EJFA Emirates Journal of Food and Agriculture

Emirates Journal of Food and Agriculture 38: e175268
doi: 10.3897/ejfa.2026.175268

REVIEW ARTICLE

A review of machine vision and deep learning applications in grain quality inspection

Baosheng Wang[1]

1 *School of Computer and Software, Nanyang Institute of Technology, Nanyang, 473004, China*

Corresponding author: Baosheng Wang (baoshengwang2016@163.com)

Academic editor: Yonis Gulzar ♦ **Received** 17 October 2025 ♦ **Accepted** 23 February 2026 ♦ **Published** 27 April 2026

Abstract

The accurate and rapid evaluation of cereal grain quality is fundamental to global food security, efficient processing, and fair trade. Traditional manual inspection is constrained by subjectivity and labor intensity, prompting the development of automated, non-destructive methods. This review critically examines the technological evolution from classical machine vision systems based on handcrafted features to deep learning architectures, and finally to vision transformers. It analyzes how early approaches leveraged morphological, color, and texture descriptors under controlled conditions to achieve high classification accuracy, yet often exhibited limited robustness under domain shift. The emergence of convolutional neural networks enabled end-to-end feature learning, improving robustness for tasks such as varietal classification, defect detection, and sorting. More recently, vision transformers introduced a paradigm shift by modeling global dependencies and capturing subtle, spatially distributed patterns, offering advantages for complex diagnostic tasks, while their data hunger and computational demands remain significant bottlenecks. Beyond summarizing architectures, this review synthesizes reported evidence into an analytical perspective on accuracy–latency–memory trade-offs, the dominant failure modes that drive lab-to-line performance gaps, and the evaluation practices required to claim deployable reliability. We highlight enabling factors such as large-scale open datasets, model lightweighting for edge deployment, and explainable AI to enhance transparency and trust, and we outline future trends toward multi-modal sensing, hybrid CNN–Transformer architectures, zero-shot learning, and fully autonomous inspection-sorting systems. By framing these developments as responses to technical and practical constraints, this review provides a structured and actionable perspective on the past, present, and future of intelligent grain quality assessment.

Keywords

Convolutional neural networks, vision transformers, image processing, non-destructive sensing, agricultural technology

Introduction

Cereal grains, including wheat, rice, maize, and barley, form the cornerstone of global food security and are pivotal commodities in international trade (Evers and Millar 2002). The quality of these grains is not an abstract concept but a critical determinant of their market value, storage stability, nutritional content, and suitability for specific end-use applications, such as milling, baking, malting, and animal feed production (Szczypiński and Zapotoczny 2012). Consequently, the ability to assess grain quality accurately, consistently, and efficiently is a fundamental imperative for the entire agri-food value chain, from plant breeders and farmers to grain processors and consumers. This imperative has driven a decades-long search for technologies that can transcend the inherent limitations of traditional assessment methods and provide objective, data-driven insights into grain quality.

For centuries, the gold standard for grain quality assessment has been manual visual inspection by trained human experts (Ahmed et al. 2021). This method, while deeply in-

Copyright *Baosheng Wang*. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

grained in industry practice, is fraught with fundamental weaknesses that compromise its reliability and scalability in a modern globalized food system. The primary critique centers on its subjectivity; quality judgments can vary significantly between inspectors due to differences in experience, training, and even fatigue, leading to inconsistent and often irreproducible results (Wang et al. 2003). This subjectivity can create significant friction in trade and lead to unfair pricing outcomes for producers (Koklu et al. 2021). Furthermore, manual inspection is a labor-intensive and time-consuming process, creating significant bottlenecks at grain receival points and processing facilities, thereby increasing operational costs and reducing overall efficiency (Koklu et al. 2021). The repetitive nature of the task also makes it prone to human error, further undermining the accuracy of the assessment (Wang et al. 2015).

The manifest deficiencies of manual inspection have catalyzed the adoption of automated, non-destructive sensing technologies, with machine vision (MV) emerging as a transformative force (Ma et al. 2016). By leveraging digital imaging and computational algorithms, MV systems offer a means to quantify grain quality based on objective, measurable parameters, thereby replacing subjective human judgment with consistent, data-driven evaluation (Du et al. 2004). Early MV systems demonstrated remarkable success in automating tasks like varietal classification and defect detection by engineering specific features based on grain morphology, color, and texture (Gupta and Waoo 2024). This transition from manual art to automated analysis represents more than a mere technological upgrade; it signifies a fundamental philosophical shift in quality assessment itself. It moves the practice from a craft-based skill, dependent on the tacit knowledge of an individual inspector, to a quantifiable science grounded in objective data, paving the way for more transparent and equitable standards in the grain industry.

While foundational, this classical MV paradigm carried its own inherent limitation: a dependency on handcrafted, human-engineered features. This approach proved brittle and lacked the adaptability to handle the vast biological variability of cereal grains and the complexity of real-world operational conditions (Gupta and Waoo 2024). This critical inflexibility set the stage for the next major technological leap: the deep learning (DL) revolution (Yang et al. 2025). Deep learning models, particularly Convolutional Neural Networks (CNNs), instigated a paradigm shift by learning relevant features automatically and directly from image data, eliminating the need for manual feature design (Lingwal et al. 2021). This capability for end-to-end learning has unlocked unprecedented levels of performance in complex grain analysis tasks.

This review provides a critical and comprehensive analysis of this technological evolution. Section 2 deconstructs the classical machine vision paradigm, examining its foundational pipeline and the principles of handcrafted feature engineering. Section 3 details the transformative impact of CNNs, reviewing their architectural principles and diverse applications in grain assessment while analyzing the performance trade-offs of different network designs. Section 4 explores the current frontier of research, focusing on the emergence of Vision Transformers (ViTs) and critically debating their potential advantages and disadvantages compared to CNNs. Finally, Section 5 synthesizes the persistent, cross-cutting challenges facing the field—including data scarcity, model interpretability, and the gap between laboratory research and industrial deployment—and outlines the most promising future research directions that will shape the next generation of intelligent grain quality assessment systems. In addition to providing a structured tutorial-style progression from classical pipelines to modern end-to-end models, the specific contributions of this review are threefold: first, we consolidate evidence across representative grain inspection tasks (varietal identification, defect/damage recognition, and sorting-oriented detection) to clarify which visual cues are predominantly local versus globally distributed and how this cue structure motivates architecture choice; second, we synthesize deployability-oriented trade-offs by jointly discussing recognition performance with practical constraints such as data scale, computational cost, and edge feasibility, which are central to translating laboratory results into online inspection systems; third, we highlight evaluation practices that should accompany deployable claims, including robustness to acquisition variability and reporting of efficiency indicators alongside accuracy, building on the broader agri-food computer-vision literature and recent grain-specific benchmarks. Despite rapid progress, several research gaps remain that repeatedly limit real-world reliability. A major unresolved issue is domain shift between controlled imaging and industrial conditions, where changes in device, lighting, occlusion, and kernel clustering can dominate error modes and degrade in-line performance; this gap is explicitly exposed by large-scale grain benchmarks formulated around fine-grained recognition, domain adaptation, and out-of-distribution settings. Another gap is the continuing data bottleneck for data-hungry models, especially transformer-based approaches, which motivates increased attention to dataset scale/quality, transfer learning, and data-efficient learning strategies. Finally, practical adoption still hinges on two under-addressed needs: (i) transparent, interpretable decision support (to build trust among inspectors and operators) and (ii) standardized, comparable reporting of latency, model size, and robustness testing to enable engineering selection under throughput constraints.

Review methodology and PRISMA-style study selection

In order to improve transparency and reproducibility, this review adopted a structured literature identification and screening procedure following the reporting principles of the PRISMA 2020 statement. The search was designed to capture both classical machine vision studies (feature engineering, segmentation, texture/morphology/color descriptors) and modern deep learning work (CNNs, object

detection frameworks, and transformer-based architectures) applied to cereal grain quality inspection.

Information sources included Web of Science Core Collection, Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink, complemented by forward and backward citation tracking and targeted searches in Google Scholar to reduce omission of influential papers. The search window was set to January 1, 2000 through June 30, 2025 to align with the historical development discussed in Sections 2–5.

Search queries combined grain/cereal terms with vision and learning terms. A representative query string was: ("grain" OR "cereal" OR wheat OR rice OR maize OR barley OR kernel OR seed) AND ("machine vision" OR "computer vision" OR "image processing" OR "deep learning" OR CNN OR "convolutional neural network" OR YOLO OR "vision transformer" OR ViT OR Swin) AND (quality OR inspection OR grading OR defect OR damage OR sorting). Search strings were adapted to each database's syntax, and duplicates were removed using reference-management software prior to screening.

