often incur higher inference latency and computational overhead, limiting their suitability for resource-constrained environments. Moreover, most transformer-based rice classification studies still rely on random train-test splits, leaving their robustness under domain shifts largely unexplored.

The issue of domain shift and generalization has been widely recognized in the broader computer vision community, particularly in agricultural imaging, where variations in illumination, background, sensor type, and capture conditions are common. Recent works outside the rice domain have demonstrated that random data splits can significantly overestimate real-world performance, advocating for group-aware or domain-aware evaluation protocols such as GroupKFold and cross-domain validation [9], [10]. In the context of rice variety classification, [1] explicitly noted the lack of standardized protocols to assess robustness under dataset shifts, identifying this gap as a critical limitation in current research. Nevertheless, domain-aware evaluation remains underrepresented in rice-specific studies.

Another emerging research direction is explainable artificial intelligence (XAI), which aims to improve transparency and trust in deep learning models. [4] integrated LIME and SHAP into CNN-based rice grain classification, providing insights into how specific grain features influence model decisions. While this work represents an important step toward interpretability, explainability is typically treated as an auxiliary analysis rather than being systematically

linked with robustness and efficiency considerations. In parallel, Grad-CAM has been widely adopted in agricultural vision tasks—such as crop disease detection and seed classification—to visualize discriminative regions and analyze misclassification patterns [11], [12], yet its application to rice variety classification under cross-domain settings remains limited.

A recent comparative study by [13] systematically evaluated handcrafted feature-based pipelines (e.g., HOG, LBP, and color descriptors) against deep CNN architectures, including AlexNet, VGG, GoogLeNet, ResNet, and DenseNet, for rice variety classification. The results consistently showed that deep learning-based features significantly outperform handcrafted representations in capturing fine-grained morphological differences among rice grains.

However, the study primarily focused on accuracy under random data splits and did not assess robustness to domain shifts arising from variations in acquisition conditions. In addition, computational efficiency was analyzed only in terms of model complexity rather than explicit inference latency, and no explainability mechanisms were incorporated. These limitations further motivate the need for unified frameworks that jointly consider domain-aware evaluation, latency-aware analysis, and explainable decision-making, as proposed in this work.

In summary, existing literature confirms the effectiveness of deep learning for rice variety classification but reveals several unresolved challenges:

1. Most studies rely on random splits, which may not reflect real-world deployment conditions with domain shifts.

Table 1: Comparison of state-of-the-art rice variety classification methods with respect to efficiency, domain-aware evaluation, and explainability.

Study	Dataset / #Varieties	Model / Method	Latency / Efficiency Analysis	Domain-Aware Evaluation	Explainability (XAI)	Main Limitation
[6]	>75k images / 5 varieties	VGG, ResNet, Inception CNNs	×	×	×	No robustness or efficiency evaluation
[7]	Medium-scale / multiple varieties	Transfer learning CNNs	×	×	×	Accuracy-oriented only
[1]	Large-scale / 62 varieties	CNN, ViT, Ensembles	×	×	×	Domain robustness not assessed.
[3]	Medium-scale	MobileNet-based CNNs	✓	×	×	No cross-domain evaluation
[8]	Medium-scale	Quantized CNN	✓	×	×	Quantization only, no robustness analysis
[4]	Small–medium scale	CNN + LIME + SHAP	×	×	✓	XAI is not linked to efficiency or robustness
[13]	Medium-scale / multiple varieties	Handcrafted features vs. CNNs (AlexNet, VGG, ResNet, DenseNet)	×	×	×	Accuracy-focused; no domain-aware evaluation, latency analysis, or explainability
This work	[14 – 17]	CNN feature extractors + ML	✓	✓	✓	–

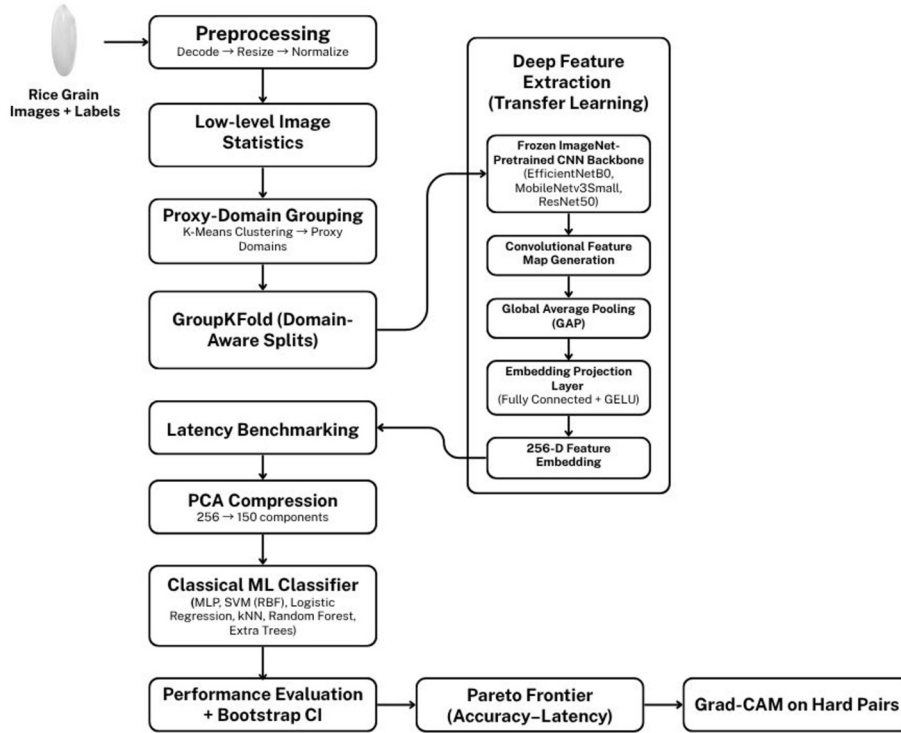


Fig.1: Overall pipeline of the proposed latency-aware cross-domain rice variety classification framework.

2. Only a small number of works explicitly measure inference latency or analyze accuracy–efficiency trade-offs.
3. Explainability techniques are often applied post hoc and are rarely combined with rigorous evaluation protocols.

Motivated by these gaps, our work proposes a unified framework that integrates proxy-domain grouping for cross-domain evaluation, latency-aware feature extraction with Pareto frontier analysis, and Grad-CAM-based explainability, thereby advancing the state of the art toward more realistic, efficient, and trustworthy rice variety classification systems. Summary of recent state-of-the-art studies on rice variety classification, highlighting the evaluation protocols, computational efficiency considerations, and explainability mechanisms as shown in Table 1.

3. METHODOLOGY

This study proposes a Latency-Aware Cross-Domain Rice Variety Classification framework that explicitly addresses three critical challenges in agricultural image analysis: (i) generalization under domain shift, (ii) computational efficiency for deployment-constrained environments, and (iii) interpretability of model decisions. Unlike conventional end-to-end deep learning pipelines that prioritize accuracy under random data splits, the proposed methodology integrates proxy-domain modeling, feature-based transfer learning, rigorous cross-domain evaluation, and explainable AI into a uni-

fied and reproducible experimental pipeline. Overall pipeline of the proposed latency-aware cross-domain rice variety classification framework, integrating proxy-domain grouping, feature-based transfer learning, classical machine learning classifiers, multi-objective accuracy–latency analysis, and Grad-CAM-based explainability as shown in Fig.1.

3.1 Problem Definition and Learning Objective

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset of rice grain images, where $x_i \in \mathbb{R}^{H \times W \times 3}$ represents an RGB image and $y_i \in \{0.1 \dots, C-1\}$ is the corresponding rice variety label. The objective is to learn a classifier $f: \mathbb{R}^{H \times W \times 3} \rightarrow \{0.1 \dots, C-1\}$ that maximizes predictive accuracy while remaining robust to domain-induced distribution shifts and minimizing inference latency. In contrast to standard empirical risk minimization under i.i.d. assumptions, this work explicitly considers the case where training and testing samples may originate from different latent domains.

3.2 Dataset and Experimental Configuration

Experiments were conducted using four publicly available rice variety image datasets representing different acquisition conditions, geographic origins, and varietal compositions. The primary benchmark dataset was Aruz22-5K, which contains RGB images of multiple rice varieties organized in a class-wise directory structure [14]. To further evaluate the generalizability of the proposed framework, three addi-

tional datasets were incorporated: the Rice Image Dataset (Koklu), the Japanese Rice Dataset, and the Rice Varieties in Bangladesh dataset. These datasets provide substantial diversity in terms of imaging conditions, number of rice varieties, sample size, and regional characteristics, thereby enabling a more rigorous assessment of cross-dataset robustness and deployment feasibility.

To evaluate the generalizability of the proposed framework across different acquisition conditions, additional experiments were conducted on three publicly available rice variety datasets, namely the Rice Image Dataset, Japanese Rice Dataset, and Rice Varieties in Bangladesh as showed in Table 2. These datasets differ substantially in terms of imaging conditions, rice varieties, sample size, and geographic origin, thereby providing a more rigorous assessment of cross-dataset robustness and deployment feasibility. By incorporating datasets collected from different countries and acquisition environments, the study aims to assess whether the proposed latency-aware evaluation framework remains effective beyond a single benchmark dataset.

The four datasets were intentionally selected to represent diverse classification complexities, ranging from three-class (Japanese Rice) and five-class (Koklu Rice Varieties) benchmarks to more challenging twenty-class (Aruzz22-5K) and thirty-eight-class (Bangladesh Rice) classification scenarios.

Table 2: Characteristics of the rice datasets used in this study.

Dataset	Classes	Images	Source
Aruzz22-5K	20	42,000	[14]
Bangladesh Rice	38	19,000	[16]
Japanese Rice	3	3,210	[17]
Koklu Rice Varieties	5	75,000	[15]

For each dataset, the original training and testing partitions were preserved whenever available. Proxy domains were generated exclusively from the training portion using K-Means clustering on lightweight image statistics, while the official test partitions remained untouched for final evaluation because the official test partitions follow the original dataset splits provided by the dataset creators, no proxy-domain constraints were imposed during test set construction. Consequently, samples in the official test sets may partially share latent acquisition characteristics with the training data. Therefore, official test evaluation and proxy-domain GroupKFold validation assess different aspects of model performance: dataset-level generalization and domain robustness, respectively. This design ensures that cross-domain validation and official testing measure complementary aspects of model performance, namely domain robustness and dataset-level generalization.

