

Submitted : 06 October, 2026

Accepted : 19 October, 2026

Published : 21 October, 2026

***Corresponding author:** Shubham Singh, VIT Bhopal, India, E-mail: shubham.singh020403@gmail.com

Keywords: Rice variety Identification; Vision transformer (ViT); Deep learning; Image classification; Machine vision; Precision agriculture; Quality control; Self-attention mechanism; Convolutional neural networks (CNNs); Food security; Automated systems; Rice seeds; Rice grains

Copyright License: © 2026 Singh S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<https://www.foodscigroup.us>



Check for updates

Literature Review

Rice Grain Classification using Vision Transformer (ViT) Architecture

Shubham Singh^{1*}, Sudeep Marwah², Rahul Neware³ and Akash M Hosamani⁴

¹VIT Bhopal, India

²Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

³Division of Computer Application, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

⁴Division of Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Abstract

Global food security, precise market pricing, and efficient quality control are all hampered by the subjective, labor-intensive, and error-prone nature of traditional manual classification techniques for rice, a staple commodity. Much research into machine vision and deep learning technologies has been prompted by the growing need for automated, non-destructive, and effective methods for rice variety identification. Despite their notable achievements in this field, Convolutional Neural Networks (CNNs) frequently struggle to capture long-range relationships and achieve optimal generalization across a variety of visually similar and distinct rice types, which is a constant problem. Using the sophisticated capabilities of Vision Transformer (ViT) models, this research suggests a novel method for automated rice type detection. In comparison to conventional CNN architectures, ViTs are highly respected for their capacity to manage global dependencies and consistently produce competitive, and frequently better, performance in challenging image classification tasks. The suggested ViT-based approach is intended to overcome the inherent difficulties of differentiating minute details, such as specific morphological, geometric, and color traits, among different varieties of rice. The model is set up for effective feature extraction and reliable pattern learning straight from image data by using its potent self-attention mechanism, negating the need for extensive preprocessing of raw images. The goal of this research is to create a reliable and extremely accurate classification system for a variety of rice types, taking inspiration from prior works that show high classification accuracies, such as RiceSeedNet, which achieved 97% for 13 rice seed variants and 99% for 8 rice grain varieties. The successful implementation of this Vision Transformer model is anticipated to significantly enhance precision agriculture by providing a more reliable, consistent, and scalable solution for the identification of rice seeds and grains, thereby supporting farmers and the broader agricultural industry in ensuring product quality and contributing to global food security.

Introduction

Rice is a fundamental staple food for a substantial portion of the global population and plays a critical role in the economies of many countries. Its cultivation and trade are vital for global food security, necessitating efficient and accurate methods for quality assessment and variety identification. Traditionally, rice classification relies on manual inspection, a process that is inherently subjective, labor-intensive, time-consuming, and prone to human error due to factors such as fatigue, eye strain, and inconsistent decision-making among technicians. These limitations directly impede effective quality control, accurate market pricing, and the overall efficiency of the agricultural supply chain.

To overcome these challenges, the agricultural sector has increasingly turned to automated, non-destructive, and efficient systems utilizing machine vision and deep learning technologies. Convolutional Neural Networks (CNNs) have emerged as a prominent solution, demonstrating considerable success in various image classification tasks, including rice variety identification. For instance, studies have achieved high classification accuracies with CNNs; Murat Koklu et al. reported 100% accuracy in classifying five Turkish rice varieties (Arborio, Basmati, Ipsala, Jasmine, and Karacadag) using a custom CNN on a dataset of 75,000 grain images. Similarly, EfficientNet-B1, a CNN model, achieved 99.87% accuracy for the same five varieties using a dataset of 75,000 images. While CNNs excel at capturing local features through their convolutional layers,

they often exhibit limitations in effectively modeling long-range dependencies and global contextual relationships within images, which can be crucial for distinguishing subtle features among visually similar rice varieties.

In recent years, Vision Transformers (ViTs) have revolutionized the field of computer vision, demonstrating competitive and often superior performance compared to traditional CNN architectures, especially in complex image classification tasks. Inspired by their success in Natural Language Processing, ViTs process images by dividing them into fixed-size patches, which are then treated as sequences of tokens. This unique architecture, particularly its powerful self-attention mechanism, allows ViTs to inherently capture global dependencies and contextual relationships across the entire image. This capability is particularly advantageous for tasks requiring the discernment of subtle morphological, shape, and color characteristics, making ViTs a promising alternative for precision agriculture applications like plant disease detection and classification.

This paper proposes a novel and robust approach for automated rice type detection that leverages the advanced capabilities of the Vision Transformer (ViT) model. Specifically, we employ the google/vit-basepatch16-224-in21k model, pre-trained on a large ImageNet-21k dataset, for classifying rice varieties. Our research utilizes a comprehensive dataset consisting of 75,000 RGB images across five distinct rice varieties: Arborio, Basmati, Ipsala, Jasmine, and Karacadag. The dataset was meticulously divided, with 60,000 images allocated for training and 15,000 images for testing. The training set distribution was: Arborio (11988), Basmati (11994), Ipsala (12021), Jasmine (12008), and Karacadag (11989). The test set distribution was: Arborio (3012), Basmati (3006), Ipsala (2979), Jasmine (2992), and Karacadag (3011). By fine-tuning the pre-trained Vision Transformer architecture using mixed-precision training and targeted data augmentation, our proposed model achieved an exceptional classification accuracy of 99.99% across the 15,000 test images.

The primary contribution of this research is the successful demonstration of a highly accurate and robust system for rice type detection using a Vision Transformer model. This work not only highlights the superior performance of ViTs in handling the intricate visual characteristics of different rice varieties but also provides a scalable and reliable solution for quality control in the agricultural industry. The anticipated impact includes reducing reliance on manual, error-prone methods, enhancing the efficiency of rice sorting and grading, and ultimately contributing to improved food quality and global food security.