Eligibility criteria were defined a priori. Studies were included if they: (i) were peer-reviewed journal articles or full conference papers; (ii) reported an imaging-based approach (RGB, multispectral, hyperspectral, fluorescence, X-ray/CT, or related machine-vision modalities) for cereal grain or kernel/seed quality assessment; and (iii) presented sufficient methodological detail (data source, task definition, model/feature pipeline, and evaluation metrics) to support synthesis. Studies were excluded if they: (i) did not involve cereal grains or closely related grain kernels/seeds; (ii) focused exclusively on non-vision sensors without imaging (e.g., point spectroscopy only) unless directly integrated into an imaging workflow; (iii) were editorials, patents, abstracts without full methods, or non-technical commentary; or (iv) addressed plant/leaf disease recognition without a grain/kernal inspection component, as this falls outside the stated scope of grain quality inspection.

Study selection proceeded in two stages. First, titles and abstracts were screened for relevance to grain quality inspection and to the methodological themes of this review (handcrafted features, CNNs, transformers, deployment constraints, and datasets). Second, full texts were assessed against the inclusion/exclusion criteria, and the final included studies formed the evidence base for Sections 2–5 and the synthesis tables.

Foundations and handcrafted feature engineering

The advent of machine vision marked the first significant step toward automating cereal grain quality assessment, establishing a structured, algorithmic approach to replace subjective human inspection. This paradigm is defined by a sequential pipeline where explicit, human-defined rules and features are used to interpret visual data. The core philosophy of this approach was not to model the full complexity of the visual world, but rather to simplify the problem by controlling the environment and extracting a small set of highly informative, handcrafted features. While ultimately superseded by more flexible deep learning methods, this foundational paradigm established the viability of automated inspection and laid the groundwork for future innovations.

The classical machine vision pipeline

A typical MV system for grain analysis follows a well-defined, multi-stage workflow, as illustrated in Fig. 1. This

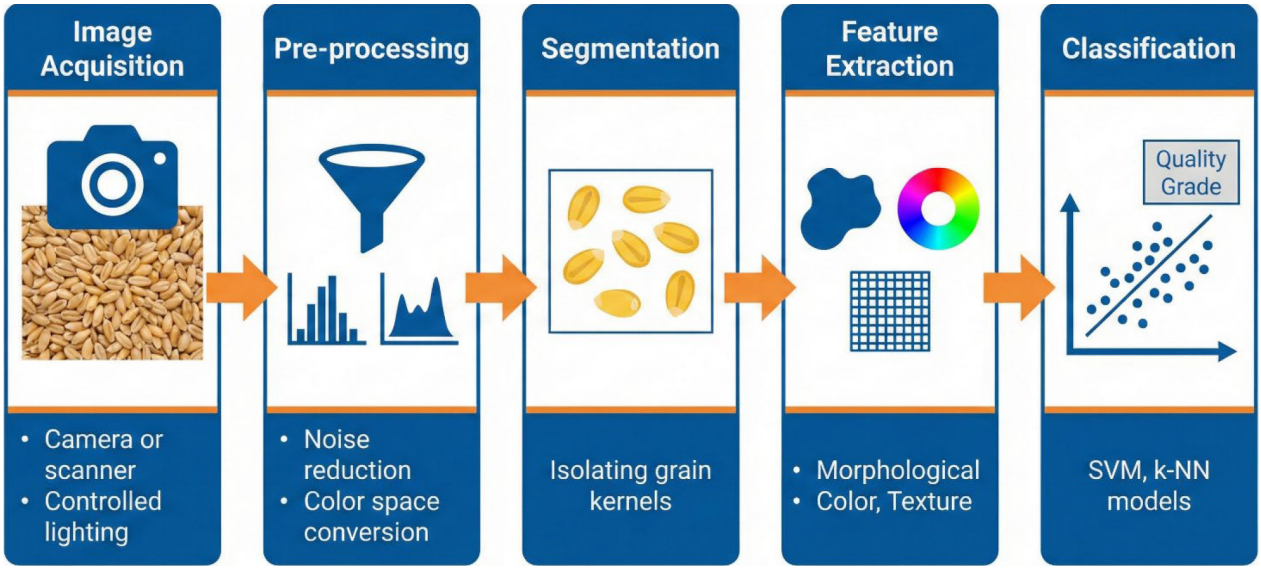

Figure 1. The classical machine vision pipeline for grain analysis. The workflow illustrates the sequential stages from image acquisition and pre-processing to segmentation, handcrafted feature extraction, and final classification using traditional machine learning models.

pipeline is designed to systematically reduce the complexity of raw image data into a final classification or quality score.

The initial stage, image acquisition, is critical for the success of the entire process. Early and effective systems often utilized high-resolution flatbed scanners or Charge-Coupled Device (CCD) cameras to capture images (Szczypiński and Zapotoczny 2012). A crucial element of this stage is the creation of a highly controlled environment to simplify subsequent analysis. This often involved using uniform, high-contrast backgrounds, such as black velvet or specialized blue conveyor belts, to make the brighter grain kernels easier to distinguish (Szczypiński and Zapotoczny 2012). Illumination was also carefully managed, with diffuse, uniform lighting sources like circular fluorescent lamps or smoked magnesium oxide domes employed to minimize shadows and specular reflections that could interfere with feature measurement (Wang et al. 2003). This deliberate engineering of the imaging environment to remove visual ambiguity is a hallmark of the classical MV philosophy. Rather than building algorithms to handle complexity, the environment was simplified so that simpler algorithms could succeed. This approach, while effective in the lab, reveals a fundamental limitation in its scalability to less controlled, real-world industrial settings.

Following acquisition, images undergo pre-processing and segmentation (Khan et al. 2022). Common pre-processing steps include converting color images from the RGB space to grayscale for morphological analysis or to other color spaces like HSI (Hue, Saturation, Intensity) for color feature extraction (Zaw et al. 2023). Noise removal, often using median filters, is applied to clean the image (Ramesh et al. 2021). The primary goal of this stage is to prepare the image for segmentation, the process of isolating each individual grain kernel as a distinct object (Shi et al. 2022). In the controlled environments typical of early MV systems, this could often be achieved with simple global thresholding techniques, such as Otsu's method, which automatically finds an optimal threshold to separate the bright kernels from the dark background (Shen et al. 2023). The resulting binary mask is then refined using morphological operations like opening and closing to remove small artifacts and fill holes within the kernel regions, ensuring a clean and accurate representation of each grain's silhouette.

Handcrafted feature extraction: the core of classical MV

The intellectual core of the classical MV paradigm lies in the feature extraction stage. Here, domain experts design algorithms to compute a set of numerical descriptors, or features, that are believed to be effective for discriminating between different grain types or quality grades. These features are typically categorized into three main groups: morphological, color, and texture.

Morphological features

Morphological features quantify the shape and size of the grain kernels, derived from their segmented binary silhouettes. These are among the most intuitive and widely used descriptors, as many cereal types have distinct and characteristic shapes. Pioneering work by Majumdar and Jayas (Majumdar and Jayas 2000) systematically explored the utility of 23 different morphological features for classifying five major Canadian cereal grains. As detailed in Table 1, these features

Table 1. Common morphological features used for grain kernel classification, including basic size measurements and shape descriptors.

Feature Category	Specific Feature Example	Description	Typical Application in Grains
Morphological	Area	Total number of pixels within the kernel boundary.	Distinguishing grain size; estimating kernel weight.
	Aspect Ratio	Ratio of the major axis length to the minor axis length.	Differentiating elongated grains (e.g., oats, rye) from rounder grains (e.g., wheat).
	Circularity	A measure of how closely the shape of the kernel resembles a perfect circle.	Identifying broken or irregularly shaped kernels.
	Fourier Descriptors	Mathematical descriptors of the kernel's boundary shape in the frequency domain.	Capturing complex shape details for fine-grained varietal classification.
Color	Mean Hue (HSI Space)	The average hue value of all pixels in the kernel, representing the dominant color.	Detecting discoloration from fungal damage or identifying different wheat classes (red vs. white).
	Saturation Variance	The statistical variance of the saturation values, indicating the purity of the color.	Assessing vitreousness in durum wheat, where non-vitreous kernels appear less saturated.
	Mean Red (RGB Space)	The average intensity of the red channel across all kernel pixels.	Classifying red wheat varieties; detecting reddish discoloration from certain diseases.
Texture	GLCM Contrast	Measures the local intensity variations in the image; high contrast indicates a rough surface.	Detecting cracks, insect damage, or rough surface textures.
	GLCM Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the diagonal.	Quantifying the smoothness of the kernel surface.
	LRE Moment (RLM)	Long Runs Emphasis moment from the Run-Length Matrix; higher for smooth areas.	Identifying and quantifying wrinkled regions on barley kernels for malting quality assessment.