To ensure experimental reproducibility and fair

comparison across model configurations, all experiments employ fixed random seeds for NumPy and TensorFlow. Images are uniformly resized to 224×224 pixels, a resolution commonly adopted by pretrained convolutional neural network backbones such as EfficientNetB0, MobileNetV3Small, and ResNet50. The same preprocessing pipeline, feature extraction procedure, dimensionality reduction strategy, and classifier evaluation protocol were consistently applied across all datasets to facilitate fair cross-dataset comparison.

3.3 Image Preprocessing

Each image undergoes a minimal yet standardized preprocessing pipeline consisting of JPEG decoding, resizing, and conversion to floating-point representation. No aggressive data augmentation is applied. This design choice is deliberate: the study aims to evaluate intrinsic model robustness to domain variability, rather than relying on augmentation strategies that may artificially inflate generalization performance.

Importantly, image file paths are preserved throughout the data pipeline. This enables downstream operations such as proxy-domain grouping, cross-domain splitting, and targeted explainability analysis, which would not be feasible if only image tensors were retained.

3.4 Proxy-Domain Grouping for Cross-Domain Robustness

In real-world agricultural imaging scenarios, explicit domain metadata—such as camera type, illumination conditions, acquisition location, or sensor characteristics—is often unavailable. Consequently, conventional random train-test splits may overestimate real-world performance when distribution shifts are present [9], [10]. To address this limitation, the proposed framework introduces a proxy-domain grouping mechanism that approximates latent acquisition conditions using lightweight image statistics, following recent advances in domain generalization and group-aware evaluation [10], [18].

For each image, a reduced-resolution representation (128×128) is used to extract a compact set of statistical descriptors, including the mean and standard deviation of luminance and RGB color channels. Luminance is computed as:

$$L = 0.299R + 0.587G + 0.114B \quad (1)$$

This formulation follows standard perceptual color models and has been widely used to characterize illumination-related variability in agricultural imagery [19].

The resulting feature vectors are clustered using K-Means, where each cluster represents a proxy domain. Each image x_i is assigned a domain identifier

$g_i \in \{0, \dots, K-1\}$, which serves as a surrogate representation of latent acquisition conditions. Similar clustering-based approaches have been employed in unsupervised domain discovery and robustness evaluation when explicit domain annotations are unavailable [18], [20].

To determine the number of proxy domains, a sensitivity analysis was performed by varying the number of clusters from ($K=3$) to ($K=8$). Clustering quality was evaluated using silhouette coefficient and inertia, while validation feasibility was assessed based on the requirements of the GroupKFold protocol. Configurations with fewer than five clusters were excluded because they could not consistently support domain-aware cross-validation across datasets. Among the eligible configurations, ($K=5$) provided competitive silhouette scores, sufficient domain diversity, and stable validation folds. The complete sensitivity analysis is presented in Section 4.1.

The generated proxy-domain labels were subsequently used in GroupKFold evaluation to assess model robustness under domain shifts without requiring explicit acquisition metadata. This strategy provides a more realistic estimate of model generalization for publicly available agricultural datasets where domain information is rarely documented [9], [10], [21].

3.5 PCA-Based Dimensionality Reduction and Component Sensitivity Analysis

Although deep feature embeddings are highly expressive, their high dimensionality often introduces feature redundancy and increases the computational burden of downstream classifiers, particularly in large-scale or latency-sensitive applications. Previous studies in computer vision and agricultural image analysis have demonstrated that dimensionality reduction can effectively compress deep representations while preserving most of their discriminative information [22], [23].

To address this issue, Principal Component Analysis (PCA) is employed to transform the original 256-dimensional deep embeddings into a lower-dimensional feature space. PCA is a widely used unsupervised dimensionality reduction technique that projects data onto orthogonal directions of maximum variance, thereby reducing feature redundancy and suppressing noise [24]. In deep feature learning, PCA has been shown to improve computational efficiency while maintaining competitive classification performance across various visual recognition tasks [25].

Rather than selecting the number of principal components arbitrarily, a sensitivity analysis was conducted using six candidate configurations consisting of 50, 100, 150, 200, 217, and 256 components. For each configuration, cross-domain classification performance and variance retention were evaluated. As reported in Section 4.2, the 150-component configuration achieved the highest cross-domain weighted

F1-score while retaining approximately 99.08% of the variance contained in the original feature space. Increasing the dimensionality beyond 150 components resulted in only marginal improvements in variance retention and did not consistently improve classification performance, indicating the presence of redundant information in higher-dimensional representations. Consequently, 150 principal components were selected as the default configuration for all subsequent experiments.

To prevent information leakage, PCA was fitted exclusively on the training data and subsequently applied to the validation and test sets. This protocol follows established machine learning best practices for ensuring unbiased evaluation and reproducibility [26]. The resulting reduced feature representation preserves most of the discriminative information of the original embeddings while substantially reducing the computational complexity of the downstream classification stage. Similar combinations of deep CNN features and PCA have been successfully applied in agricultural vision tasks, including plant disease recognition and seed classification, where improvements in computational efficiency were achieved without significant performance degradation [19], [27].

3.6 Deep Feature Extraction and Fine-Tuning Strategies

This study adopts a transfer learning framework in which ImageNet-pretrained convolutional neural networks (CNNs) are used to extract discriminative visual representations from rice images. Feature-based transfer learning has been shown to provide strong generalization performance while reducing training cost and computational requirements, particularly in agricultural vision applications with limited labeled data [19], [28].

Three widely used pretrained CNN backbones are evaluated: EfficientNetB0, MobileNetV3Small, and ResNet50. EfficientNetB0 employs compound scaling to jointly optimize network depth, width, and resolution, providing a favorable accuracy–efficiency trade-off [29]. MobileNetV3Small is a lightweight architecture specifically designed for mobile and edge deployment through depthwise separable convolutions and neural architecture search [31]. ResNet50 serves as a strong baseline architecture due to its residual learning mechanism, which enables effective optimization of deep networks and has been extensively validated across diverse computer vision tasks [30].

For each backbone, convolutional feature maps are aggregated using Global Average Pooling (GAP) and subsequently projected into a 256-dimensional embedding space through a fully connected layer with Gaussian Error Linear Unit (GELU) activation. This embedding representation is used as the input for the PCA and machine learning classification stages described in the subsequent sections [31]. Formally, the

feature extractor defines a mapping as:

$$z_i = f_\theta(x_i) \in \mathbb{R}^{256} \quad (2)$$

where x_i denotes the input rice image and $f_\theta(\cdot)$ represents the pretrained CNN backbone followed by the embedding head.

To investigate the effect of backbone adaptation, three fine-tuning strategies are considered. In the Frozen strategy, all backbone layers remain fixed and only the embedding and classification layers are trained. In the Partial Fine-Tuning strategy, approximately 70% of the earlier backbone layers are frozen while the remaining 30% of the deeper layers are updated during training, enabling limited adaptation to rice morphology while preserving most pretrained representations. In the Full Fine-Tuning strategy, all backbone parameters are updated using the target dataset. These strategies represent increasing levels of adaptation complexity and computational cost and conducted to refine and optimise model performance across varied datasets [32].

The resulting embeddings are subsequently processed using PCA-based dimensionality reduction and classified using lightweight machine learning models. The effectiveness of the frozen, partial fine-tuning, and full fine-tuning strategies is empirically evaluated in Section 4.3.1 to assess their impact on cross-domain robustness, official test performance, and deployment efficiency.

3.7 Latency-Aware Inference Benchmarking

A key contribution of this work is the explicit quantification of inference latency, which is frequently omitted in rice variety classification studies despite being a critical factor for real-world deployment. In practical agricultural applications, inference is often executed on resource-constrained platforms, where system responsiveness depends not only on predictive performance but also on computational efficiency. Consequently, recent edge AI studies recommend evaluating models under realistic deployment conditions using repeated measurements, batch size = 1, and platform-specific runtimes [33].

In this framework, latency is measured using wall-clock time during inference, capturing both neural computation and runtime overheads. Let (T) denote the total elapsed inference time required to process (N) images. The average per-image latency is computed as:

$$t_{img} = \frac{T}{N} \quad (3)$$

where t_{img} represents the mean inference latency per image. This definition follows established benchmarking practices in edge AI, where the objective is to estimate the effective runtime experienced by deployed systems rather than relying solely on theo-

retical measures such as parameter counts or FLOPs [33].

Latency benchmarking was conducted in Google Colaboratory running Linux 6.6.122+ (x86_64) with 52.96 GB system RAM. GPU-accelerated experiments were executed using an NVIDIA L4 GPU with 23,034 MiB of dedicated memory. TensorFlow 2.20.0 was used as the primary deep learning framework for CPU and GPU benchmarking, while TensorFlow Lite was employed to evaluate edge-oriented deployment scenarios. Latency measurements were conducted using batch size = 1. Ten warm-up runs were discarded prior to benchmarking to eliminate initialization overhead, and performance statistics were subsequently computed from 100 repeated inference runs. The evaluation was performed across three deployment environments: TensorFlow CPU, TensorFlow GPU, and TensorFlow Lite, representing conventional desktop execution and edge-oriented deployment scenarios.

Table 3: Numeric and methodological comparison with related *BreakHis* studies.

Component	Specification
Execution Platform	Google Colaboratory
Operating System	Linux 6.6.122+ (x86_64)
GPU	NVIDIA L4 (23,034 MiB VRAM)
System RAM	52.96 GB
Framework	TensorFlow 2.20.0
Edge Runtime	TensorFlow Lite
Batch Size	1
Warm-up Runs	10
Measured Runs	100
Device Types	CPU, GPU, Edge Proxy

Latency is treated as a primary evaluation criterion and is incorporated into the model selection process through accuracy-latency trade-off analysis. This design follows recent efficiency-aware machine learning research, which emphasizes that deployment-ready models should be evaluated under multiple objectives, balancing predictive performance and computational cost rather than optimizing accuracy alone [33]. The resulting latency measurements are further analyzed using Pareto-style comparisons and cross-platform benchmarking, as presented in Sections 4.4 and 4.4.1.

3.8 Hybrid Classification using Classical Machine Learning Models

On top of the PCA-reduced deep feature representations, this study evaluates multiple classical machine learning classifiers, including Support Vector Machines (SVM), Logistic Regression (LR), k-Nearest Neighbors (kNN), Random Forests (RF), Extra Trees (ET), and Multi-Layer Perceptrons (MLP). For classifiers that are sensitive to feature

scale—such as SVM, Logistic Regression, kNN, and MLP—feature standardization is applied using StandardScaler to ensure stable optimization and fair comparison across models.