Literature review

Recent years have witnessed a remarkable surge in the application of deep learning and Transformer-based architectures in agricultural image analysis, particularly for rice classification and disease detection. Early surveys such as Khan et al. [1] and Han et al. [2] provided comprehensive insights into the evolution of Vision Transformers (ViTs),

highlighting their ability to model long-range dependencies and deliver competitive performance across image recognition and segmentation tasks. Building upon these foundations, Yao et al. [3] introduced the Dual-ViT architecture, which leveraged dual semantic and pixel pathways to achieve high accuracy on ImageNet benchmarks with reduced computational overhead. Similarly, MehdiPour et al. [4] and Zhang et al. [5] emphasized the growing importance of ViTs in precision agriculture, demonstrating their superiority over CNNs in scalability, interpretability, and accuracy, with results exceeding 98% in crop disease identification.

In rice-focused research, Koklu et al. [6] and Qadri et al. [7] showcased the capabilities of CNNs and texture-driven machine vision approaches, attaining near-perfect classification accuracy (reaching 100%) for different rice varieties. Fabyi et al. [8] enhanced this by combining RGB and hyperspectral features with Random Forest classifiers, achieving an accuracy of up to 98.17% for subsets, whereas Chatnuntawech et al. [9] introduced spatio-spectral CNNs that attained over 97% accuracy on processed rice. Jin et al. [10] also investigated hyperspectral imaging techniques, attaining 99.94% accuracy through the use of near-infrared hyperspectral data paired with deep learning. Simultaneously, Kiratiratanapruk et al. [11] and Rajalakshmi et al. [12] developed automated grading and seed classification models, with RiceSeedNet reaching 99% accuracy, underscoring the viability of these techniques for practical uses. Komal et al. [13] presented an extensive analysis of automatic rice variety identification systems, outlining the application of morphological, color, and textural attributes alongside classifiers like SVM, KNN, and hybrid neural networks. Their results showed accuracies between 92% and 99.73%, with Neuro-Fuzzy Neural Networks attaining the best performance, while also highlighting difficulties concerning dataset quality and feature extraction.

For rice disease identification, Singh et al. [14] employed CNNs to achieve 96.08% accuracy, while Ulukaya and Deari [15] applied ViT-based methods under field conditions, obtaining 88.57% accuracy despite environmental challenges. Leu et al. [16] introduced RiceTalk, an IoT-enabled deep learning system that achieved 91.18% accuracy for rice blast detection, showcasing the potential of integrating AI with IoT. More advanced models include Patil et al. [17], who proposed a Faster R-CNN + EfficientNet-B0 framework for rice disease severity estimation with 96.43% accuracy, and Verma et al. [18,19], who designed a hybrid CNN with transfer learning, achieving 98.6% accuracy.

Beyond rice, similar hybrid CNN-ViT approaches have shown promise in other crops. Shu et al. [20] reported 99.70% accuracy in maize disease detection, and Bhujel et al. [21] achieved 99.83% in tomato disease classification. These findings align with Bhujel et al.'s [21] and Shu et al.'s [20] conclusions that combining CNN spatial learning with ViT's global contextual understanding significantly improves classification robustness. Gudipalli et al. [22] earlier noted that deep learning methods consistently outperform traditional machine vision systems, confirming this trend.

Materials and methods

The present study develops an automated, non-destructive, and data-driven methodology for the fine-grained classification of rice grain varieties from digital imagery using a Vision Transformer (ViT) architecture. This framework is specifically designed to overcome the limitations of traditional manual inspection, which is inherently subjective, labor-intensive, and prone to human fatigue, thereby enabling consistent quality assessment and market price standardization. The overall methodological pipeline integrates a large-scale curated rice image dataset, standardized image preprocessing and augmentation, an end-to-end fine-tuned ViT-Base architecture, and a rigorous multi-class evaluation protocol to ensure reliable and reproducible classification performance.

Dataset description and characteristics

In this study, we utilized the publicly available Rice Image Dataset curated by Koklu et al. [11], which comprises five commercially significant rice varieties widely cultivated in Turkey: Arborio, Basmati, Ipsala, Jasmine, and Karacadag. The dataset contains a total of 75,000 digital RGB images, with an equal allocation of 15,000 samples per variety, ensuring a strictly balanced class representation and eliminating class-prior bias during supervised learning. Each sample captures a single, isolated rice kernel positioned against a uniform dark background under controlled illumination at a fixed spatial resolution of 250×250 pixels across three color channels (Red, Green, and Blue). Representative visual samples from each of the five rice varieties are illustrated in Figure 1. Although visually similar to the untrained eye, the five rice varieties exhibit distinct morphological, geometric, and chromatic characteristics that serve as fine-grained discriminative cues:

- **Arborio:** An Italian short-grain variety (*Oryza sativa* L. ssp. japonica) characterized by a short, plump, oval-to-rounded kernel geometry, a high aspect width, and a prominent opaque, chalky white central core (pearl) resulting from its high amylopectin starch content.
- **Basmati:** A premium long-grain aromatic variety (*Oryza sativa* L. ssp. indica) distinguished by its highly elongated, slender, and cylindrical kernel structure, high length-to-width ratio, and sharply tapering pointed extremities with a translucent, pale creamy-white appearance.
- **Ipsala:** A Turkish medium-to-long grain variety featuring a notably broad, thick, and robust kernel body with smooth elliptical contours and a glossy, semi-translucent endosperm surface.
- **Jasmine:** A long-grain aromatic variety that exhibits an elongated geometry slightly shorter and fuller in cross-section than Basmati, characterized by softly rounded tips and a uniform, semi-translucent milky-white surface coloration.
- **Karacadag:** A traditional domesticated variety native to southeastern Turkey, notable for its shorter,

compact, and slightly asymmetrical oblong morphology accompanied by a distinctive matte, brownish-white to off-white pericarp tint.