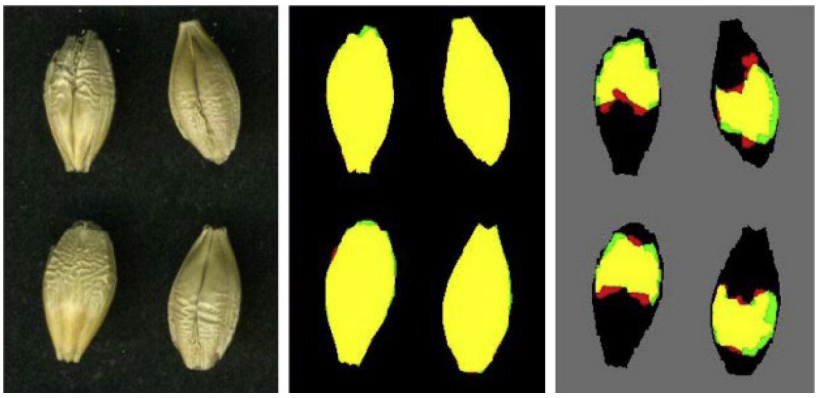

Figure 2. Example of morphological feature extraction from a barley kernel. Source: Figure by Szczypiński and Zapotoczny (Szczypiński and Zapotoczny 2012).

include basic measurements like area (the number of pixels in the kernel), perimeter, major and minor axis lengths (approximating length and width), and aspect ratio (the ratio of major to minor axis length) (Douik and Abdellaoui 2010). More complex shape descriptors such as circularity and compactness, which measure how closely the grain's shape resembles a perfect circle, were also employed. These features proved highly effective; using just the 10 most significant morphological features, their model achieved classification accuracies of 99.9% for oats and 98.9% for Canada Western Red Spring (CWRS) wheat, demonstrating the power of shape for distinguishing between morphologically distinct grains. Fig. 2 provides a visual example of how such features are derived from a kernel image.

Color features

Color features provide crucial information related to the chemical composition and condition of the grain surface (Tan et al. 2022). Algorithms for color feature extraction typically begin by analyzing the pixel intensity values across different channels of a color space (Fidan and Kaya 2020; Guan-shi et al. 2022). While the default RGB (Red, Green, Blue) space is commonly used, many studies transform the image into perceptual color spaces like HSI or LAB, which separate color information (hue and saturation) from brightness (intensity or lightness) (Fu et al. 2020). This separation is advantageous because hue and saturation are often less sensitive to minor variations in illumination than raw RGB values. The features extracted are typically statistical measures, such as the mean, variance, and range of pixel values within the kernel region for each color channel (Liu et al. 2024). These features have been instrumental in assessing quality attributes that manifest as color changes (Hassanisaadi et al. 2021; Gurău et al. 2023). For instance, a machine vision system using HSI features was developed to determine the vitreousness of durum wheat, a key quality parameter related to protein content that affects its suitability for pasta production. Similarly, color features have been used to detect fungal infections or discoloration caused by overheating, which are difficult to capture with morphology alone (Bao et al. 2023).

Texture features

Texture features describe the spatial arrangement of pixel intensities on the grain's surface, capturing properties like smoothness, coarseness, and regularity. These features are particularly valuable for identifying surface defects such as cracks, wrinkling, or the presence of fungal mycelia. Two of the most common methods for texture feature extraction are the Gray-Level Co-occurrence Matrix (GLCM) and the Run-Length Matrix (RLM). The GLCM characterizes texture by calculating how often pairs of pixels with specific brightness values and in a specified spatial relationship occur in an image. From this matrix, statistical measures like contrast, correlation, energy, and homogeneity are derived. The RLM, used effectively by Szczypiński and Zapotoczny (Szczypiński and Zapotoczny 2012) to detect wrinkled regions on barley kernels, instead counts runs of consecutive pixels with similar gray-level values. The lengths and prevalence of these runs are used to compute texture descriptors. For example, a smooth surface will have long runs of similar pixels, while a rough or wrinkled surface will have many short runs.

The systematic evaluation of these feature sets, both individually and in combination, represents a key contribution of the classical MV era. The work of Majumdar and Jayas (Majumdar and Jayas 2000) is again exemplary. After evaluating models based on morphology, color, and texture independently, they developed combined models. Their results consistently demonstrated a principle of information fusion: that complex biological objects are best characterized by multiple modalities. The combined morphology-color-texture model achieved the highest mean classification accuracy of 99.7%, outperforming any single-feature model. This iterative process of feature combination is a microcosm of the broader evolution in pattern recognition, showing that integrating diverse sources of information leads to more robust and accurate systems. This principle remains highly relevant today in the context of multi-modal deep learning.

The efficacy and inflexibility of handcrafted features

The handcrafted feature engineering paradigm undeniably achieved significant success, proving that

automated, objective grain inspection was not only feasible but could also be highly accurate for well-defined problems. These systems were computationally efficient compared to modern deep learning models and offered a high degree of interpretability, as each feature corresponded to a physically meaningful property of the grain. However, this paradigm was built on a foundation that was ultimately too rigid to support broad, real-world application.

The core argument against the handcrafted feature approach is its fundamental lack of adaptability. These systems are "brittle" because their performance is inextricably tied to the specific features chosen by a human designer (Gupta and Waoo 2024). If a new grain variety is introduced, or an unforeseen type of defect appears, or imaging conditions shift slightly, the pre-defined features may no longer be optimal or even relevant, causing performance to degrade catastrophically. The process of designing and selecting features is itself a laborious, trial-and-error process that requires deep domain expertise. This dependency on a priori knowledge means the system can only identify what it has been explicitly programmed to look for. Furthermore, as noted by Szczypiński and Zapotoczny (Szczypiński and Zapotoczny 2012), the correlation between these carefully engineered visual features and the actual end-use technological properties of the grain, such as the quality of flour produced, can often be unsatisfactory. This "semantic gap" between what can be easily measured and what truly matters represents a profound limitation. It was this inherent inflexibility and the challenge of bridging the semantic gap that created the imperative for a new approach—one where features are not designed by humans but learned from data.

The deep learning revolution: CNNs in grain analysis

The limitations of the handcrafted feature paradigm—its rigidity, labor-intensive design process, and difficulty in capturing complex visual patterns—precipitated a fundamental shift in the field of computer vision, leading to the rise of deep learning. CNNs, in particular, have emerged as the dominant architecture for image analysis tasks, revolutionizing cereal grain quality assessment by replacing manual feature engineering with automated, hierarchical feature learning. This end-to-end approach has unlocked new levels of accuracy and robustness, enabling the development of systems capable of tackling far more complex and nuanced quality assessment challenges.

Automated feature learning

Unlike traditional machine vision systems that require a separate, human-designed feature extraction step, CNNs integrate feature extraction and classification into a single, unified learning framework. The architecture of a CNN is inspired by the human visual cortex and is specifically designed to process pixel data. As depicted in Fig. 3, a typical CNN consists of a series of interconnected layers that progressively transform an input image into a final output, such as a class label.

The core building blocks of a CNN are the convolutional layers. Each convolutional layer contains a set of learnable filters, or kernels, which are small matrices of weights. During the forward pass, each filter is convolved—slid across the input image or the feature map from the previous layer—computing a dot product at each

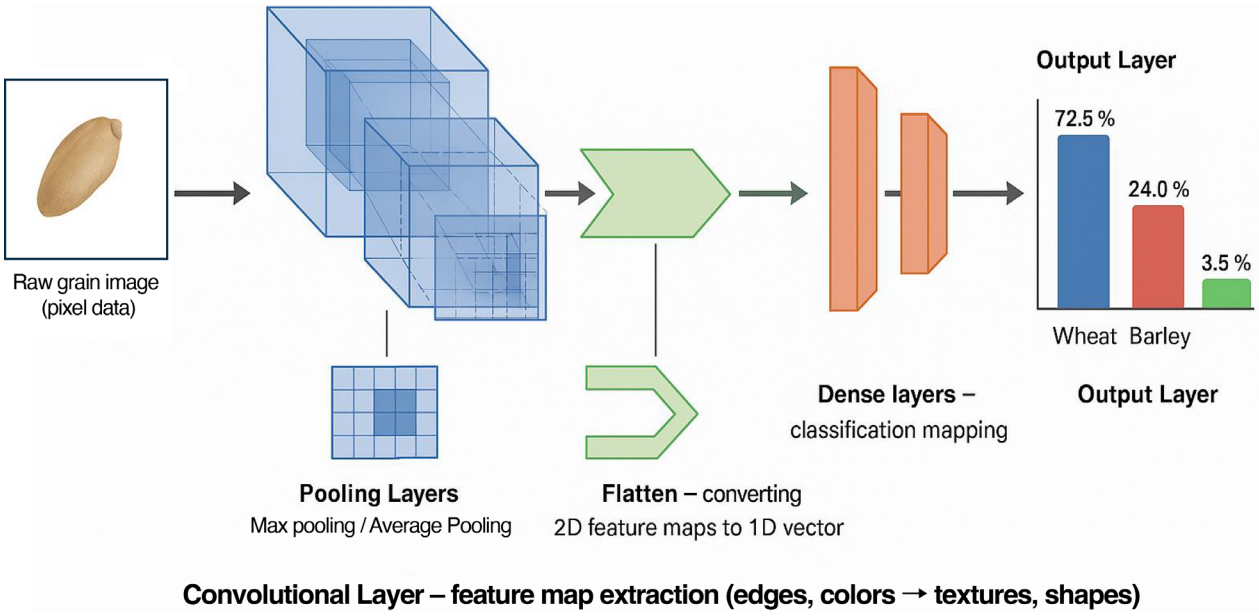

Figure 3. A generic CNN architecture. The schematic demonstrates the hierarchical transformation of raw pixel data through convolutional and pooling layers into high-level feature maps, culminating in dense layers for probability-based classification.