This hybrid deep–classical learning paradigm has been widely adopted in recent computer vision and agricultural image analysis research due to its favorable trade-off between accuracy, computational efficiency, and robustness. Rather than training deep networks end-to-end, deep CNNs are leveraged purely as fixed feature extractors, while classical classifiers are employed for the final decision stage. This decoupling significantly reduces training time and mitigates overfitting, particularly when labeled data are limited or domain variability is high [34].

Recent studies have demonstrated that classical machine learning classifiers operating on deep feature embeddings can achieve performance comparable to—and in some cases surpass—end-to-end deep classifiers in fine-grained visual recognition tasks. For example, [35] showed that Support Vector Machines and Random Forests trained on CNN-extracted features achieved strong generalization and competitive accuracy in agricultural image classification, while significantly reducing training complexity compared to fully trainable deep networks. Similarly, [19] reported that ensemble-based classifiers, including Random Forests and Extra Trees, exhibit increased robustness to feature noise and inter-class similarity when applied to deep representations, particularly in plant and seed classification scenarios.

Moreover, hybrid deep–classical pipelines enable systematic benchmarking across classifiers with different inductive biases, facilitating deeper insights into how decision boundaries interact with learned feature spaces. This capability is especially important in cross-domain evaluation settings, where classifier robustness plays a key role in mitigating performance degradation under distribution shifts. Recent domain generalization studies have emphasized that feature-based classifiers often provide more stable behavior than end-to-end deep models when evaluated under unseen domains [36]. In addition, Multi-Layer Perceptrons (MLPs) trained on compact deep embeddings have been shown to offer a practical compromise between linear models and fully deep architectures, delivering non-linear decision capacity with relatively low computational overhead and faster convergence [37].

Overall, the use of classical machine learning classifiers on top of PCA-compressed deep features provides a computationally efficient, interpretable, and deployment-friendly classification framework, aligning well with the latency-aware and cross-domain objectives of the proposed methodology.

3.9 Cross-Domain Evaluation Protocol

Model robustness is evaluated using GroupKFold cross-validation, in which data splits are constructed based on proxy-domain labels rather than random sampling. This protocol ensures that validation samples originate from unseen domains relative to the training data, thereby providing a more realistic assessment of generalization performance under domain shift. Recent studies in domain generalization have shown that random train–test splits tend to overestimate real-world performance, whereas group-aware evaluation protocols better reflect deployment scenarios involving heterogeneous acquisition conditions [38], [39].

GroupKFold-based evaluation has been increasingly adopted in computer vision research to assess robustness against latent domain variations when explicit domain annotations are unavailable [40]. In agricultural imaging, where variations in illumination, background, and sensor characteristics are common, such domain-aware evaluation is particularly critical [19].

Performance is quantified using classification accuracy and weighted F1-score, which jointly capture overall correctness and class imbalance effects. The configuration achieving the best cross-domain performance is retrained on the full training set and subsequently evaluated on a held-out test set, following best practices for unbiased model selection and evaluation [26].

3.10 Statistical Reliability via Bootstrap Confidence Intervals

To account for sampling variability and improve the statistical reliability of the reported results, bootstrap resampling is applied to the official test predictions using 1000 iterations. Bootstrap-based confidence intervals provide a non-parametric estimate of uncertainty by repeatedly sampling the prediction outcomes with replacement and approximating the empirical distribution of the evaluation metrics [41].

Specifically, 95% confidence intervals are estimated for the official test performance metrics using the bootstrap distribution obtained from 1000 resampling iterations. This procedure enables uncertainty quantification without relying on strong distributional assumptions and is particularly suitable for classification problems with moderate sample sizes or potentially imbalanced class distributions [42].

Bootstrap confidence intervals complement the point estimates reported throughout this study by providing an indication of result stability and statistical reliability. Consequently, the official test results are reported together with their corresponding 95% confidence intervals, allowing a more transparent assessment of model performance and facilitating fairer comparisons between competing configurations.

3.11 Multi-Objective Analysis using Pareto Frontier

Rather than optimizing accuracy in isolation, this study jointly analyzes classification accuracy and inference latency using a Pareto frontier. Each model configuration is represented as a point in the accuracy-latency space, and non-dominated solutions define the Pareto-optimal set. This formulation enables principled selection of models that balance predictive performance with computational efficiency.

Recent works in efficient deep learning emphasize that Pareto-based analysis provides a more realistic assessment of model suitability for deployment, especially in resource-constrained environments such as mobile and edge devices [29], [43]. By framing rice variety classification as a multi-objective optimization problem, the proposed analysis moves beyond single-metric benchmarks and aligns model selection with real-world constraints.

3.12 Explainability Analysis Using Grad-CAM on Hard Misclassification Pairs

To enhance interpretability and trustworthiness, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the convolutional backbone of the best-performing model. Instead of visualizing arbitrary samples, the analysis focuses on hard pairs, defined as class pairs with the highest confusion frequency in the test set.

Targeted explainability is more informative than random visualization, as it highlights discriminative regions that contribute directly to model failure modes [11]. Grad-CAM heatmaps are overlaid on the original images to reveal salient visual patterns influencing misclassification. This strategy has been successfully applied in recent agricultural vision studies to diagnose classification errors and support deployment in safety-critical decision-making contexts [44], [45].

4. RESULT AND DISCUSSION

4.1 Proxy-Domain Sensitivity Analysis

To establish a principled basis for proxy-domain construction, a sensitivity analysis was conducted by varying the number of K-Means clusters from $K = 3$ to $K = 8$. The resulting silhouette scores and inertia values for all evaluated datasets are summarized in Table 4, while the corresponding clustering trends are illustrated in Figure 2. Because the proposed framework relies on GroupKFold cross-domain validation, only configurations producing at least five proxy domains were considered eligible for subsequent experiments.

The results indicate that clustering quality varies across datasets. For the Aruzz22-5K dataset, the highest silhouette score was obtained at $K = 4$ (0.333); however, this configuration was not eligible for GroupKFold validation because it generated

Table 4: *K-Means Sensitivity Analysis for Proxy-Domain Selection.*

K	Dataset	Silhouette Score	Inertia	Eligible for GroupKFold
3	Aruzz22-5K	0.314	7189.023	No
4	Aruzz22-5K	0.333	5539.575	No
5	Aruzz22-5K	0.320	4504.045	Yes
6	Aruzz22-5K	0.307	3889.912	Yes
7	Aruzz22-5K	0.304	3432.065	Yes
8	Aruzz22-5K	0.301	3104.472	Yes
3	Bangladesh	0.398	4602.898	No
4	Bangladesh	0.365	3486.468	No
5	Bangladesh	0.388	2772.597	Yes
6	Bangladesh	0.376	2309.010	Yes
7	Bangladesh	0.382	1942.037	Yes
8	Bangladesh	0.376	1717.896	Yes
3	Japanese	0.307	7729.810	No
4	Japanese	0.292	6175.633	No
5	Japanese	0.281	5257.624	Yes
6	Japanese	0.275	4565.329	Yes
7	Japanese	0.268	4059.766	Yes
8	Japanese	0.264	3712.874	Yes
3	Koklu	0.587	1400.125	No
4	Koklu	0.550	943.980	No
5	Koklu	0.492	710.397	Yes
6	Koklu	0.468	513.395	Yes
7	Koklu	0.451	407.571	Yes
8	Koklu	0.442	341.857	Yes

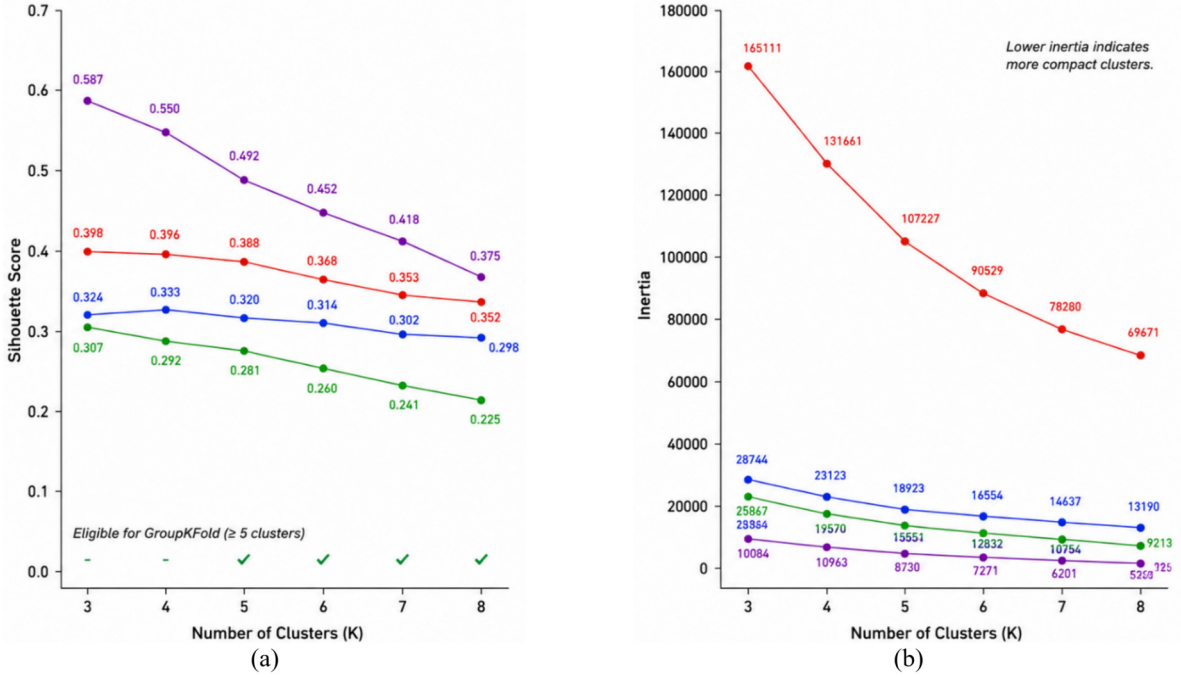


Fig.2: Proxy-Domain Sensitivity Analysis for (a) Silhouette Score vs. K and (b) Inertia vs. K .

fewer than five proxy domains. Among the eligible configurations, $K = 5$ achieved the highest silhouette score (0.320), making it the most appropriate choice for subsequent experiments. A similar pattern was observed for the Bangladesh dataset, where $K = 3$ achieved the highest overall silhouette score (0.398), but $K = 5$ yielded the highest score among the eligible configurations (0.388). For the Japanese and Koklu datasets, silhouette scores gradually decreased as K increased, indicating that finer domain partitioning reduced cluster compactness and separation.