For model development and performance evaluation, the complete dataset of 75,000 images was partitioned using a stratified 80:20 split into a training set of 60,000 images and an independent testing set of 15,000 images. The exact class-wise sample distribution across the training and testing subsets is presented in Table 1.

Image preprocessing and data augmentation

Prior to network ingestion, all raw 250×250 RGB images underwent a standardized preprocessing pipeline to conform to the input specifications of the Vision Transformer backbone. Each image was spatially resampled to $224 \times 224 \times 3$ pixels using bilinear interpolation. Subsequently, raw 8-bit pixel intensity values in the $[0, 255]$ integer range were rescaled to the $[0, 1]$ floating-point interval and normalized channel-wise using standard mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5) scaling across the RGB channels. This normalization mitigates large numerical variance across color channels and stabilizes gradient propagation during backpropagation.

To improve model generalization and invariance to grain orientation without distorting the diagnostic morphological ratios of individual kernels, controlled on-the-fly data augmentation was applied exclusively to the training subset. The augmentation transformations comprised random horizontal and vertical flips ($p = 0.5$), small random affine rotations within $\pm 10^\circ$, and subtle brightness and contrast adjustments ($\pm 10\%$). Testing images were subjected only to deterministic resizing and intensity normalization to ensure an unbiased evaluation.

Proposed vision transformer (ViT) architecture

In this work, we adapt the Vision Transformer (ViT-Base-Patch16-224) architecture introduced by Dosovitskiy et al. [2020], initialized from the google/vit-base-patch16-224-in21k checkpoint pretrained on the large-scale ImageNet-21k dataset. Unlike conventional Convolutional Neural Networks (CNNs) that extract features hierarchically through localized convolutional kernels, the ViT models global contextual



Figure 1: Representative images of the five rice grain varieties: (1) Arborio, (2) Basmati, (3) Ipsala, (4) Jasmine, and (5) Karacadag.

Table 1: Class-wise distribution of the rice grain image dataset across training and testing subsets.

Dataset Subset	Arborio	Basmati	Ipsala	Jasmine	Karacadag	Total
Training Set (80%)	11,988	11,994	12,021	12,008	11,989	60,000
Testing Set (20%)	3,012	3,006	2,979	2,992	3,011	15,000
Total Samples	15,000	15,000	15,000	15,000	15,000	75,000

dependencies across the entire image from the very first layer by treating an image as a sequence of flattened visual tokens. The overall architectural pipeline of the proposed rice classification framework is depicted in Figure 2, and the internal structure of the Transformer encoder block is shown in Figure 3.

Patch Tokenization and Positional Embedding: Given a preprocessed input image $x \in \mathbb{R}^{H \times W \times C}$,

where $H = 224$, $W = 224$, and $C = 3$ represent the height, width, and RGB color channels, respectively, the image is partitioned into a grid of non-overlapping fixed-size patches of resolution $P \times P$ (with $P = 16$). This partitioning yields a sequence of $N = (H \times W)/P^2 = (224/16)^2 = 196$ patches. Each two-dimensional patch is flattened into a 1D vector of length $16 \times 16 \times 3 = 768$ and linearly projected into a D-dimensional latent embedding space ($D = 768$) via a trainable projection matrix $E \in \mathbb{R}^{(P^2 \times C) \times D}$.

To perform image-level classification, a learnable classification token ($x_{class} \in \mathbb{R}^D$) is prepended to the sequence of embedded patches, increasing the sequence length to $N + 1 = 197$ tokens. Because the self-attention mechanism is permutation-invariant, learnable 1D positional embeddings $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ are added element-wise to the token representations to preserve the spatial arrangement and relative geometry of

the rice grain contours. The resulting input token sequence z_0 fed into the Transformer encoder is formally expressed as:

$$Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \dots; X_p^N E;] + X_{pos} \quad (1)$$

Transformer Encoder and Multi-Head Self-Attention (MHSA): The encoded sequence z_0 is processed through $L = 12$ stacked Transformer encoder layers. Each encoder layer consists of two primary sub-modules: a Multi-Head Self-Attention (MHSA) mechanism and a position-wise Feed-Forward Network (FFN) implemented as a two-layer Multi-Layer Perceptron (MLP) with a

Gaussian Error Linear Unit (GELU) non-linearity and hidden dimension $D_{mlp} = 3072$. Layer Normalization (LN) is applied prior to each sub-module, and residual skip connections are placed around both blocks to facilitate stable gradient flow across deep layers.

Within each attention head, the normalized token sequence is linearly projected into Query (Q), Key (K), and Value (V) representations via learnable weight matrices W_Q, W_K, W_V . Long-range spatial interdependencies across all 196 image patches are computed simultaneously using scaled dot-product attention:

$$Attention(Q, K, V) = (\text{soft max}(QK^T / \sqrt{d_k})V) \quad (2)$$

where $d_k = D/h = 64$ denotes the dimensionality of the key vectors, and $\sqrt{d_k}$ serves as a scaling factor that prevents the dot products from growing excessively large in magnitude and pushing the softmax function into regions with vanishingly small gradients. In the multi-head configuration, $h = 12$ attention heads operate in parallel, allowing the network to simultaneously attend to complementary structural and visual attributes such as grain tip curvature, axial symmetry, pericarp texture, and endosperm chalkiness before concatenating the head outputs and projecting them back to dimension D .