location (Benos et al. 2021). This operation effectively creates a new feature map that highlights where a specific pattern (e.g., a horizontal edge, a particular color, a curved shape) is present in the input. The power of this approach lies in the fact that the filters' weights are not predefined; they are learned automatically during the training process through backpropagation (Meghraoui et al. 2024). The network learns which features are most discriminative for the given task directly from the data.

This process is hierarchical. The first convolutional layer might learn to detect simple features like edges and color gradients. The feature maps from this layer are then fed into the next convolutional layer, which learns to combine these simple features into more complex patterns, such as textures or simple shapes like circles and corners. Deeper layers in the network learn to combine these mid-level features into even more abstract and complex representations, such as the overall shape of a specific grain variety or the characteristic pattern of a particular type of fungal damage (Lingwal et al. 2021). This automatic learning of a feature hierarchy is the key innovation of CNNs, freeing researchers from the constraints of manual feature design. Furthermore, the architecture possesses inherent properties that make it exceptionally well-suited for image data. The use of shared weights in the convolutional filters makes the network highly parameter-efficient, and the combination of convolution and pooling operations provides a degree of invariance to translation and small scale variations. However, unlike translation, rotation invariance is not inherent to standard CNN architectures. In the context of grain inspection, where kernels appear in random orientations, this capability is typically induced through data augmentation strategies, such as random rotations during training, rather than being an intrinsic property of the convolution operation itself.

Applications of CNNs in grain quality assessment

The adoption of CNNs has led to significaements across a wide spectrum of grain quality assessment tasks, consistently outperforming traditional MV methods. To make these advances easier to navigate and to enable meaningful within-task comparison, we group CNN-based work into three major application classes: varietal classification (Section 4.2.1), defect and damage inspection (Section 4.2.2), and object detection for counting/sorting (Section 4.2.3). In addition to organizing the narrative, this task-oriented structure supports a comparative synthesis in which we summarize typical reported performance ranges within each task and interpret those values in light of dataset scale and difficulty.

Varietal classification

One of the most prominent applications has been in the classification of cereal varieties, a task that is often challenging due to subtle morphological similarities. Numerous studies have demonstrated the superior capability of CNNs in this domain. For instance, in the classification of barley varieties, where traditional methods struggled to exceed 75% accuracy, CNN models achieved accuracies of over 93% (Kozłowski et al. 2019). Researchers have explored a range of architectures, with models like AlexNet, GoogleNet, and VGG19 often showing strong performance (Ma et al. 2016). While model fusion—combining the outputs of multiple CNNs—has demonstrated accuracy improvements of up to 7.2% in cotton cultivar classification, this approach inherently multiplies the computational overhead and memory footprint, making it less viable for high-speed industrial sorting. In direct contrast, transfer learning offers a more computationally elegant and data-efficient solution. By fine-tuning a single CNN pre-trained on a massive dataset like ImageNet onto a specific grain dataset, researchers bypass the exhaustive feature-learning phase required by fusion models. For example, Loddo and Di Ruberto (Loddo and Di Ruberto 2021) demonstrated in seed species classification that while ensemble networks achieve marginal accuracy gains, transfer learning utilizing architectures like ResNet50 provides a superior practical trade-off, achieving comparable precision with a fraction of the inference latency. This critical distinction highlights that for real-time agricultural applications, optimizing a single pre-trained network often presents a more scalable solution than concatenating multiple bespoke models. This leverages the powerful, generalized features learned from millions of images and has been shown to significantly boost classification performance, especially when labeled agricultural data is limited. The broader applicability and robustness of this approach across agricultural domains are well-supported by recent advances. For instance, recent applications of transfer learning in sunflower disease detection, as well as comparative studies on papaya leaf disease classification using pre-trained deep learning models (such as the AI-driven PapNet architecture), demonstrate that fine-tuning pre-trained weights drastically reduces training time while maintaining high diagnostic accuracy (Ayoub et al. 2025; Haque et al. 2025). These cross-domain successes provide strong evidence that similar pre-trained architectures can be seamlessly and effectively adapted to capture the nuanced visual features of cereal grains.

Defect and damage detection

CNNs have proven to be exceptionally adept at identifying a diverse array of grain defects, from gross physical damage to subtle signs of disease. Their ability to learn complex and nuanced features allows them to detect defects that are often difficult to define with explicit rules. For example, custom CNNs have been developed to classify wheat kernels into categories such as 'Good grain', 'Damaged Grain', 'Broken Grain', and 'Foreign Particles' with accuracies reaching 95% (Shedole et al. 2022). These models have successfully identified wheat grains damaged by sunn pests, a task requiring the detection of subtle surface marks and deformities. Similarly, CNNs have been

combined with hyperspectral imaging (HSI) to classify wheat seeds damaged by fungi (Qi et al. 2025) and to detect rice diseases from leaf images with accuracies as high as 98% (Ahad et al. 2023). These applications underscore the power of data-driven feature learning to capture the complex visual signatures of biological stress and damage.

Object detection for counting and sorting

Beyond simple classification, more advanced CNN-based frameworks like You Only Look Once (YOLO) and Faster R-CNN perform object detection, which involves both classifying and localizing objects within an image (van Vliet et al. 2025). The period between 2023 and 2025 has seen a significant surge in the application of these models for complex phenotyping and counting tasks. This capability is crucial for applications that require counting or singulating kernels. For instance, YOLOv8 has been used to separate individual wheat grains from each other in dense images before classifying them as healthy or damaged. Recent 2024 advancements have further optimized these architectures for cluttered, real-world agricultural environments; notably, improved YOLOv8 models incorporating attention mechanisms have successfully achieved high-precision wheat seed detection even in densely stacked scenes with impurities. Object detection models are also the core technology for automated sorting systems, where they can identify the location of defective grains or foreign materials on a conveyor belt, enabling a robotic arm or air-jet ejector to remove them in real-time (Berkmanas 2024). From an engineering perspective, a deployable vision sorter is not only a detection model but a synchronized closed loop consisting of (i) image acquisition under controlled and shielded illumination, (ii) low-latency inference, (iii) object tracking (to preserve identity under partial occlusion and kernel clustering), and (iv) actuator triggering (air nozzles or robotic end-effectors) with precise spatial–temporal alignment to belt motion. Optical belt sorting studies emphasize that ejection accuracy depends on calibration of belt speed, nozzle timing windows, and the kinematic mapping between image coordinates and physical ejector positions; accordingly, inference latency must be considered together with sensor triggering and actuation delay rather than reported in isolation. This is consistent with engineering analyses of optical belt sorters where speed-adaptive parameterization (e.g., nozzle activation duration and ejection position) is required to maintain separation performance under variable belt speeds.

The "Model Zoo" and performance trade-offs

The rapid development of deep learning has resulted in a vast "model zoo" of different CNN architectures, each with its own set of strengths, weaknesses, and computational requirements. While deeper and more complex networks like ResNet and DenseNet often achieve state-of-the-art accuracy on academic benchmarks, a critical debate has emerged within the agricultural technology community regarding their practical utility (Ahad et al. 2023). The high computational cost and memory footprint of these large models present a significant barrier to their deployment in real-world agricultural settings, particularly on low-cost, resource-constrained devices for in-field analysis or high-speed, on-line sorting systems (Guo et al. 2025).