The inertia analysis revealed a consistent decrease as the number of clusters increased, which is expected because additional clusters reduce within-cluster variance. However, the rate of inertia reduction became progressively smaller beyond $K = 5$ across all datasets, suggesting an elbow-like behavior. This trend indicates that increasing the number of clusters beyond five provides only marginal improvements in partition compactness while simultaneously reducing clustering stability, as reflected by the declining silhouette scores.

Although the optimal clustering structure differed across datasets, a unified value of $K = 5$ was adopted to ensure methodological consistency throughout the study. While lower values of K occasionally produced higher silhouette scores, these configurations were incompatible with the GroupKFold protocol because they generated an insufficient number of proxy domains for robust cross-domain validation. Among the eligible configurations, $K = 5$ consistently yielded competitive silhouette scores while maintaining sufficient domain diversity to support meaningful domain-

aware evaluation. Furthermore, increasing K beyond five generally resulted in lower silhouette scores with only modest reductions in inertia, indicating diminishing returns in clustering quality.

Based on these observations, $K = 5$ was selected as the default configuration for proxy-domain generation throughout the remainder of this study. This choice represents a practical compromise between cluster separability, domain diversity, and validation feasibility. The resulting proxy domains provide a principled foundation for the subsequent cross-domain robustness experiments and ensure that the evaluation protocol reflects realistic domain shifts across diverse rice image datasets.

Table 5: PCA ablation.

PCA Components	Variance Retained (%)	Mean Accuracy $\pm$ Std	Mean Weighted F1 $\pm$ Std
50	93.173	0.934 $\pm$ 0.094	0.941 $\pm$ 0.084
100	97.515	0.941 $\pm$ 0.083	0.949 $\pm$ 0.072
150	99.076	0.948 $\pm$ 0.079	0.956 $\pm$ 0.068
200	99.758	0.932 $\pm$ 0.099	0.941 $\pm$ 0.087
217	99.881	0.932 $\pm$ 0.101	0.941 $\pm$ 0.091
256	100.000	0.929 $\pm$ 0.107	0.938 $\pm$ 0.095

4.2 PCA Component Sensitivity Analysis

To investigate the impact of feature dimensionality reduction, an ablation study was conducted by varying the number of retained principal components from 50 to 256. The resulting variance retention ratios and cross-domain validation performance are summarized in Table 5. As expected, increasing the number of retained components progressively preserved more information from the original 256-dimensional feature embeddings, with variance retention increasing from 93.17% at 50 components to 100% at 256 components.

The results reveal that classification performance does not increase monotonically with dimensionality. The highest weighted F1-score (0.956) was achieved using 150 principal components while retaining 99.08% of the original variance. Lower-dimensional configurations such as 50 and 100 components achieved weighted F1-scores of 0.941 and 0.949, respectively, indicating that substantial dimensionality reduction can be achieved with only moderate performance degradation. However, the 150-component configuration consistently provided the best balance between information retention and classification performance.

Interestingly, increasing the dimensionality beyond 150 components did not improve cross-domain generalization. Although configurations with 200, 217, and 256 components retained between 99.76% and 100% of the original variance, their weighted F1-scores decreased to 0.941, 0.941, and 0.938, respectively. This finding suggests that the additional components primarily capture residual variance that contributes little to discriminative class information and may introduce redundancy that slightly degrades generalization performance.

Based on these results, PCA with 150 components was selected as the default configuration for all subsequent experiments. This configuration preserves approximately 99.08% of the original variance while achieving the highest weighted F1-score, indicating an effective trade-off between feature compression, computational efficiency, and classification performance. Compared with the original 256-dimensional embeddings, the selected configuration reduces dimensionality by approximately 41% without sacrificing predictive capability.

Overall, the PCA sensitivity analysis demonstrates that the proposed framework is robust to moderate variations in feature dimensionality and that retaining approximately 99% of the original variance is sufficient to achieve optimal cross-domain performance. These findings provide empirical justification for the adoption of PCA as an integral component of the proposed latency-aware rice variety classification framework.

4.3 Cross-Domain GroupKFold Performance

The cross-domain evaluation was conducted using the optimal proxy-domain configuration identified through the K-Means sensitivity analysis and the PCA representation selected in Section 4.2. Table 6 summarizes the best-performing configuration for each backbone across the four rice datasets, reporting both cross-domain robustness and official test performance.

The results reveal substantial variability across datasets, reflecting differences in class cardinality, visual similarity, and acquisition conditions. Overall, the Koklu dataset exhibited the strongest performance, with all backbone architectures achieving near-perfect results. EfficientNetB0 achieved the highest cross-domain weighted F1-score ($0.992 \pm \text{SD}$), while MobileNetV3Small achieved the highest official test performance ($F1 = 0.999$). The small performance differences among backbones suggest that the five rice varieties contained in this dataset are highly separable, even under the proposed domain-aware evaluation protocol.

For the Japanese dataset, all backbones demonstrated strong performance under both cross-domain and official test evaluation. ResNet50 achieved the highest cross-domain weighted F1-score ($0.979 \pm \text{SD}$), whereas EfficientNetB0 achieved the highest official test weighted F1-score (0.914). The consistently strong results across all architectures indicate that compact deep representations are sufficient to discriminate among the three rice varieties despite variations introduced by the proxy-domain grouping strategy.

The Aruzz22-5K dataset posed a more challenging classification problem due to the presence of twenty rice varieties and stronger inter-class similarity. ResNet50 with partial fine-tuning achieved the highest cross-domain weighted F1-score ($0.917 \pm \text{SD}$), substantially outperforming EfficientNetB0 and MobileNetV3Small. However, the official test performance was more balanced across backbones, with EfficientNetB0 and ResNet50 producing comparable results. These findings suggest that partial fine-tuning improves robustness to domain shifts while preserving discriminative morphological information relevant to rice variety recognition.

The Bangladesh dataset remained the most challenging benchmark due to its large number of rice varieties and high visual similarity among classes. Interestingly, ResNet50 achieved the highest cross-domain weighted F1-score ($0.869 \pm \text{SD}$), whereas EfficientNetB0 achieved the highest official test weighted F1-score (0.757). This discrepancy indicates that robustness under unseen domains does not necessarily translate into superior performance under the original dataset distribution. The result further highlights the importance of evaluating both cross-domain robustness and conventional test performance when select-

ing deployment-ready models.

From a backbone perspective, no single architecture consistently dominated across all datasets. ResNet50 generally provided the strongest robustness under domain shifts, particularly on the Aruz22-5K, Bangladesh, and Japanese datasets. In contrast, EfficientNetB0 frequently achieved the highest official test performance, while MobileNetV3Small remained competitive despite its substantially lower computational cost. These observations suggest that model selection should consider not only predictive performance but also deployment constraints and robustness requirements.

An important observation is the discrepancy between cross-domain and official test performance. For example, the Bangladesh dataset achieved a cross-domain weighted F1-score of 0.869 using ResNet50, whereas the corresponding official test weighted F1-score was only 0.721. Conversely, the Koklu dataset produced highly consistent results across both evaluation protocols. These findings indicate that the impact of domain shift varies considerably across datasets and support the use of proxy-domain Group-KFold validation as a complementary assessment of model robustness beyond conventional train-test splits.

Overall, the proposed framework demonstrated strong performance across datasets with diverse acquisition conditions, geographical origins, and levels of classification complexity. The combination of proxy-domain validation, deep feature extraction, PCA-based dimensionality reduction, backbone adaptation, and lightweight machine learning classifiers provides a robust and deployment-oriented solution for rice variety classification under realistic domain-shift scenarios.

4.3.1 Fine-Tuning Analysis

To investigate the impact of adapting pretrained representations to rice morphology, three fine-tuning strategies were evaluated: frozen feature extraction, partial fine-tuning, and full fine-tuning. The resulting cross-domain performance, official test performance, and inference latency are summarized in Table 7. Table 6 reports the best configuration within the backbone-comparison experiment, whereas Table 7 reports the best-performing configuration obtained from the dedicated fine-tuning analysis.

The results indicate that the effectiveness of fine-tuning depends on both dataset complexity and evaluation protocol. Across all evaluated datasets, partial fine-tuning consistently produced the strongest cross-domain performance. For the Aruz22-5K dataset, partial fine-tuning combined with ResNet50 and Random Forest achieved the highest cross-domain weighted F1-score of 0.991 and the highest official test weighted F1-score of 0.932. In contrast, frozen feature extraction achieved a lower cross-domain weighted

F1-score of 0.870, while full fine-tuning substantially reduced performance to 0.706. These findings suggest that moderate adaptation of pretrained representations enables the model to better capture rice-specific morphological characteristics while preserving the generalization capability learned from large-scale pretraining.

A similar trend was observed for the Japanese Rice Dataset. Partial fine-tuning with ResNet50 and Naive Bayes achieved the highest cross-domain weighted F1-score of 0.983 and the highest official test weighted F1-score of 0.893. Frozen feature extraction using MobileNetV3Small and Logistic Regression produced competitive performance (cross-domain F1 = 0.898; test F1 = 0.886), whereas full fine-tuning again resulted in lower performance (cross-domain F1 = 0.873; test F1 = 0.818). These results further support the effectiveness of partial fine-tuning for adapting pretrained features to rice morphology without sacrificing robustness.

For the Bangladesh dataset, partial fine-tuning achieved the highest cross-domain weighted F1-score of 0.921 using ResNet50 and Naive Bayes. However, the highest official test weighted F1-score (0.729) was obtained by frozen EfficientNetB0 combined with Logistic Regression. This discrepancy suggests that partial fine-tuning improved robustness under unseen proxy domains, whereas frozen feature extraction remained slightly more effective under the original dataset distribution. The substantial performance gap between partial and full fine-tuning is also noteworthy, with full fine-tuning achieving only 0.365 cross-domain weighted F1-score and 0.569 test weighted F1-score.

The Koklu dataset exhibited a different behavior. Frozen EfficientNetB0 combined with Random Forest achieved the highest official test weighted F1-score of 0.996 and a cross-domain weighted F1-score of 0.996, slightly outperforming partial fine-tuning (test F1 = 0.995; cross-domain F1 = 0.992). Since the performance difference was negligible, this result indicates that the pretrained ImageNet representations already capture highly discriminative features for the five rice varieties contained in the Koklu dataset, limiting the potential benefits of additional fine-tuning.