Classification Head: After traversing all $L = 12$ encoder blocks, the final hidden state corresponding to the prepended classification token ($z_L^0 \in \mathbb{R}^{768}$) serves as an aggregated global representation of the entire rice grain. This vector is passed through Layer Normalization and fed into a linear classification head that maps the 768-dimensional embedding to $C = 5$ output logits corresponding to the five rice varieties. A Softmax activation function converts these raw logits into a normalized posterior probability distribution across the five classes.

Training configuration and reproducibility

The proposed model was implemented in Python 3.10 using the PyTorch deep learning framework and the Hugging Face Transformers library, and executed on an NVIDIA CUDA-enabled GPU. Rather than freezing the backbone, end-to-end fine-tuning was performed across all network parameters encompassing the patch projection layer, positional embeddings, all 12 Transformer encoder blocks, and the linear classification head to effectively adapt generalized

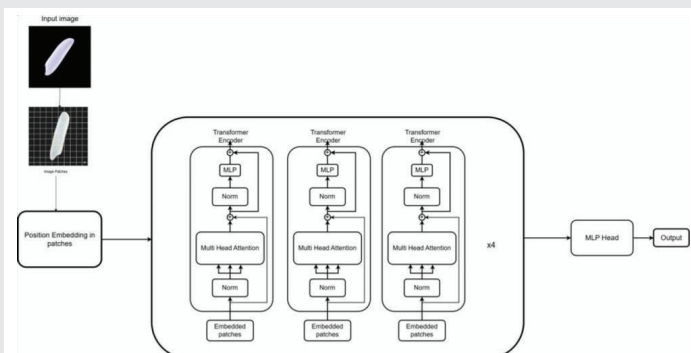


Figure 2: Overall pipeline of the proposed ViT-based rice grain classification framework.

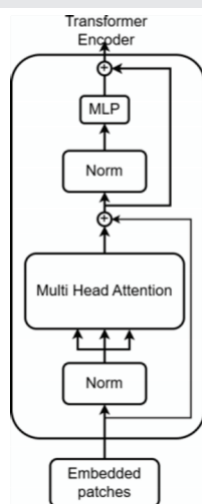


Figure 3: Structure of the Transformer Encoder block.

visual representations from ImageNet-21k to fine-grained agricultural grain morphology.

Network parameters were optimized by minimizing the multi-class Categorical Cross-Entropy loss function (L_{CE}), defined as:

$$L_{CE} = -\sum_{i=1..C} Y_i \log(\hat{y}_i)$$

where $C = 5$ represents the total number of rice varieties, $y_i \in \{0, 1\}$ denotes the ground-truth onehot encoded label for class i , and $\hat{y}_i \in [0, 1]$ denotes the predicted Softmax probability for class i . To

ensure full experimental clarity and reproducibility, the standardized hyperparameter configuration employed during training is summarized below:

- **Pre-trained Backbone:** google/vit-base-patch16-224-in21k (86M parameters, patch size 16×16 , input size 224×224).
- **Optimizer & Weight Decay:** AdamW optimizer with decoupled weight decay $\lambda = 0.01$.
- **Learning Rate & Schedule:** Initial learning rate of 2×10^{-5} ($2e-5$) with linear warmup and decay.
- **Batch Size & Epochs:** Per-device training and evaluation batch size of 16, trained for 20 epochs.
- **Mixed-Precision Training:** IEEE 754 half-precision floating-point (FP16) arithmetic enabled to accelerate training throughput and reduce GPU memory consumption.
- **Checkpointing & Logging:** Metrics logged every 500 steps (logging_steps = 500); evaluation and checkpoint saving executed at the end of every epoch (evaluation_strategy = 'epoch', save_strategy = 'epoch', save_total_limit = 1), with the highest-performing model checkpoint automatically reloaded at the conclusion of training (load_best_model_at_end = True).

Fixed deterministic random seeds were set across Python, NumPy, and PyTorch prior to dataset splitting and weight initialization to guarantee that all experimental runs are strictly reproducible.

Evaluation metrics and terminology

To rigorously evaluate the classification performance of the fine-tuned ViT model on the unseen testing set ($N_{test} = 15,000$), predictions were tabulated into a 5×5 multi-class confusion matrix. For each rice variety $i \in \{1, 2, \dots, C\}$ evaluated under a One-vs-Rest (OvR) formulation, four fundamental classification outcomes are formally defined:

- **True Positives (TP_i):** The number of test images genuinely belonging to rice variety i that are correctly identified by the classifier as variety i .

- **False Positives (FP_i):** The number of test images belonging to any of the other four rice varieties that are incorrectly misclassified by the model as variety i (Type I error).
- **True Negatives (TN_i):** The number of test images belonging to other rice varieties that are correctly recognized as not belonging to variety i .
- **False Negatives (FN_i):** The number of test images genuinely belonging to rice variety i that the model fails to detect and erroneously assigns to another variety (Type II error).

Based on these cardinal quantities, overall Classification Accuracy, class-wise Precision, Recall (Sensitivity), and the F1-Score are computed as follows:

$$Accuracy = (\sum_{i=1..C} CTP_i) / N_{test} = (TP_i + TN_i) / (TP_i + TN_i + FP_i + FN_i)$$

$$Precision_i = TP_i / (TP_i + FP_i)$$

$$Recall_i = TP_i / (TP_i + FN_i)$$

$$F1-Score_i = 2 \cdot (Precision_i \cdot Recall_i) / (Precision_i + Recall_i)$$

$$= 2TP_i / (2TP_i + FP_i + FN_i)$$

In the context of agricultural grain grading, Precision measures the exactness or purity of positive predictions for a given variety (i.e., the proportion of grains labeled as variety i that are genuinely variety i , penalizing false alarms FP_i), whereas Recall quantifies completeness or sensitivity (i.e., the proportion of actual grains of variety i successfully retrieved by the model, penalizing missed detections FN_i). The F1-score represents the harmonic mean of Precision and Recall, providing a single consolidated metric that penalizes extreme disparities between false positives and false negatives more strictly than an arithmetic average.