This practical constraint has fueled a strong research trend towards the development of lightweight architectures. This shift in focus marks a maturation of the field, moving from an initial phase of "proof-of-concept"—simply demonstrating that deep learning could solve the problem—to a phase of "practical implementation," where the goal is to create solutions that are economically and computationally viable for widespread adoption. Models like MobileNet and custom-designed lightweight networks are engineered to balance recognition performance with efficiency (Ahad et al. 2023). A compelling example is the LWheatNet model, which reports 98.59% accuracy for wheat seed classification while having a model size of only 1.33 MB, making it well-suited for deployment on mobile devices (Guo et al. 2025). Similarly, the WDD-Net was designed for wafer defect detection with a model size of just 307 KB and an inference speed of 105.6 frames per second, demonstrating the feasibility of real-time analysis on edge devices (Chen et al. 2020). However, headline accuracy (or mAP) values reported across different papers should not be interpreted as directly comparable, because they are conditioned on heterogeneous datasets, acquisition setups, class definitions, split strategies, and metric conventions; such dataset and protocol bias is well recognized in the broader vision literature. Therefore, Table 2 is intended as a structured catalog of representative tasks, model families, and reported outcomes within each study's experimental setting, rather than a unified benchmark ranking. When quantitative comparison is required, it should be performed on shared benchmarks or under harmonized evaluation protocols (same dataset and split, consistent metrics, and—if deployment is claimed—cross-domain and out-of-distribution tests).

Beyond convolutions: the emergence of vits for cereal grain assessment

While CNNs have become the workhorse of modern computer vision, achieving remarkable success in localized feature extraction, their core architectural principle imposes an inherent limitation regarding global context. CNNs rely on localized convolutional filters (kernels) that process information in small, restricted windows (e.g., 3x3 or 5x5 pixels). Although deeper layers theoretically expand the effective receptive field, empirical studies suggest that the effective receptive field often remains much smaller than the theoretical maximum, limiting the network's ability to capture long-range dependencies. This results in a phenomenon of "contextual blindness," where

Table 2. Summary of representative CNN architectures applied in cereal grain quality assessment, detailing target cereal types, primary tasks, performance outcomes, and notable strengths or weaknesses as reported in the literature.

CNN Architecture	Cereal Type	Assessment Task	Reported Accuracy / mAP	Key Strengths / Weaknesses
AlexNet, GoogleNet, VGG19	Cotton Cultivars	Varietal Classification	>90% (Fused models)	High accuracy, foundational deep architectures. Computationally heavy.
Custom CNN	Wheat	Quality Classification (Good, Damaged, Broken, Foreign)	95%	Tailored architecture for specific task, effective pre-processing.
Custom CNN	Barley	Varietal Classification	>93%	Significantly outperforms traditional MV methods (<75% accuracy).
DenseNet121, InceptionV3, MobileNetV2	Rice	Leaf Disease Classification	98% (Ensemble)	Ensemble of diverse architectures yields highest accuracy. MobileNetV2 is lightweight.
LWheatNet	Wheat	Varietal Classification	98.59%	Extremely lightweight (1.33 MB model size), suitable for mobile/edge devices.
YOLOv8	Wheat	Detection of Sunn Pest Damage	98.61% (on augmented data)	High accuracy in object detection for identifying and localizing damaged kernels.
ResNet50 + SVM	Rice	Nutrient Deficiency Detection	99.05%	Uses a pre-trained CNN as a feature extractor for a traditional classifier.
YOLOv8-HD (Improved)	Wheat	Seed Detection (in stacked scenes)	99.3% mAP	Incorporates Transformer components to improve feature extraction in complex scenes.

the model struggles to correlate distant visual cues. This is particularly detrimental in grain pathology, such as identifying early-stage fungal infections like *Fusarium* Head Blight (FHB), where the disease manifests not as a distinct edge or shape, but as a subtle, diffuse discoloration pattern distributed across the entire kernel surface. A CNN may correctly identify a small discolored patch but fail to recognize the global distribution pattern that confirms the disease, leading to misclassification. To overcome this inability to model explicit long-range relationships, the field has shifted toward an architecture specifically designed to capture global context through self-attention mechanisms: the Vision Transformer (ViT).

The ViTs architecture: a global perspective

The ViTs fundamentally re-imagines how an image is processed. Instead of applying localized filters, a ViT treats an image as a sequence of patches, analogous to how a language Transformer treats a sentence as a sequence of words. As illustrated in Figs 4, 5, the standard ViT pipeline begins by dividing an input image into a grid of non-overlapping patches. Each patch is flattened into a vector and linearly projected into an embedding space. Crucially, positional encodings are added to these patch embeddings to retain spatial information, as the subsequent processing is permutation-invariant (Patra et al. 2025).

The core of the ViTs is the Transformer encoder block, which is composed of two main sub-layers: a multi-head self-attention mechanism and a feed-forward multi-layer perceptron (MLP) (Patra et al. 2025). The self-attention mechanism is the key innovation. It allows the model to weigh the importance of all other patches in the sequence when processing a given patch. In essence, it computes a set of "attention scores" that determine how much focus to place on different parts of the image. This enables the model to capture global dependencies and intricate spatial relationships, directly addressing the limited receptive field of CNNs. By stacking multiple Transformer encoder blocks, the model can learn increasingly complex and abstract relationships between different parts of the grain image.

CNNs vs. transformers

The introduction of ViTs has ignited a vigorous debate about the optimal architectural foundation for computer vision tasks, including grain quality assessment. Each architecture embodies a different set of assumptions and trade-offs.

The core argument for Transformers is their superior ability to model global context. This is particularly relevant for diagnosing systemic health issues in grains that may not have a single, localized tell-tale sign. For example, early-stage drought stress or a systemic fungal infection like Fusarium head blight (FHB) might manifest as very subtle, distributed changes in color and texture across the entire kernel surface. A CNN, with its local focus, might struggle to aggregate these weak, spatially distant signals into a coherent feature representing "disease." A Transformer, by contrast, is architecturally designed to find these long-range dependencies, making it potentially more powerful for such diagnostic tasks. Evidence from recent studies supports this view, revealing distinct performance divergences between these architectures on complex diagnostic tasks. In a study on corn quality classification, a model based on the Swin Transformer achieved 99.89% accuracy, significantly outperforming baseline CNNs. Rather than simply extracting localized texture features—such as isolated mold patches, which CNNs handle proficiently—the Swin Transformer's self-attention mechanism successfully mapped the spatial distribution of these defects across the entire kernel

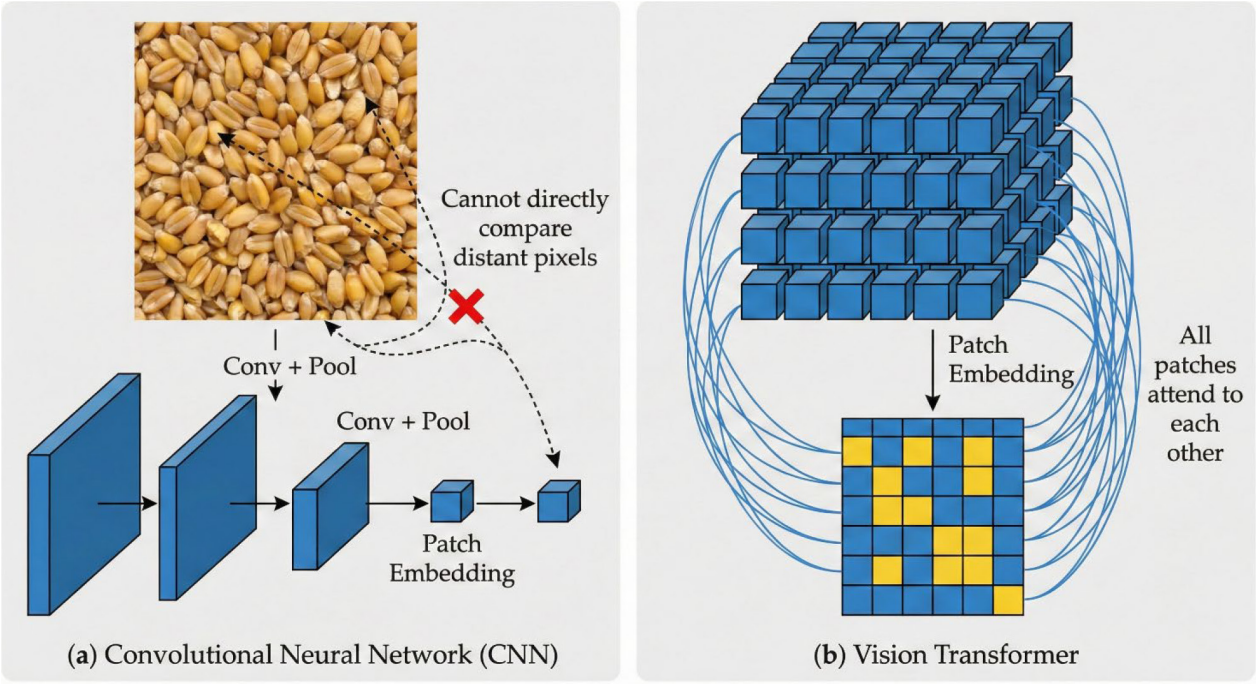

Figure 4. Comparison of processing mechanisms in CNNs versus ViTs. **a.** CNNs utilize hierarchical convolutional filters with local receptive fields, whereas; **b.** ViTs employ a self-attention mechanism to allow simultaneous global interaction between all image patches.