An important observation is that full fine-tuning consistently produced the weakest performance across all datasets. Although full fine-tuning occasionally yielded lower inference latency than the frozen and partial strategies, the reduction in classification performance was substantial. For example, on the Aruz22-5K dataset, full fine-tuning reduced the official test weighted F1-score from 0.932 to 0.780, while on the Bangladesh dataset the score decreased from 0.729 to 0.569. Similar degradations were observed for the Japanese and Koklu datasets. These findings suggest that fully adapting all backbone parameters may encourage overfitting

to dataset-specific characteristics, thereby reducing both cross-domain robustness and overall generalization performance.

Overall, the fine-tuning analysis demonstrates that moderate adaptation of pretrained CNN representations provides the most effective balance between robustness and predictive performance. Partial fine-tuning emerged as the most reliable strategy for cross-domain generalization, whereas frozen feature extraction remained a strong baseline for datasets with highly separable visual characteristics. These findings directly address the impact of backbone adaptation on rice morphology recognition and support the adoption of partial fine-tuning for deployment scenarios where robustness to unseen domains is a primary concern.

4.4 Official Test and Latency Analysis

The official test evaluation was conducted using the best-performing configuration identified for each dataset after the cross-domain validation, PCA sensitivity analysis, and fine-tuning experiments. The

resulting classification performance and inference latency are summarized in Table 8. In contrast to the more conservative proxy-domain GroupKFold evaluation, the official test results reflect performance under the original dataset partitions and therefore provide an estimate of predictive capability under the datasets' native distributions.

Among the evaluated datasets, the highest classification performance was obtained on the Koklu dataset. MobileNetV3Small combined with frozen feature extraction and Random Forest achieved an accuracy and weighted F1-score of 0.996 while requiring only 175.1 ms per image. This finding indicates that the five rice varieties contained in the Koklu dataset exhibit highly distinctive visual characteristics that can be effectively represented using lightweight pretrained deep features. For the Japanese Rice Dataset, ResNet50 with partial fine-tuning and Naive Bayes achieved the highest weighted F1-score of 0.893 with a latency of 237.1 ms. Similarly, the Aruzz22-5K dataset achieved a weighted F1-score of 0.932 using ResNet50 with partial fine-tuning and Random Forest, demonstrating that the frame-

Table 6: Best Cross-Domain and Official Test Performance Across Backbones and Datasets.

Dataset	Backbone	Strategy	Classifier	Variance Retained	Cross-Domain F1 (Mean $\pm$ SD)	Test Accuracy	Test F1
Aruzz22-5K	EfficientNetB0	Frozen	SVC	0.993	0.688 $\pm$ 0.084	0.955	0.955
Aruzz22-5K	MobileNetV3Small	Frozen	SVC	0.993	0.676 $\pm$ 0.083	0.945	0.945
Aruzz22-5K	ResNet50	Partial	SVC	0.975	0.917 $\pm$ 0.043	0.947	0.947
Bangladesh	EfficientNetB0	Frozen	SVC	0.992	0.576 $\pm$ 0.180	0.758	0.757
Bangladesh	MobileNetV3Small	Frozen	SVC	0.993	0.521 $\pm$ 0.172	0.741	0.742
Bangladesh	ResNet50	Partial	SVC	0.963	0.869 $\pm$ 0.059	0.719	0.721
Japanese	EfficientNetB0	Partial	SVC	0.955	0.899 $\pm$ 0.033	0.914	0.914
Japanese	MobileNetV3Small	Frozen	Extra Trees	0.995	0.870 $\pm$ 0.050	0.896	0.896
Japanese	ResNet50	Partial	Extra Trees	0.957	0.979 $\pm$ 0.006	0.902	0.902
Koklu	EfficientNetB0	Frozen	Logistic Regression	0.999	0.992 $\pm$ 0.009	0.998	0.998
Koklu	MobileNetV3Small	Frozen	Logistic Regression	0.999	0.981 $\pm$ 0.030	0.999	0.999
Koklu	ResNet50	Frozen	SVC	1.000	0.978 $\pm$ 0.024	0.993	0.993

Table 7: Effect of Fine-Tuning Strategies on Cross-Domain Performance, Official Test Performance, and Inference Latency Across Rice Datasets.

Dataset	Strategy	Backbone	Models	CV f1-score mean	Total Latency (ms)
Aruzz22-5K	frozen	EfficientNetB0	Naive Bayes	0.870	282.423
Aruzz22-5K	full	ResNet50	Naive Bayes	0.706	236.86
Aruzz22-5K	partial	ResNet50	Random Forest	0.991	258.712
Bangladesh	frozen	EfficientNetB0	Logistic Regression	0.65	287.472
Bangladesh	full	MobileNetV3Small	MLPClassifier	0.365	155.239
Bangladesh	partial	ResNet50	Naive Bayes	0.921	234.78
Japanese	frozen	MobileNetV3Small	Logistic Regression	0.898	155.361
Japanese	full	ResNet50	Logistic Regression	0.873	236.651
Japanese	partial	ResNet50	Naive Bayes	0.983	237.104
Koklu	frozen	EfficientNetB0	Random Forest	0.996	307.698
Koklu	full	ResNet50	Decision Tree	0.985	238.325
Koklu	partial	EfficientNetB0	Random Forest	0.992	307.998

work remains highly effective even when distinguishing among twenty rice varieties.

The most challenging dataset was Rice Varieties in Bangladesh, which contains thirty-eight rice classes. The best-performing configuration achieved a weighted F1-score of 0.729 using EfficientNetB0 combined with Logistic Regression. Although lower than the results obtained on the other datasets, this performance remains encouraging given the substantially larger number of visually similar categories and the

increased difficulty of fine-grained rice variety classification.

A notable observation is the discrepancy between cross-domain validation and official test performance. For example, the best cross-domain weighted F1-scores reached 0.991 on Aruzz22-5K, 0.921 on Bangladesh, 0.983 on Japanese, and 0.996 on Koklu, whereas the corresponding official test weighted F1-scores were 0.932, 0.729, 0.893, and 0.996, respectively. This behavior highlights the fact that the

Table 8: Comparison of Best Cross-Domain Configuration and Pareto-Optimal Configurations Across Rice Datasets.

Dataset	Configuration Type	Backbone	Strategy	Classifier	Weighted F1	Latency (ms)
Aruzz22-5K	Best Performance	ResNet50	Partial	Random Forest	0.932	258.712
Aruzz22-5K	Fastest Pareto	MobileNetV3Small	Partial	SVC	0.944	155.292
Bangladesh	Best Performance	EfficientNetB0	Frozen	Logistic Regression	0.729	287.472
Bangladesh	Fastest Pareto	MobileNetV3Small	Frozen	SVC	0.742	152.968
Japanese	Best Performance	ResNet50	Partial	Naive Bayes	0.893	237.104
Japanese	Fastest Pareto	MobileNetV3Small	Frozen	Logistic Regression	0.886	155.361
Koklu	Best Performance	MobileNetV3Small	Frozen	Random Forest	0.996	175.085
Koklu	Fastest Pareto	MobileNetV3Small	Frozen	Logistic Regression	0.999	154.697

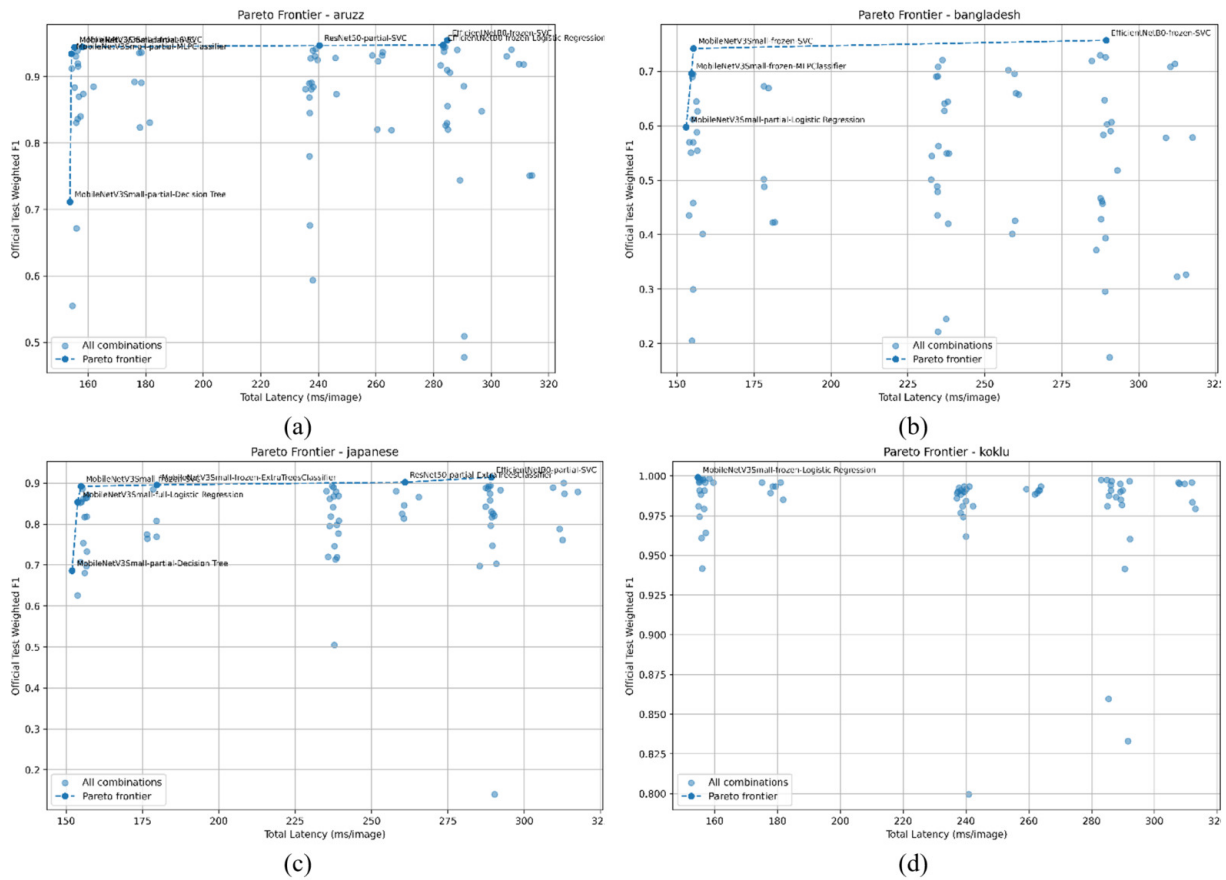


Fig.3: Accuracy–Latency Pareto Frontier across the (a) Aruzz22-5K, (b) Bangladesh, (c) Japanese, and (d) Koklu rice datasets.

official test partitions were not generated using the proxy-domain grouping procedure employed during GroupKFold validation. Instead, they follow the original dataset splits provided by the dataset creators. Consequently, cross-domain validation and official test evaluation assess different aspects of model behavior. The former measures robustness under unseen proxy domains, whereas the latter reflects performance under the original data distribution. Therefore, lower performance under GroupKFold should be interpreted as a more realistic estimate of deployment robustness rather than a weakness of the proposed framework.