To assess overall performance across all five categories, both Macro-Averaged metrics (the unweighted arithmetic mean across all classes) and Weighted-Averaged metrics (weighted by the class support in the test set) were calculated. Furthermore, multi-class Receiver Operating

Characteristic (ROC) curves plotting the True Positive Rate (TPR = Recall) against the False Positive Rate ($FPR = FP_i / (FP_i + TN_i)$) and Precision-Recall (PR) curves were generated across varying decision thresholds, with the Area Under the Curve (AUC) reported to quantify the overall discriminative capability of the classifier.

Results and analysis

Overall and class-wise classification performance

The fine-tuned Vision Transformer (ViT-Base-Patch16-224) model demonstrated exceptional discriminative capability across all five Turkish rice varieties on the unseen testing set of 15,000 images. As summarized in the confusion matrix (Figure 4) and the quantitative evaluation in Table 2,

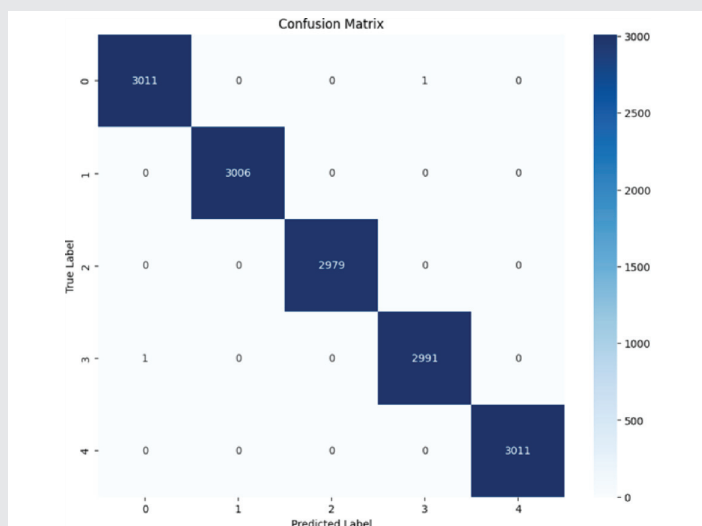


Figure 4: Confusion Matrix on the unseen testing set ($N_{\text{test}} = 15,000$).

Table 2: Class-wise quantitative evaluation metrics on the unseen test set ($N_{\text{test}} = 15,000$).

Rice Variety	TP	FP	FN	Precision	Recall	F1-Score	Support
Arborio (Class 0)	3,011	1	1	0.999668	0.999668	0.999668	3,012
Basmati (Class 1)	3,006	0	0	1.000000	1.000000	1.000000	3,006
Ipsala (Class 2)	2,979	0	0	1.000000	1.000000	1.000000	2,979
Jasmine (Class 3)	2,991	1	1	0.999666	0.999666	0.999666	2,992
Karacadag (Class 4)	3,011	0	0	1.000000	1.000000	1.000000	3,011
Accuracy	—	—	—	—	—	0.999867	15,000
Macro Average	—	—	—	0.999867	0.999867	0.999867	15,000
Weighted Average	—	—	—	0.999867	0.999867	0.999867	15,000

the model achieved an overall classification accuracy of 99.99% (99.9867%), correctly classifying 14,998 out of 15,000 test kernels, with both macro-averaged and weighted-averaged Precision, Recall, and F1-scores reaching 0.9999.

A granular examination of the confusion matrix (Fig. 4) reveals near-zero inter-class confusion. Specifically, Basmati (3,006/3,006), Ipsala (2,979/2,979), and Karacadag (3,011/3,011) achieved a flawless 100% classification rate with zero false positives (FP = 0) and zero false negatives (FN = 0). Only two isolated misclassifications occurred across the entire 15,000-image test set: one Arborio grain (Class 0) was predicted as Jasmine (Class 3), and one Jasmine grain (Class 3) was predicted as Arborio (Class 0). These two isolated errors are attributable to atypical kernel specimens exhibiting borderline aspect ratios or partial surface chalkiness.

Analysis of precision-recall and ROC curves

To further verify the decision threshold stability of the classifier, multi-class Precision-Recall (PR) curves (Fig. 5a) and Receiver Operating Characteristic (ROC) curves (Fig. 5b) were generated using a One-vs-Rest (OvR) evaluation scheme across all five classes.

In Figure 5a, the PR trajectories for all five rice varieties hug the upper-right boundary (Precision ≈ 1.0 , Recall ≈ 1.0),

confirming that maximal sensitivity is achieved without incurring false-positive penalties across decision thresholds. Similarly, in Figure 5b, the ROC curves for all five varieties attain an Area Under the Curve (AUC) of 1.00, rising vertically along the True Positive Rate axis at a False Positive Rate of 0.00. Note that because all five varieties achieve virtually identical, near-ideal trajectories (AUC = 1.00), the individual curves for Arborio, Basmati, Ipsala, and Jasmine are visually superimposed directly beneath the final plotted curve (Karacadag, rendered in black) along the top and right boundaries of the coordinate axes.