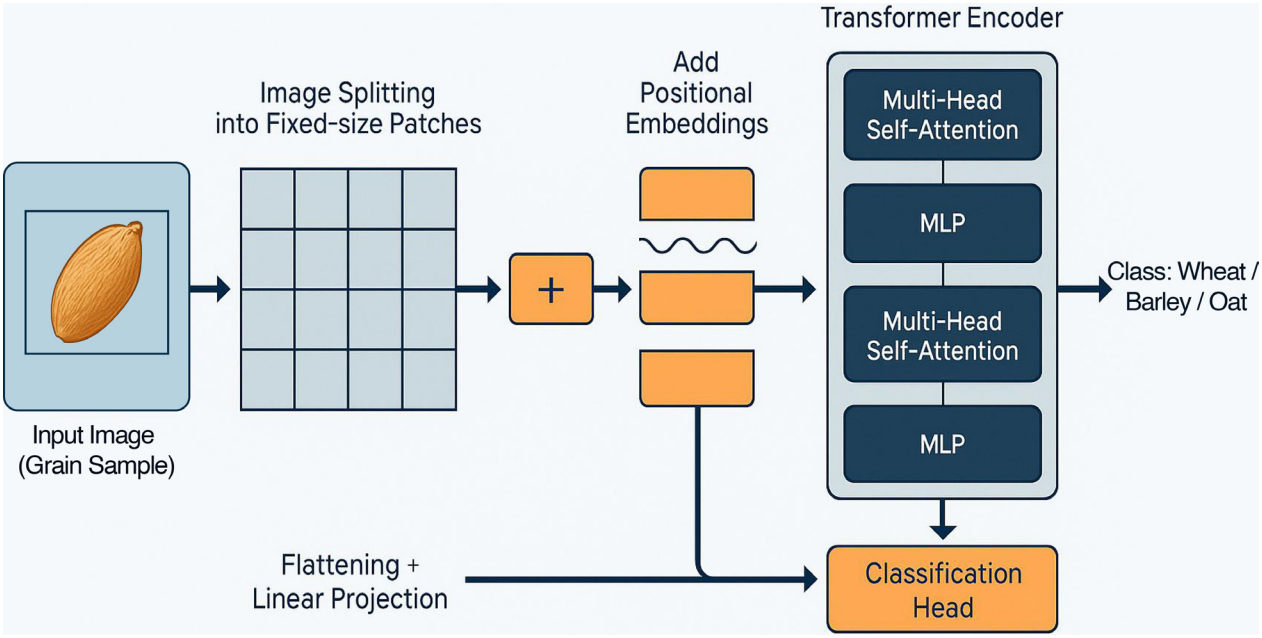

Figure 5. The standard ViT architecture. The diagram depicts the data flow where the input image is split into patches, embedded with positional information, and processed through stacked transformer encoder blocks containing multi-head self-attention to generate a classification output.

geometry. This global contextual awareness directly addresses the "contextual blindness" of CNNs described earlier. This mechanistic advantage is further corroborated by Bi et al. (Bi et al. 2022), who observed that while standard ResNet architectures degraded in performance when identifying structurally deformed seeds due to overlapping kernel edges, Transformer-based models maintained high precision. They achieved this by correlating distant visual cues to distinguish true biological deformation from artificial occlusion. Thus, Transformers do not merely offer incremental accuracy improvements; they provide a fundamentally different, globally aware feature representation that is uniquely suited for diagnosing diffuse biological stress.

Conversely, the primary argument for the continued relevance of CNNs lies in their strong and highly effective "inductive bias" (Han et al. 2022; Wang et al. 2023). CNNs

are built with the inherent assumptions of locality (pixels close together are related) (Bagherzadeh et al. 2022) and translation equivariance (a feature is the same regardless of where it appears in the image) (Jiang et al. 2022). These assumptions are extremely well-suited for natural images and make CNNs remarkably data-efficient (Krichen 2023). Transformers, lacking this built-in bias, must learn these properties from scratch, which typically requires enormous amounts of training data. This "data-hungry" nature is a major drawback in specialized domains like agriculture, where large, labeled datasets are often scarce (Rezaei et al. 2024). Furthermore, the quadratic complexity of the self-attention mechanism with respect to the number of patches makes ViTs computationally expensive (Yuan et al. 2022), posing a barrier to practical deployment (Hou et al. 2024).

This debate is not merely academic; it has profound implications for the future of agricultural technology. The current research landscape suggests that the immediate future may not involve a wholesale replacement of one architecture by the other, but rather a period of hybridization. Researchers are actively developing models that combine the strengths of both paradigms. This strategy leverages the data-efficient local feature extraction capabilities of CNNs as a front-end to generate rich feature maps, which are then fed into a transformer back-end to model the global context between these learned features. For example, one study improved the YOLOv8 object detector by incorporating a ViTs with a deformable attention mechanism into its CNN backbone, boosting detection accuracy for wheat seeds in cluttered scenes by 16.8% (Ma et al. 2024). Another proposed an "enhanced transformer context bridge" to connect a CNN-based encoder and decoder for FHB segmentation. This hybrid approach offers a pragmatic and powerful path forward, balancing the proven efficiency of convolutions with the global modeling power of self-attention. To provide a clearer comparison between the two paradigms, Table 3 summarizes the key architectural attributes of CNNs and ViTs, along with their practical implications for grain quality assessment. The table highlights how differences in receptive field, inductive bias, feature extraction mechanisms, and computational cost influence their respective strengths and weaknesses in detecting localized versus diffuse grain defects. It also underscores the trade-off between CNNs' proven data efficiency and ViTs' capacity for modeling global context, offering guidance for selecting or designing architectures tailored to specific agricultural imaging challenges.

Advanced transformer applications in cereal grains

Beyond classification, the unique capabilities of Transformers are being explored for other advanced tasks in grain analysis, representing a major research frontier in 2024 and 2025. The Swin transformer, which introduces a hierarchical structure and uses self-attention within non-overlapping, shifted windows, has emerged as a more efficient and scalable alternative to the original ViT (Tian et al. 2024). Recent 2025 studies have demonstrated its state-of-the-art performance in complex evaluations, such as high-precision maize quality detection, heavily outperforming traditional convolutional networks. Similarly, the application of hybrid Vision Transformer models has proven highly effective in recent 2025 literature for advanced rice grain classification (Ahmad et al. 2025), offering superior precision and recall by merging local feature extraction with global contextual awareness. In the realm of image enhancement, a novel 2024 network called T-transformer net has been used for super-resolution of low-quality wheat grain images (Zhang et al. 2025).

Synthesis of challenges, innovations, and future research directions

The progression from manual inspection to handcrafted machine vision and now to end-to-end deep learning represents a remarkable technological journey. However, despite the significant advances in accuracy and capability,

Table 3. Comparative attributes of CNN and ViTs architectures and their implications for grain quality assessment.

Attribute	CNN Approach	ViTs Approach	Implication for Grain Quality Assessment
Receptive Field	Local; builds hierarchically.	Global; all patches attend to all others in one layer.	ViTs may be superior for detecting diffuse, non-local defects (e.g., systemic disease), while CNNs excel at localized defects (e.g., a single crack).
Inductive Bias	Strong (locality, translation equivariance).	Weak; must learn visual properties from data.	CNNs are highly data-efficient and learn quickly on smaller datasets. ViTs are "data-hungry" and require large datasets to perform well.
Feature Extraction	Implicitly learns hierarchical features from simple to complex.	Models relationships between patch embeddings.	CNNs are naturally suited for learning visual hierarchies. ViTs are more flexible but less structured without massive data.
Computational Cost	Generally lower, scales linearly with image size.	High, scales quadratically with the number of patches (image size).	CNNs are better suited for resource-constrained edge devices and real-time sorting. ViTs are computationally expensive.
Key Strength	Data efficiency and excellent performance on a wide range of visual tasks.	Modeling long-range dependencies and global context.	CNNs are the robust, proven workhorse. ViTs offer a new capability for problems requiring a holistic view of the grain.

several cross-cutting challenges persist that hinder the widespread, robust deployment of these systems in real-world agricultural and industrial environments. Addressing these challenges through targeted innovation is the central focus of current research and will define the future trajectory of the field. The ultimate goal is to move beyond isolated laboratory demonstrations and create integrated, intelligent ecosystems for grain quality assessment that are accessible, reliable, and transparent.