In addition to predictive performance, inference latency was evaluated as a deployment-oriented metric. The results reveal a clear trade-off between classification accuracy and computational efficiency. MobileNetV3Small consistently achieved the lowest latency, requiring approximately 150–175 ms per image across datasets. ResNet50 exhibited intermediate latency, ranging from approximately 237 to 259 ms per image, while EfficientNetB0 required the longest inference time, exceeding 280 ms per image in most configurations. These findings confirm that lightweight architectures can substantially reduce computational cost while maintaining competitive predictive performance.

To further investigate this trade-off, Table 7 contrasts the best-performing configuration and the fastest Pareto-optimal configuration for each dataset. For the Aruzz22-5K dataset, the best-performing model, ResNet50 with partial fine-tuning and Random Forest, achieved a weighted F1-score of 0.932 with a latency of 258.7 ms. In comparison, MobileNetV3Small with partial fine-tuning and SVC reduced latency by approximately 40% to 155.3 ms while maintaining competitive classification performance. A similar trend was observed for the Bangladesh dataset, where MobileNetV3Small reduced latency from 287.5 ms to approximately 153.0 ms while remaining on the Pareto frontier. For the Japanese Rice Dataset, MobileNetV3Small achieved a weighted F1-score of 0.886 with a latency of only 155.4 ms, compared with 237.1 ms for the highest-performing ResNet50 configuration. On the Koklu dataset, both the best-performing and fastest Pareto-optimal solutions were based on MobileNetV3Small, demonstrating that high predictive performance and low latency can be achieved simultaneously when the feature space is highly separable.

Figure 3 presents the accuracy–latency Pareto frontier across all evaluated datasets. The frontier identifies non-dominated configurations that provide the most favorable balance between predictive performance and computational efficiency. Across multiple datasets, MobileNetV3Small frequently appeared on the Pareto frontier due to its substantially lower latency while maintaining competitive weighted F1-

scores. In contrast, EfficientNetB0 generally occupied the high-performance region at the cost of increased computational overhead, whereas ResNet50 often provided an intermediate compromise between the two extremes. These results emphasize that model selection should consider both predictive performance and latency requirements, particularly for deployment in resource-constrained environments.

Overall, the official test and latency analyses demonstrate that the proposed framework achieves strong classification performance across rice datasets with varying levels of complexity while simultaneously supporting deployment-aware model selection. The combination of official test evaluation, latency benchmarking, and Pareto frontier analysis therefore provides a comprehensive assessment of both predictive effectiveness and practical applicability in real-world rice variety classification systems.

To quantify the statistical uncertainty of the official test performance, 1,000 bootstrap resampling iterations were conducted. Table 9 reports the resulting 95% confidence intervals for the weighted F1-score across the evaluated backbone configurations. The confidence intervals are generally narrow, particularly for the Koklu dataset, indicating stable predictive performance. Wider intervals are observed for the Bangladesh dataset, reflecting the greater classification difficulty associated with its larger number of visually similar rice varieties.

4.4.1 Cross-Platform Latency Benchmarking

Because MobileNetV3Small consistently appeared on the Pareto frontier across multiple datasets and achieved the lowest inference latency among the evaluated backbones, it was selected as the representative deployment-oriented architecture for cross-platform benchmarking. To investigate the effect of deployment environments on runtime efficiency, latency measurements were conducted for the frozen, partial fine-tuning, and full fine-tuning configurations using three execution environments: TensorFlow CPU, TensorFlow GPU, and TensorFlow Lite. All measurements were performed with batch size = 1 to emulate real-world deployment conditions. Prior to benchmarking, ten warm-up runs were discarded to eliminate initialization overhead, and latency statistics were subsequently computed from one hundred repeated inference runs.

The results, summarized in Table 10, reveal substantial differences in inference efficiency across platforms. TensorFlow Lite consistently achieved the lowest latency regardless of the fine-tuning strategy, reducing inference time from approximately 150–280 ms observed on conventional TensorFlow CPU and GPU execution to approximately 7 ms per image. This corresponds to a speedup of more than one order of magnitude compared with standard TensorFlow inference. The reduction in latency is primar-

Table 9: Bootstrap-Based Statistical Reliability of Official Test Performance.

Dataset	Backbone	Strategy	Classifier	Weighted F1	95% Bootstrap CI	Latency (ms)
Aruzz22-5K	EfficientNetB0	Frozen	MLP	0.919	[0.903–0.935]	303.500
Aruzz22-5K	MobileNetV3Small	Frozen	Logistic Regression	0.905	[0.886–0.921]	174.009
Aruzz22-5K	ResNet50	Frozen	Logistic Regression	0.892	[0.874–0.910]	235.317
Bangladesh	EfficientNetB0	Frozen	MLP	0.638	[0.610–0.664]	292.353
Bangladesh	MobileNetV3Small	Frozen	MLP	0.639	[0.612–0.667]	157.455
Bangladesh	ResNet50	Frozen	MLP	0.598	[0.572–0.626]	234.732
Japanese	EfficientNetB0	Frozen	Logistic Regression	0.866	[0.839–0.893]	295.802
Japanese	MobileNetV3Small	Frozen	Logistic Regression	0.857	[0.832–0.883]	166.301
Japanese	ResNet50	Frozen	Logistic Regression	0.843	[0.813–0.871]	237.837
Koklu	EfficientNetB0	Frozen	Logistic Regression	0.996	[0.992–0.999]	300.431
Koklu	MobileNetV3Small	Frozen	Logistic Regression	0.991	[0.986–0.996]	158.077
Koklu	ResNet50	Frozen	Logistic Regression	0.989	[0.983–0.995]	236.258

ily attributable to the lightweight execution engine and optimization mechanisms employed by TensorFlow Lite, which are specifically designed for deployment on resource-constrained devices.

Across all evaluated platforms, GPU execution generally outperformed CPU execution, although the magnitude of improvement varied according to the fine-tuning strategy. Interestingly, the latency differences among frozen, partial, and full fine-tuning configurations were relatively small compared with the differences introduced by the deployment platform itself. This observation suggests that deployment environment selection has a greater impact on inference efficiency than the choice of backbone adaptation strategy.

These findings provide additional evidence supporting the deployment-oriented nature of the proposed framework. While the official experiments were conducted using conventional TensorFlow environments, the TensorFlow Lite results demonstrate that the framework can be effectively deployed in edge-oriented scenarios with substantially reduced inference time while maintaining the predictive performance reported in the previous sections. Consequently, MobileNetV3Small combined with TensorFlow Lite represents a particularly attractive solution for real-time rice variety classification applications where computational resources and response time are critical constraints.

Table 10: Cross-Platform Latency Benchmarking of MobileNetV3Small Across Fine-Tuning Strategies and Deployment Environments.

Fine-Tuning Strategy	TensorFlow CPU (ms)	TensorFlow GPU (ms)	TensorFlow Lite CPU (ms)
Frozen	267.03 ± 152.99	161.49 ± 25.68	7.57 ± 0.95
Partial	225.45 ± 69.02	154.20 ± 18.11	5.10 ± 0.54
Full	211.50 ± 24.76	151.97 ± 19.96	5.78 ± 0.67

4.5 Explainability Analysis

To better understand the remaining classification errors and investigate whether misclassifications occur systematically, an explainability analysis was conducted using Grad-CAM visualizations on the most frequently confused class pairs. Table 11 summarizes the top three misclassification pairs identified from the best-performing model for each dataset. These hard pairs were subsequently analyzed using Grad-CAM to examine the visual regions that contributed most strongly to the model’s predictions.

The results indicate that the frequency and severity of misclassification varied substantially across datasets. The Japanese Rice Dataset exhibited the highest concentration of confusion, particularly between the Koshihikari and Akitakomachi varieties. The most frequent error involved Koshihikari being classified as Akitakomachi, occurring 40 times, followed by the reverse confusion with 28 instances.

These findings suggest that the two varieties share highly similar visual characteristics, making them difficult to distinguish even under the best-performing configuration.

For the Aruzz22-5K dataset, the most common misclassification occurred between Gutisharna and Paijam, with 14 incorrectly classified samples. Interestingly, the reverse confusion was also observed among the top three hard pairs, indicating a bidirectional relationship between these varieties. Additional confusion was observed between BR29 and BR28, which are morphologically related rice varieties. These results suggest that the remaining errors are concentrated among visually similar classes rather than being randomly distributed across the dataset.

The Bangladesh dataset exhibited a broader distribution of misclassifications due to its larger number of classes. The most frequent confusion occurred between BD56 and BD30, followed by Binadhan7 and Binadhan14. Despite the increased classification complexity associated with thirty-eight rice varieties, the number of errors within each hard pair remained relatively small, suggesting that the proposed framework successfully learned discriminative features for

most classes.

The Koklu dataset demonstrated the highest degree of class separability. The most frequent misclassification, Basmati being predicted as Jasmine, occurred only three times. All remaining hard pairs were associated with a single error. These findings are consistent with the near-perfect classification performance reported in the previous section and indicate that the learned feature representations are highly discriminative for this dataset.

Figures 4–7 present representative Grad-CAM visualizations generated from the best-performing model for the four most challenging misclassification pairs. The heatmaps reveal that the best-performing model consistently focuses on grain shape, texture distribution, and boundary-related characteristics when making predictions. In several cases, the highlighted regions overlap substantially between the true and predicted classes, providing a plausible explanation for the observed confusion. Importantly, the Grad-CAM results suggest that the model attends to meaningful morphological structures rather than irrelevant background regions, indicating that the learned representations remain semantically rele-

Table 11: Top-3 Most Frequent Misclassification Pairs Identified from the Best-Performing Models.