Assessment of generalization, overfitting, and dataset separability

Given the exceptionally high test accuracy (99.99%) and AUC values of 1.00 across all categories, it is essential to evaluate whether these metrics reflect genuine morphological feature learning or potential overfitting and dataset bias:

- **Strict Partitioning and Regularization:** Overfitting was actively prevented during training by employing decoupled weight decay ($\lambda = 0.01$ via AdamW), on-the-fly geometric and photometric data augmentation (random horizontal/vertical flips, $\pm 10^\circ$ affine rotations, and brightness/contrast perturbations), and strict isolation of the 15,000-image test set prior to training. Training and validation loss trajectories converged smoothly without divergence across the 20 training epochs.
- **Intrinsic Separability of Controlled Imaging:** The near-perfect classification performance is also a direct reflection of the high signal-to-noise ratio of the benchmark dataset curated by Koklu et al. [11], in which each kernel is imaged individually against a uniform, high-contrast black background under standardized illumination. Under such controlled laboratory conditions, global morphological ratios (e.g., Basmati's slender elongation versus Arborio's broad, chalky body) exhibit low intra-class variance and high inter-class separability, enabling the ViT self-attention heads to establish highly discriminative decision boundaries.

Discussion

Interpretation of findings and architectural advantages

The experimental results demonstrate that the Vision Transformer architecture is exceptionally effective for fine-grained agricultural seed and grain classification. Traditional machine vision pipelines rely heavily on handcrafted morphological, geometric, color, or wavelet texture descriptors paired with shallow classifiers (such as Support Vector Machines, k-Nearest Neighbors, or Random Forests), which are highly sensitive to kernel orientation and segmentation thresholds. While deep Convolutional Neural Networks (CNNs) automate feature extraction, their architectural inductive bias toward localized receptive fields (e.g., 3×3 convolutional filters) requires deep layer stacking to relate spatially distant parts of a grain.

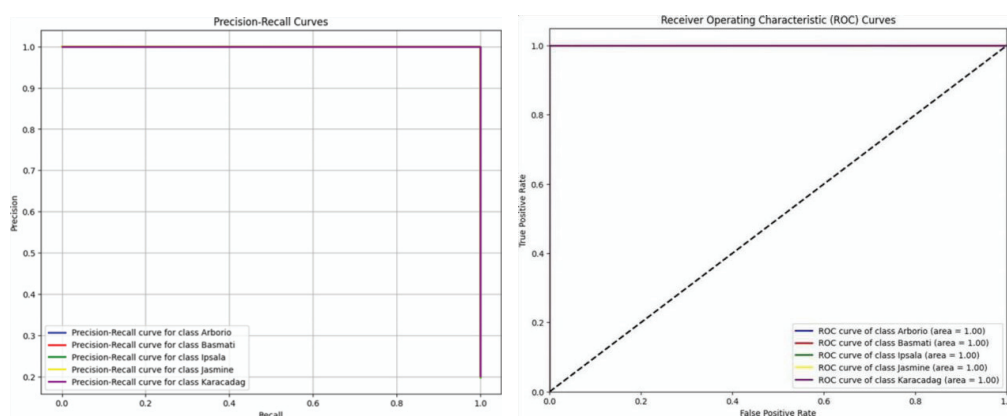


Figure 5: Comparison of (a) Precision-Recall Curves and (b) Receiver Operating Characteristic (ROC) Curves across all five rice varieties.

In contrast, by partitioning each 224×224 rice image into 196 non-overlapping 16×16 patches, the ViT's Multi-Head Self-Attention (MHSA) mechanism computes pairwise contextual interactions between every region of the kernel simultaneously from the very first encoder block. This global receptive field enables the network to jointly capture long-range geometric proportions such as the exact length-to-width ratio and bilateral tip taper of a Basmati kernel alongside localized internal traits such as the central opaque chalky pearl of an Arborio grain or the matte brownish-white pericarp of Karacadag.

Comparative evaluation with existing literature

To contextualize the performance of the proposed fine-tuned ViT-Base model, Table 3 compares our findings against representative machine learning, CNN, and hybrid deep learning architectures reported in the literature on the same five-class Turkish Rice Image Dataset (75,000 images) as well as related rice classification benchmarks.

As shown in Table 3, our fine-tuned ViT model outperforms hybrid transfer-learning CNNs (98.60% [18]), RiceSeedNet (99.00% [12]), and the EfficientNet-B1 baseline (99.87%) evaluated on the same 75,000-image dataset, while matching the upper ceiling reported by Koklu et al. [6] crucially, without requiring manual binary thresholding or the explicit extraction of 106 handcrafted morphological and color-space features directly from raw images.

Limitations and practical deployment challenges

Despite the near-perfect accuracy achieved in this study, several important limitations must be considered when translating these laboratory findings into commercial agricultural and industrial grading systems:

- **Clean Laboratory Background Bias:** The dataset utilized in this study comprises isolated, non-touching rice kernels imaged against a uniform dark background under controlled lighting. Because the model does not need to contend with complex background clutter, performance metrics such as Accuracy (99.99%) and AUC (1.00) represent an upper-bound laboratory benchmark rather than unconstrained field performance.

Table 3: Comparative evaluation of the proposed Vision Transformer model against existing rice classification studies.