The data bottleneck

The single greatest impediment to the development and validation of advanced deep learning models for agriculture has historically been the scarcity of large, high-quality, publicly available, and well-annotated datasets (Guo et al. 2025). Beyond mere data volume, the inherent quality and resolution of the dataset are paramount. Recent investigations into the impact of dataset quality on deep learning models for dragon fruit and leaf health classification underscore that label noise, inconsistent lighting, and low-resolution inputs can severely degrade a model's feature extraction capabilities (Zangana et al. 2025). In the context of grain inspection, ensuring rigorous annotation protocols and high-fidelity imaging is similarly critical to preventing the network from learning artifactual noise rather than genuine morphological or color-based quality attributes. This challenge has transitioned from a hurdle to a critical bottleneck with the advent of Vision Transformers. As detailed in Section 4.2, unlike CNNs which possess a strong inductive bias towards locality and translation invariance, ViTs must learn these visual properties from scratch (Bishaw et al. 2025). Consequently, ViTs require significantly larger volumes of data to converge than their CNN counterparts. Early research relying on small, privately collected datasets was sufficient for lighter CNNs but proved inadequate for unlocking the global context modeling capabilities of Transformers. This architectural dependency on massive data underscores the importance of recent innovations such as GrainSpace, a repository containing 5.25 million images, which effectively bridges the gap between the "data-hungry" nature of modern algorithms and the practical realities of agricultural data collection.

This data bottleneck has begun to be addressed by a transformative innovation: the creation and release of large-scale public datasets specifically for grain analysis. The most significant of these is GrainSpace, a massive repository containing 5.25 million images of wheat, maize, and rice, collected from multiple countries and captured using different devices to ensure diversity (Fan et al. 2022). Datasets like GrainSpace, along with others such as PaddyDoctor and various repositories on platforms like Kaggle (summarized in Table 4), are game-changers (Pabamalie and Premaratne 2010). Their availability does more than just provide training data; it fundamentally reshapes the research methodology of the entire field. It enables a shift from bespoke, difficult-to-compare studies to a standardized, benchmark-driven approach. This fosters reproducibility, facilitates fair and direct comparison between novel algorithms, and lowers the barrier to entry for researchers, thereby accelerating the pace of collective innovation.

From lab to line - the generalization gap

A persistent and critical challenge is the "generalization gap" between model performance in controlled laboratory settings and in dynamic industrial environments. Models trained on perfectly illuminated, non-overlapping kernels often fail when deployed on a high-speed conveyor belt where grains are clumped together, partially occluded, and subject to variable lighting (Ma et al. 2016). Furthermore, standard resizing operations required by many CNN architectures can obscure critical, fine-grained morphological details. Insights from recent efforts dedicated

Table 4. Summary of major publicly available datasets for cereal grain and plant image analysis, including cereal types, scale, image characteristics, annotated quality attributes, availability, and references.

Dataset Name	Cereal Type(s)	Number of Images / Classes	Image Type	Key Quality Attributes Annotated	Availability/ Link	Reference
GrainSpace	Wheat, Maize, Rice	5.25 million images / Multiple fine-grained classes	RGB	Damaged, unsound (e.g., fusarium, moldy, pest-damaged), broken, foreign material.	GitHub	(Fan et al. 2022)
Rice Grains Image Dataset	Rice	75,000 images / 5 varieties	RGB	Varietal classification (Arborio, Basmati, Ipsala, Jasmine, Karacadag).	Kaggle	(Pabamalie and Premaratne 2010)
PaddyDoctor	Rice	10,407 images / 10 classes	RGB	Healthy and nine types of disease from leaf images.	Publicly available	(Xu et al. 2024)
Wheat Grains - 15 Varieties	Wheat	15,000 images / 15 varieties	RGB	Varietal classification.	Custom dataset, availability not specified.	(Lingwal et al. 2021)
Wheat Leaf Dataset	Wheat	407 images / 3 classes	RGB	Healthy, Septoria, Stripe Rust from leaf images.	Publicly available	(Xu et al. 2024)
Bangladesh Rice Varieties	Rice	4,730 original images (23,650 augmented) / 20 varieties	RGB	Varietal classification of local Bangladeshi strains.	Publicly available	(Islam et al. 2024)

to mitigating spatial scale loss in CNN-based fine-grained image classification for date fruit grading highlight that preserving aspect ratios and employing scale-aware feature extraction are vital for distinguishing subtle grade differences (Gulzar 2025a). This discrepancy between training and deployment conditions, compounded by such spatial scaling issues, remains a major barrier to practical adoption (Gulzar 2025b). This discrepancy between training and deployment conditions is a major barrier to practical adoption. To strengthen the analysis beyond qualitative statements, this review connects this gap to benchmark evidence from large-scale datasets designed explicitly to expose domain shift. GrainSpace formulates grain appearance inspection as fine-grained recognition, domain adaptation, and out-of-distribution recognition and reports systematic performance degradation across acquisition devices and domains, illustrating that real deployment variability can dominate error modes even when large numbers of images are available (Fan et al. 2022).

Innovations to bridge this gap focus on building more robust and adaptable models. Data augmentation is a key strategy, where the training dataset is artificially expanded by applying random transformations like rotations, scaling, brightness changes, and even simulated occlusions to teach the model to be invariant to these variations (Supreetha et al. 2024). Furthermore, researchers are designing models specifically to handle these real-world complexities. For instance, recent improvements to the YOLOv8 model have focused on enhancing its ability to detect individual wheat seeds in densely stacked scenes with impurities, a scenario that closely mimics industrial conditions (Ma et al. 2024).

Computational constraints and model interpretability

For many applications, particularly real-time sorting and in-field analysis on mobile devices, computational efficiency is as important as accuracy (Guo et al. 2025). However, the architectural evolution described in this review presents a conflict with this requirement. As noted in Table 3, the self-attention mechanism in ViTs scales quadratically with image resolution, creating a high computational overhead that challenges edge deployment. This necessitates the explicit development of lightweight architectures, such as the LWheatNet and MobileNet variants discussed in Section 3.3, which are engineered to balance the trade-off between the depth required for feature extraction and the latency constraints of industrial sorting lines. The trend towards increasingly deep and complex models runs counter to these practical needs. In parallel, the "black box" nature of deep learning models presents a significant barrier to trust and adoption. End-users, such as grain inspectors and farmers, are often hesitant to rely on a system whose decision-making process is opaque In engineering terms, the relevant question is whether end-to-end latency (acquisition + preprocessing + inference + post-processing + communication to the actuator controller) fits within the belt-motion time budget between imaging and the physical reject point. Recent online grain-processing quality inspection work illustrates the type of reporting needed for deployment claims; for example, an improved YOLOv8n-based system for rice processing quality reported real-time performance with an inference speed on the order of tens of frames per second and a compact model size in the single-digit megabyte range, supporting feasibility on edge-oriented compute without sacrificing detection accuracy. Such studies demonstrate that model size and FPS/latency can be presented together with task-specific quality metrics, enabling engineers to translate algorithmic performance into throughput planning and actuator timing margins (Shuai et al. 2024).

The innovations addressing these challenges are twofold. First, the development of lightweight architectures, as discussed in Section 3, is a direct response to the need for computational efficiency, enabling high-performance deep learning on resource-constrained hardware (Guo et al. 2025). Second, the emerging field of explainable AI (XAI) aims to make model predictions transparent and interpretable. Techniques like visualizing the attention maps of a ViTs can show which parts of a grain image the model focused on to make its classification, providing a visual rationale for its decision (Patra et al. 2025). Other methods, such as SHapley Additive exPlanations (SHAP), can quantify the contribution of different input features to the final prediction (Shuai et al. 2024). Incorporating XAI is crucial for building trust and facilitating the adoption of AI-driven systems in the agricultural sector.

Future research directions and frontiers

Looking forward, the field is poised for significant advancements driven by the integration of multiple technologies and the pursuit of more sophisticated AI capabilities. The future of grain quality assessment lies not in a single sensor or algorithm, but in a distributed, intelligent ecosystem composed of complementary technologies.

A primary frontier is multi-modal data fusion. Combining data from different sensing technologies can provide a more holistic and accurate assessment than any single modality alone. The fusion of standard RGB imaging, which captures spatial and morphological information, with HSI, which captures detailed spectral information related to chemical composition, is particularly promising. Studies have consistently shown that models trained on fused RGB-HSI data outperform models trained on either data type alone for tasks like classifying rice varieties or assessing milling quality. This approach allows for the simultaneous evaluation of physical attributes (shape, size, defects) and internal quality parameters (protein, moisture, starch content). Fig. 6 illustrates a conceptual framework for such a system.