Dataset	Rank	True Class	Predicted Class	Count
Aruzz22-5K	1	Gutisharna	Paijam	14
Aruzz22-5K	2	BR29	BR28	8
Aruzz22-5K	3	Paijam	Gutisharna	8
Bangladesh	1	BD56	BD30	8
Bangladesh	2	Binadhan7	Binadhan14	7
Bangladesh	3	BD30	BD49	5
Japanese	1	Koshihikari	Akitakomachi	40
Japanese	2	Akitakomachi	Koshihikari	28
Japanese	3	Yangdao-8	Akitakomachi	1
Koklu	1	Basmati	Jasmine	3
Koklu	2	Arborio	Karacadag	1
Koklu	3	Ipsala	Arborio	1

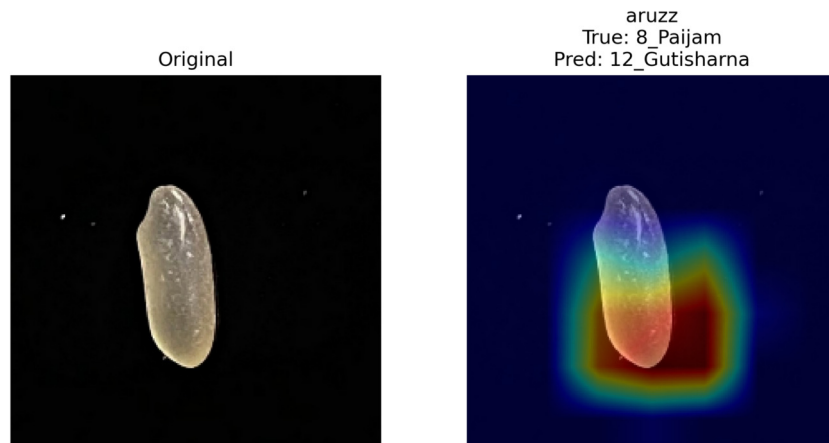


Fig.4: Grad-CAM visualization for the most frequent misclassification pair identified in the Aruzz22-5K dataset.

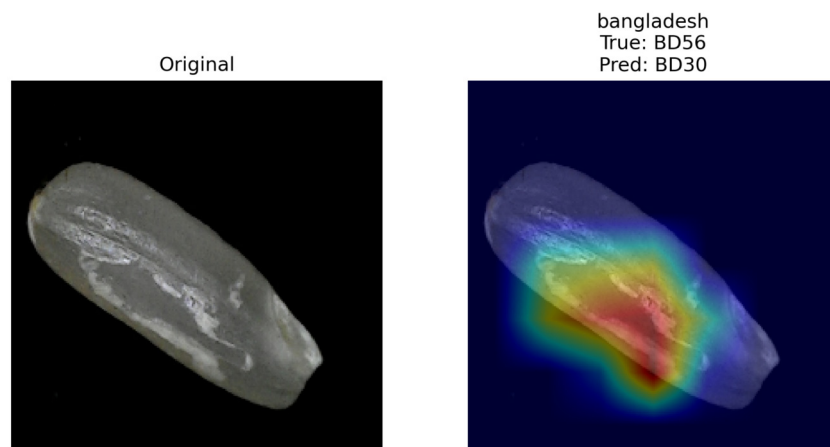


Fig.5: Grad-CAM visualization for the most frequent misclassification pair identified in the Bangladesh dataset.

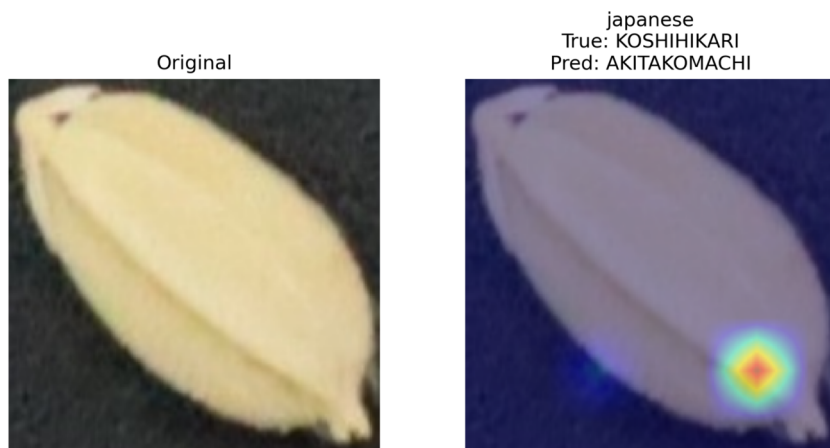


Fig.6: Grad-CAM visualization for the most frequent misclassification pair identified in the Japanese dataset .

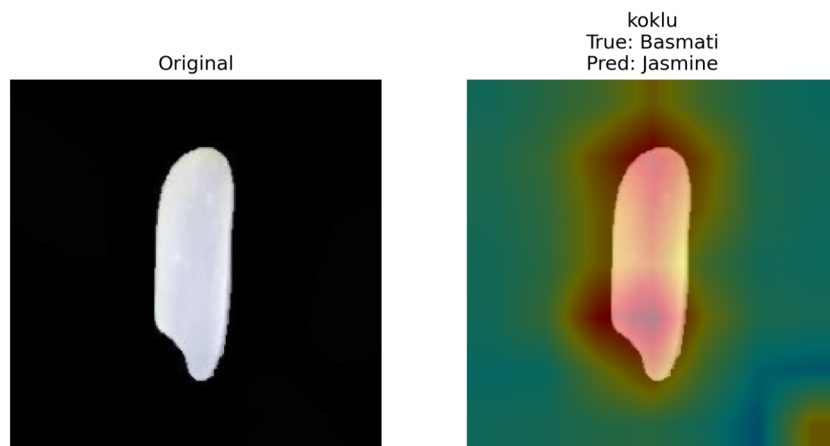


Fig.7: Grad-CAM visualization for the most frequent misclassification pair identified in the Koklu dataset .

vant even when misclassification occurs.

Overall, the explainability analysis demonstrates that classification errors are rare and highly systematic, predominantly occurring between visually similar rice varieties. The combination of confusion analysis and Grad-CAM visualization therefore provides additional evidence that the proposed framework learns meaningful and interpretable rice morphology representations while maintaining strong predictive performance across diverse datasets.

4.6 Discussion and Implications

Compared to existing rice variety classification studies, the proposed framework extends the state of the art by jointly addressing robustness, efficiency, and interpretability within a unified evaluation pipeline. Previous large-scale studies based on deep CNNs and transformer architectures primarily focused on maximizing classification accuracy under conventional train–test splits, without explicitly considering domain shifts, deployment constraints, or explainability [1], [6], [7]. As a result, the robustness of these models under realistic distribution shifts remains insufficiently understood.

A key contribution of this work is the introduction of proxy-domain GroupKFold evaluation. Unlike conventional random splits, the proposed strategy approximates latent acquisition domains using lightweight image statistics and evaluates model generalization across unseen proxy domains. The sensitivity analysis demonstrated that ($K=5$) provides an effective balance between cluster separability, domain diversity, and validation feasibility. This finding highlights the importance of domain-aware validation in agricultural vision applications where acquisition metadata are rarely available.

The study also demonstrates the effectiveness of PCA-based dimensionality reduction for deployment-oriented classification. While deep embeddings are highly expressive, the sensitivity analysis revealed that retaining 150 principal components preserves approximately 99.08% of the original variance while achieving the strongest cross-domain performance. Increasing dimensionality beyond this point provided only marginal gains in variance retention and did not consistently improve classification accuracy, indicating the presence of redundant information in higher-dimensional representations.

From a representation-learning perspective, the fine-tuning analysis revealed that partial fine-tuning generally provides a better trade-off between robustness and computational complexity than either fully frozen or fully trainable backbones. In particular, ResNet50 with partial fine-tuning consistently achieved the strongest cross-domain performance on the more challenging Aruzz22-5K, Bangladesh, and Japanese datasets. These findings suggest that limited adaptation of pretrained representations is often

sufficient to capture rice-specific morphological characteristics while preserving the generalization capability learned from large-scale ImageNet pretraining.

Comparative studies evaluating handcrafted and deep representations have consistently demonstrated the superiority of CNN-based features over traditional color, texture, and shape descriptors [13]. However, most prior works remained accuracy-centric and did not consider latency, domain-aware validation, or statistical reliability. Similarly, lightweight and quantized CNN approaches explored efficiency improvements [3], [8], whereas explainability-oriented studies such as [4] incorporated post hoc interpretation techniques without explicitly evaluating robustness under domain shifts.

In contrast, the proposed framework combines proxy-domain validation, PCA optimization, fine-tuning analysis, latency-aware benchmarking, and Grad-CAM-based explainability within a single evaluation protocol. Cross-platform benchmarking further demonstrated that TensorFlow Lite substantially reduces inference latency compared with conventional CPU and GPU execution, supporting efficient edge deployment without sacrificing predictive performance. Moreover, bootstrap confidence intervals provide additional evidence regarding the stability and reliability of the reported results.

Finally, the Grad-CAM analysis of hard misclassification pairs revealed that most remaining classification errors arise from intrinsic visual similarities among rice varieties rather than arbitrary model behavior. By focusing on systematically confused class pairs, the proposed explainability strategy offers more actionable insights than generic saliency visualizations and helps bridge the gap between predictive performance and model trustworthiness.

Overall, the proposed framework provides a more realistic and deployment-oriented evaluation methodology for rice variety classification than existing accuracy-driven approaches. By jointly considering robustness, efficiency, statistical reliability, and interpretability, the framework offers a practical foundation for developing trustworthy rice classification systems in real-world agricultural environments [1], [3], [4], [6], [7], [8], [13].

5. CONCLUSION

This study presented a latency-aware cross-domain rice variety classification framework that integrates proxy-domain grouping, GroupKFold validation, PCA-based dimensionality reduction, deep feature extraction, fine-tuning strategies, latency benchmarking, and explainability analysis within a unified evaluation pipeline. Unlike conventional rice classification studies that rely primarily on random train–test splits, the proposed framework explicitly evaluates robustness under domain shifts while simultaneously considering deployment efficiency and

model interpretability.

The proxy-domain sensitivity analysis demonstrated that $K = 5$ provides an effective balance between cluster separability, domain diversity, and GroupKFold feasibility. Likewise, the PCA sensitivity analysis showed that retaining 150 principal components preserves approximately 99.08% of the original variance while achieving the highest cross-domain weighted F1-score, reducing feature dimensionality by approximately 41% relative to the original embeddings.

Experimental results across four rice datasets revealed that no single backbone consistently dominates all evaluation scenarios. ResNet50 combined with partial fine-tuning generally provided the strongest cross-domain robustness, particularly on the Aruzz22-5K, Bangladesh, and Japanese datasets, whereas EfficientNetB0 frequently achieved the highest official test performance. MobileNetV3Small consistently offered the most favorable accuracy–latency trade-off and frequently appeared on the Pareto frontier, making it particularly attractive for deployment-constrained environments.