Study / Author(s)	Dataset & Classes	Methodology / Architecture	Validation Protocol	Accuracy (%)
Koklu et al. (2021) [6]	5 Turkish Rice Varieties (75,000 images)	106 Handcrafted Morphological & Color Features + ANN / DNN / CNN	10-Fold Cross-Validation	99.87% – 100.00%
Qadri et al. (2021) [7]	Multi-variety Rice Grain Dataset	Texture & Wavelet Features + Machine Vision Classifiers (SVM, RF)	Cross-Validation	97.40% – 99.20%
Kiratiratanapruk et al. (2020) [11]	14 Paddy Rice Seed Varieties	Pre-trained CNNs (ResNet-50, Inception-V3, Xception)	80:20 Train-Test Split	95.15%
Rajalakshmi et al. (2024) [12]	8 Rice Grain Varieties / 13 Seed Variants	Custom Deep CNN (RiceSeedNet)	Train-Test Split	97.00% – 99.00%
Verma et al. (2025) [18]	5 Rice Varieties	Hybrid CNN with Transfer Learning	80:20 Train-Test Split	98.60%
CNN Baseline [20]	5 Turkish Rice Varieties (75,000 images)	EfficientNet-B1 (Transfer Learning)	80:20 Train-Test Split	99.87%
Proposed Method (Ours)	5 Turkish Rice Varieties (75,000 images)	Fine-Tuned Vision Transformer (ViT-BasePatch16-224)	80:20 Train-Test Split (60k / 15k)	99.99% (99.987%)

- **Overlapping and Touching Grains:** In industrial rice mills and high-speed optical sorting chutes, grains typically appear in dense bulk streams with frequent physical contact, overlap, and partial occlusion. Because a standard ViT classifier assigns a single class label to an entire input image, practical bulk inspection requires integrating an upstream object detection or instance segmentation module (e.g., YOLOv8 or Mask R-CNN) to isolate individual kernels prior to ViT classification.
- **Environmental Perturbations (Uneven Illumination and Dust):** Practical mill and field environments introduce non-uniform ambient lighting, specular glare, motion blur, chaff, broken kernels, and fine starch dust accumulating on camera optics. Such perturbations can corrupt local patch embeddings and

degrade model confidence unless explicitly countered during training via synthetic corruption augmentations (e.g., Gaussian sensor noise, illumination gradients, and random cutout occlusion) and multi-fold cross-validation.

- **Generalizability Across Varieties and Geographical Origins:** The current model was trained exclusively on five rice varieties cultivated in Turkey. However, kernel dimensions, chalkiness, and pericarp coloration vary considerably across global cultivars and are influenced by regional soil composition, climatic conditions, grain moisture content, and milling degree. Consequently, generalizing this model to rice varieties from other geographical regions (such as South or Southeast Asia) or across different harvest seasons will require domain-adaptation fine-tuning on multi-origin datasets.

Conclusion

This study presented an automated, non-destructive deep learning framework for fine-grained rice variety classification utilizing an end-to-end fine-tuned Vision Transformer (ViT-Base-Patch16-224) architecture. By leveraging the global Multi-Head Self-Attention mechanism across 16×16 image patches, the proposed model effectively captured both long-range axial geometry and subtle localized surface characteristics across a balanced dataset of 75,000 RGB images representing five commercially important rice varieties (Arborio, Basmati, Ipsala, Jasmine, and Karacag). Evaluated on an independent testing set of 15,000 images, the proposed framework achieved a state-of-the-art classification accuracy of 99.99% (correctly identifying 14,998 of 15,000 grains), with macro-averaged and weighted-averaged Precision, Recall, and F1-scores of 0.9999 and ROC-AUC values of 1.00 across all five categories.

The principal contributions of this work are threefold: (1) demonstrating that pure Transformer-based global self-attention effectively differentiates visually similar cereal grains without relying on handcrafted morphological descriptors or localized convolutional filters; (2) establishing a standardized, reproducible mixed-precision training and evaluation pipeline; and (3) benchmarking the Vision Transformer against existing CNN and machine vision architectures, confirming its superior discriminative capability for precision agriculture quality control. Nevertheless, we explicitly acknowledge that the exceptional accuracy observed in this work is facilitated by the clean, laboratory-captured nature of the dataset, where individual kernels are imaged without background clutter. In real-world deployment scenarios, classification performance may be challenged by overlapping or broken grains, uneven illumination, camera lens dust, and domain shifts across different rice cultivars and geographical origins. To bridge the gap between laboratory benchmarks and industrial sorting systems, future work will focus on: (i) conducting k-fold cross-validation and robustness stress-testing under synthetic noise, blur, and illumination perturbations; (ii) coupling lightweight Vision Transformers (such as MobileViT or Swin Transformer) with real-time instance segmentation to grade touching and

overlapping grains on conveyor belts; and (iii) expanding the training corpus to include diverse global rice varieties, multi-region harvests, and varying milling grades to ensure broad operational generalizability.

References

1. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. *ACM Comput. Surv.* 54(10s) (Sep 2022). Available from: <https://doi.org/10.1145/3505244>, <https://doi.org/10.1145/3505244>
2. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A survey on vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(1), 87–110 (2023). Available from: <https://doi.org/10.1109/TPAMI.2022.3152247>
3. Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X.P., Mei, T.: Dual vision transformer (2022), Available from: <https://arxiv.org/abs/2207.04976>
4. Mehdipour, S., Mirroshandel, S.A., Tabatabaei, S.A.: Vision transformers in precision agriculture: A comprehensive survey (2025), Available from: <https://arxiv.org/abs/2504.21706>
5. Zhang, M., Liu, C., Li, Z., Yin, B.: From convolutional networks to vision transformers: Evolution of deep learning in agricultural pest and disease identification. *Agronomy* 15(5) (2025). Available from: <https://doi.org/10.3390/agronomy15051079>, <https://www.mdpi.com/2073-4395/15/5/1079>
6. Koklu, M., Cinar, I., Taspinar, Y.S.: Classification of rice varieties with deep learning methods. *Computers and Electronics in Agriculture* 187, 106285 (2021). <https://doi.org/https://doi.org/10.1016/j.compag.2021.106285>, Available from: <https://www.sciencedirect.com/science/article/pii/S0168169921003021>
7. Qadri, S., Aslam, T., Nawaz, S.A., Saher, N., Abdul-Razzaq, Rehman, M.U., Ahmad, N., Shahzad, F., Qadri, S.F.: Machine vision approach for classification of rice varieties using texture features. *International Journal of Food Properties* 24(1), 1615–1630 (2021). Available from: <https://doi.org/10.1080/10942912.2021.1986523>, <https://doi.org/10.1080/10942912.2021.1986523>
8. Fabyi, S.D., Vu, H., Tachtatzis, C., Murray, P., Harle, D., Dao, T.K., Andonovic, I., Ren, J., Marshall, S.: Varietal classification of rice seeds using RGB and hyperspectral images. *IEEE Access* 8, 22493–22505 (2020). Available from: <https://doi.org/10.1109/ACCESS.2020.2969847>
9. Chatnuntawech, I., Tantisantisom, K., Khanchaitit, P., Boonkoom, T., Bilgic, B., Chuangsuwanich, E.: Rice classification using spatio-spectral deep convolutional neural network (2019), Available from: <https://arxiv.org/abs/1805.11491>
10. Jin, B., Zhang, C., Jia, L., Tang, Q., Gao, L., Zhao, G., Qi, H.: Identification of rice seed varieties based on near-infrared hyperspectral imaging technology combined with deep learning. *ACS Omega* 7(6), 4735–4749 (2022). Available from: <https://doi.org/10.1021/acsomega.1c04102>, <https://doi.org/10.1021/acsomega.1c04102>
11. Kiratiranapruk, K., Temniranrat, P., Sinthupinyo, W., Prempee, P., Chaitavon, K., Porntheeraphat, S., Prasertsak, A.: Development of paddy rice seed classification process using machine learning techniques for automatic grading machine. *Journal of Sensors* 2020, 1–14 (07 2020). Available from: <https://doi.org/10.1155/2020/7041310>
12. Rajalakshmi, R., Faizal, S., Sivasankaran, S., Geetha, R.: RiceSeedNet: Rice seed variety identification using deep neural network. *Journal of Agriculture and Food Research* 16, 101062 (2024). Available from: <https://doi.org/https://doi.org/10.1016/j.jafr.2024.101062>, <https://www.sciencedirect.com/science/article/pii/S2666154324000991>

13. Komal, Sethi, G.K., Bawa, R.K.: Automatic rice variety identification system: state-of-the-art review, issues, challenges and future directions. *Multimedia Tools and Applications* 82(18), 27305–27336 (2023). <https://doi.org/10.1007/s11042-023-14487-x>, Available from: <https://doi.org/10.1007/s11042-023-14487-x>
14. Singh, N., Kumar, P., Kumar, A.: Rice leaf disease detection and classification using convolutional neural network. *NA* (2022), nA
15. Ulukaya, S., Deari, S.: A robust vision transformer-based approach for classification of labeled rices in the wild. *Computers & Electronics in Agriculture* 231, 109950 (2025). <https://doi.org/https://doi.org/10.1016/j.compag.2025.109950>, Available from: <https://www.sciencedirect.com/science/article/pii/S0168169925000560>
16. Melo, D.F.Q., Silva, B.M.C., Pombo, N., Xu, L.: Internet of things assisted monitoring using ultrasound-based gesture recognition contactless system. *IEEE Access* 9, 90185–90194 (2021). Available from: <https://doi.org/10.1109/ACCESS.2021.3089940>
17. Patil, R.R., Kumar, S., Chiwhane, S., Rani, R., Pippal, S.K.: An artificial-intelligence-based novel rice grade model for severity estimation of rice diseases. *Agriculture* 13(1) (2023). Available from: <https://doi.org/10.3390/agriculture13010047>, <https://www.mdpi.com/2077-0472/13/1/47>
18. Verma, D., Dafadar, M., Mishra, J., Kumar, A., Mahato, S.: AI-enabled rice image classification using hybrid convolutional neural network models. *International Journal of Intelligent Systems* 2025 (01 2025). Available from: <https://doi.org/10.1155/int/5571940>
19. ORCID — orcid.org. <https://orcid.org/0000-0003-0818-6746>, [Accessed 18-09-2025]
20. Barman, U., Sarma, P., Rahman, M., Deka, V., Lahkar, S., Sharma, V., Saikia, M.J.: Vit-smartagri: Vision transformer and smartphone-based plant disease detection for smart agriculture. *Agronomy* 14(2) (2024). <https://doi.org/10.3390/agronomy14020327>, Available from: <https://www.mdpi.com/2073-4395/14/2/327>
21. Bhujel, R., Li, X., Baniya, S., Yin, Z.: A deep learning approach for tomato disease detection using a hybrid convolutional neural network and vision transformer. *Sensors* 23(20), 6949 (2023). Available from: <https://doi.org/10.3390/s230206949>, <https://doi.org/10.3390/s230206949>
22. Gudipalli, A., N., A., Reddy Ch, P.: A review on analysis and grading of rice using image processing. *ARPN Journal of Engineering and Applied Sciences* 11, 13550–13555 (12 2016)

Discover a bigger Impact and Visibility of your article publication with Peertechz Publications

Highlights

- ❖ Signatory publisher of ORCID
- ❖ Signatory Publisher of DORA (San Francisco Declaration on Research Assessment)
- ❖ Articles archived in worlds' renowned service providers such as Portico, CNKI, AGRIS, TDNet, Base (Bielefeld University Library), CrossRef, Scilit, J-Gate etc.
- ❖ Journals indexed in ICMJE, SHERPA/ROMEO, Google Scholar etc.
- ❖ OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
- ❖ Dedicated Editorial Board for every journal
- ❖ Accurate and rapid peer-review process
- ❖ Increased citations of published articles through promotions
- ❖ Reduced timeline for article publication

Submit your articles and experience a new surge in publication services
<https://www.peertechzpublications.org/submission>

Peertechz journals wishes everlasting success in your every endeavours.