Cross-platform latency benchmarking further demonstrated that TensorFlow Lite substantially improves deployment efficiency, reducing inference latency to approximately 5–8 ms per image while preserving predictive performance. In addition, bootstrap confidence intervals confirmed the statistical reliability of the reported results, and Grad-CAM analysis showed that most remaining classification errors arise from intrinsic visual similarities among rice varieties rather than arbitrary model behavior.

Overall, the proposed framework provides a reproducible and deployment-oriented methodology for rice variety classification that jointly considers robustness, efficiency, and explainability. Future work may extend the framework to larger multi-country rice datasets, incorporate explicit domain adaptation techniques, and evaluate deployment on real mobile and edge devices under field conditions.

AUTHOR CONTRIBUTIONS

Conceptualization, N.S. and G.H.M.; methodology, N.S. and G.H.M.; software, G.H.M.; validation, N.S. and G.H.M.; formal analysis, N.S. and G.H.M.; investigation, N.S. and G.H.M.; resources, N.S.; data curation, G.H.M.; writing—original draft preparation, N.S. and G.H.M.; writing—review and editing, N.S. and G.H.M.; visualization, G.H.M.; supervision, N.S.; project administration, N.S. All authors have read and agreed to the published version of the manuscript.

References

- [1] M. M. Islam, M. S. Islam, M. G. Moazzam and M. S. Uddin, “A comprehensive deep learning framework for rice variety classification with real-time deployment,” *Smart Agricultural Technology*, vol. 13, p. 101710, 2026.
- [2] P. Saxena, K. Priya, S. Goel, P. K. Aggarwal, A. Sinha and P. Jain, “Rice varieties classification using machine learning algorithms,” *Journal of Pharmaceutical Negative Results*, vol. 13, no. 7, pp. 3762–3772, 2022.
- [3] M. J. Iqbal *et al.*, “On application of lightweight models for rice variety classification and their potential in edge computing,” *Foods*, vol. 12, no. 21, pp. 1–16, 2023.
- [4] M. J. Asif *et al.*, “Exploring Convolutional Neural Networks for Rice Grain Classification: An Explainable AI Approach,” *2025 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, pp. 1–6, 2025.
- [5] A. Rahman and S. Lee, “Rice seed varieties classified using diffusion convolutional neural networks,” 2024. [Online]. Available: https://sensors.myu-group.co.jp/sm_pdf/SM4066.pdf.
- [6] S. Hazra, S. Karforma, A. Bandyopadhyay, A. Banerjee and D. Chakraborty, “Deep learning-based classification of rice varieties: A comprehensive study,” *Journal of Survey in Fisheries Sciences*, vol. 10, no. 3, pp. 411–419, 2023. [Online]. Available: <https://www.researchgate.net/publication/385142789>.
- [7] M. Koklu, I. Cinar and Y. S. Taspinar, “Classification of rice varieties with deep learning methods,” *Computers and Electronics in Agriculture*, 2021.
- [8] M. Tasci, A. Istanbulu, S. Kosunalp, T. Iliev, I. Stoyanov and I. Beloev, “An efficient classification of rice variety with quantized neural networks,” *Electronics*, vol. 12, no. 10, pp. 1–16, 2023.
- [9] I. Gulrajani and D. Lopez-Paz, “In search of lost domain generalization,” in *Proc. Int. Conf. Learning Representations (ICLR)*, pp. 1–29, 2020.
- [10] P. W. Koh *et al.*, “WILDS: A benchmark of in-the-wild distribution shifts,” in *Proc. 38th Int. Conf. Machine Learning (ICML)*, 2021, pp. 5637–5664. [Online]. Available: <https://proceedings.mlr.press/v139/koh21a.html>.
- [11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, 2020.
- [12] M. Ryo, “Explainable artificial intelligence and interpretable machine learning for agricultural data analysis,” *Artificial Intelligence in Agriculture*, vol. 6, pp. 257–265, 2022.
- [13] V. T. Hoang, D. P. Van Hoai, T. Surinwarangkoon, H.-T. Duong and K. Meethongjan, “A comparative study of rice variety classifica-

- tion based on deep learning and hand-crafted features,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 14, no. 1, pp. 1–10, 2020.
- [14] G. M. S. Islam, M. M. Himel, M. G. Moazzam and M. S. Uddin, “An image dataset of rice varieties,” *Mendeley Data*. [Online]. Available: <https://data.mendeley.com/datasets/3mn9843tz2/3>.
- [15] M. Tahsin *et al.*, “An extensive image dataset for classifying rice varieties in Bangladesh,” *Mendeley Data*, ver. 1, 2024.
- [16] N. Rathnayake, A. Miyazaki, T. L. Dang, and Y. Hoshino, “Age classification of rice seeds in Japan using gradient-boosting and ANFIS algorithms,” *Sensors*, vol. 23, Art. no. 2828, 2023.
- [17] I. Cinar and M. Koklu, “Identification of rice varieties using machine learning algorithms,” *Journal of Agricultural Science*, vol. 28, no. 2, pp. 307–325, 2022.
- [18] S. Sagawa, T. B. Hashimoto, P. W. Koh and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2020, pp. 1–19. [Online]. Available: <https://arxiv.org/abs/1911.08731>.
- [19] J. G. A. Barbedo, “Factors influencing the use of deep learning for plant disease recognition,” *Biosystems Engineering*, 2018.
- [20] S. Gururangan *et al.*, “Scaling expert language models with unsupervised domain discovery,” *arXiv preprint arXiv:2303.14177*, 2023.
- [21] M. M. Islam, G. M. S. Himel, M. G. Moazzam and M. S. Uddin, “Artificial intelligence-based rice variety classification: A state-of-the-art review and future directions,” *Smart Agricultural Technology*, vol. 10, Art. no. 100788, 2025.
- [22] M. Wang, B. Zheng, G. Wang, J. Yang, J. Lu and W. Zhang, “A principal component analysis-based feature optimization network for few-shot fine-grained image classification,” *Mathematics*, vol. 13, no. 7, Art. no. 1098, 2025.
- [23] W. Jia, M. Sun, J. Lian, and S. Hou, “Feature dimensionality reduction: A review,” *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.
- [24] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, Art. no. 20150202, 2016.
- [25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, pp. 1–14, 2015.
- [26] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” *arXiv preprint arXiv:1811.12808*, 2020.
- [27] T. Gill, S. K. Gill, D. K. Saini, Y. Chopra, J. P. de Koff and K. S. Sandhu, “A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping,” *Phenomix*, vol. 2, no. 3, pp. 156–183, 2022.
- [28] Y. Bengio, A. Courville and P. Vincent, “Representation Learning: A Review and New Perspectives,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [29] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. 36th Int. Conf. Machine Learning (ICML)*, 2019.
- [30] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
- [31] D. Hendrycks and K. Gimpel, “Gaussian error linear units (GELUs),” *arXiv preprint arXiv:1606.08415*, 2016.
- [32] G. H. Martono and N. Sulistianingsih, “Fine-tuning transformer models for emotion detection in Indonesian tweets indicating depression,” *Vietnam Journal of Computer Science*, pp. 1–22, 2025.
- [33] E. Njor, C. Banbury and X. Fafoutis, “Fast data aware neural architecture search via supernet accelerated evaluation,” *arXiv preprint arXiv:2502.12690*, 2025.
- [34] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.
- [35] R. Linero-Ramos, C. Parra-Rodríguez and M. Gongora, “SVMobileNetV2: A hybrid and hierarchical CNN-SVM network architecture utilising UAV-based multispectral images and IoT nodes for the precise classification of crop diseases,” *AgriEngineering*, vol. 7, no. 10, Art. no. 341, 2025.
- [36] H. Jalil, A. Qazi, and A. Iqbal, “NeuRN: Neuro-inspired domain generalization for image classification,” in *Proc. AAAI Conf. Artificial Intelligence*, pp. 1–14, 2025.
- [37] I. H. Rather, S. Kumar and A. H. Gandomi, “Breaking the data barrier: A review of deep learning techniques for democratizing AI with small datasets,” *Artificial Intelligence Review*, vol. 57, no. 9, 2024.
- [38] S. Zhao, M. Gong, T. Liu, H. Fu and D. Tao, “Domain generalization via entropy regularization,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 16096–16107.
- [39] T. Matsuura and T. Harada, “Domain general-

ization using a mixture of multiple latent domains,” in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 11749–11756.

- [40] Z. Cheng, M. Liu, D. Yang, Z. Zhao, C. Yan and S. Wang, “Domain generalization for image classification with dynamic decision boundary,” *Pattern Recognition*, 2025.
- [41] B. Efron and R. Tibshirani, “The bootstrap method for assessing statistical accuracy,” *Behaviormetrika*, vol. 17, pp. 1–35, 1985.
- [42] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [43] A. Howard *et al.*, “Searching for MobileNetV3,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1314–1324.
- [44] A. Bailly *et al.*, “Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models,” *Computer Methods and Programs in Biomedicine*, vol. 213, Art. no. 106504, 2022.
- [45] M. U. I. Tamim, S. A. Hamim, S. Malik, M. F. Mridha and S. Mahmood, “InsightNet: A deep learning framework for enhanced plant disease detection and explainable insights,” *Plant Direct*, vol. 9, no. 7, pp. 1–18, 2025.



turer at Bumigora University. She can be reached at neny.sulistianingsih@universitasbumigora.ac.id.

Neny Sulistianingsih received a Bachelor of Informatics (S.Kom) and a Master of Computer (M.Kom) from the Universitas Islam Indonesia, Yogyakarta. She completed her Doctoral studies in 2023 at the Universitas Gadjah Mada, Yogyakarta. Her research interests include Machine Learning, Data Mining, Graph Mining, Big Data, Social Network Analysis, and Natural Language Processing. She is currently a Lecturer at Bumigora University. She can be reached at neny.sulistianingsih@universitasbumigora.ac.id.



Lecturer at Bumigora University. He can be reached at galih.hendro@universitasbumigora.ac.id.

Galih Hendro Martono received a Bachelor of Informatics (S.Kom) from Universitas Islam Indonesia and a Master of Engineering from Universitas Gadjah Mada, Yogyakarta. He completed her Doctoral studies in 2023 at the Universitas Gadjah Mada, Yogyakarta. Her research interests include Data Mining, Big Data, Social Network Analysis, and Natural Language Processing. He is currently a Lecturer at Bumigora University. He can be reached at galih.hendro@universitasbumigora.ac.id.