Another critical research direction is developing models that can handle novelty and data scarcity. Current supervised learning models can only identify the specific

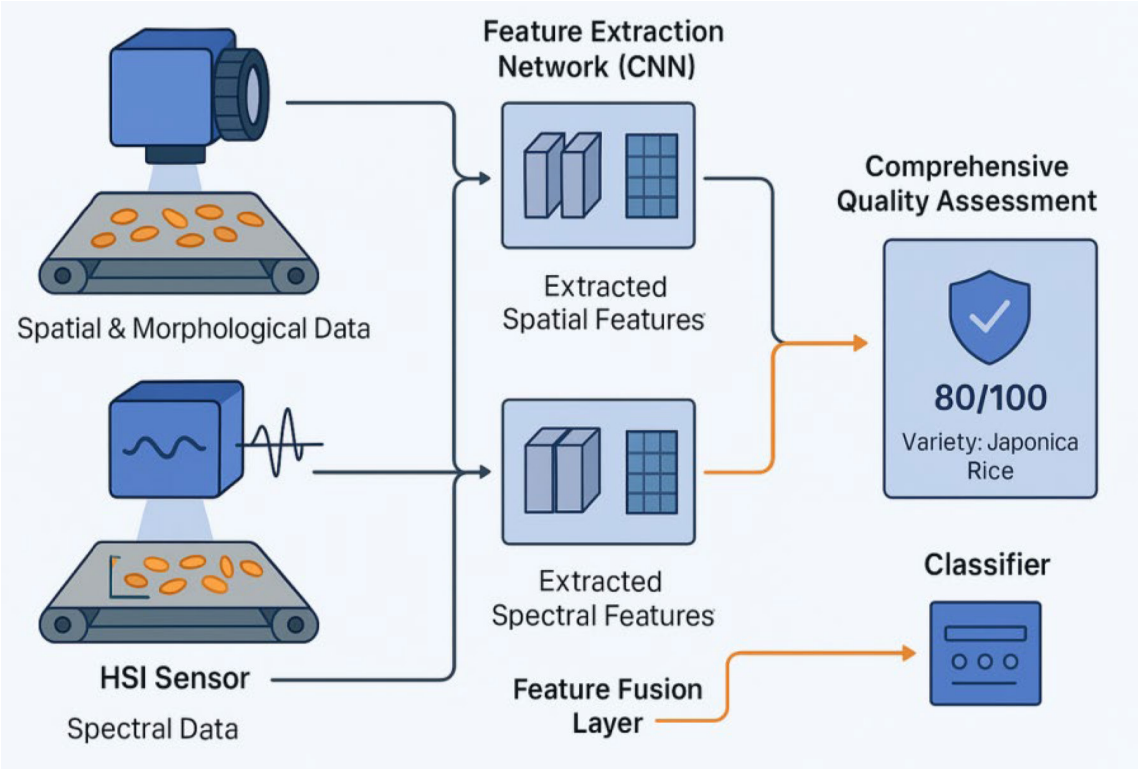

Figure 6. Multi-modal data fusion framework. The system integrates spatial data from RGB imaging and spectral data from HSI via distinct feature extraction networks, concatenating them in a fusion layer for comprehensive physico-chemical quality assessment.

classes of grains or defects they were trained on. They are brittle when faced with a new, unseen disease or a rare type of foreign material. To overcome these limitations, the field is beginning to explore advanced generative techniques. As highlighted in a recent systematic review of methods, applications, and future prospects of diffusion models for agricultural imaging, these architectures possess immense potential for generating highly realistic, synthetic agricultural data (Gulzar 2025c). By leveraging diffusion models to artificially synthesize high-fidelity images of rare grain defects or novel disease manifestations, researchers can robustly augment training datasets. This generative approach significantly enhances the model's capacity to handle edge cases and unseen environments without requiring exhaustive and costly manual data collection. Future systems must be able to generalize beyond their training data. This has led to research into generalized zero-shot and few-shot learning. These advanced machine learning paradigms aim to build models that can recognize novel classes without requiring extensive retraining. For example, frameworks based on a Food Safety Knowledge Graph (FSKG) have shown the ability to detect new, unseen food safety hazards with high accuracy by learning a generalizable relationship between visual attributes and hazard categories (Ma et al. 2016).

Finally, the ultimate goal is the creation of fully autonomous inspection and sorting systems. This involves the seamless integration of advanced deep learning vision modules with robotic actuators. As shown in Fig. 7, such a system would use a high-speed camera to capture images of grains on a conveyor, a powerful deep learning model running on an edge computer to identify and classify each kernel in real-time, and a robotic system—such as a collaborative robot arm or a bank of high-speed air-jet ejectors—to physically separate the grains into different quality grades. The development of these closed-loop, intelligent systems represents the culmination of the technological trends discussed in this review and holds the promise of transforming grain processing into a highly efficient, precise, and fully automated operation.

An Analytical decision framework for deployable grain inspection

A recurring theme across the literature is that architecture choice is not a purely algorithmic preference but a consequence of the interaction between task structure, imaging conditions, and deployment constraints. We therefore synthesize the reviewed studies into a practical analytical framework that explains when a method is likely to succeed and what evidence is required to support deployable claims. First, the nature of the discriminative signal determines the most suitable model family. Tasks dominated by local and high-contrast cues (e.g., chipped kernels, obvious discoloration, broken grain) tend to be well served by efficient CNN detectors or segmenters because the required evidence is spatially localized and can be captured with hierarchical convolutional features. Conversely, tasks involving diffuse or spatially distributed cues (e.g.,

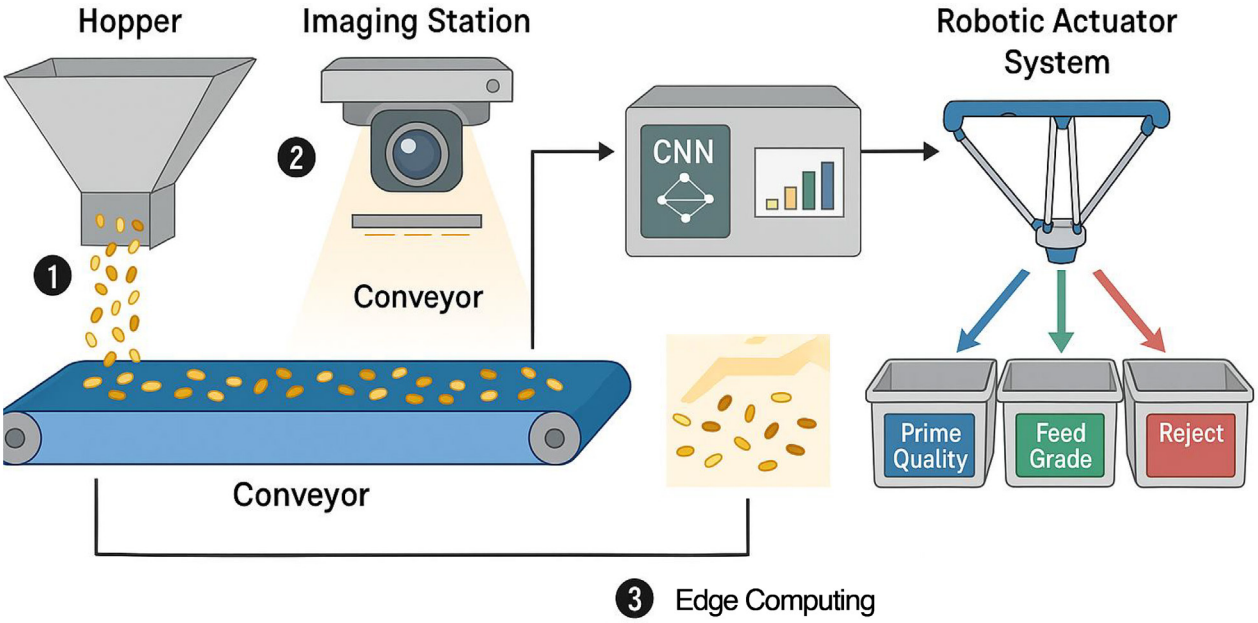

Figure 7. An integrated autonomous system combining a deep learning vision module with a robotic sorting actuator. This diagram illustrates a futuristic, fully automated grain sorting line.

mold patterns, subtle systemic defects, fine-grained varietal signatures) often benefit from stronger global context modeling, which motivates attention-based mechanisms and, in some cases, transformer-style architectures. Recent transformer-based grain quality studies report very high recognition accuracy on curated datasets, but they also illustrate that such gains come with increased sensitivity to data scale and computational cost, reinforcing the need to justify transformers with task-specific signal characteristics and deployment feasibility rather than novelty alone (Zhang et al. 2025).

Second, controlled-condition accuracy must be interpreted as an upper bound rather than a deployment guarantee. Classical machine vision can achieve extremely high performance when the imaging pipeline is standardized; for example, morphology-based grain classification has reported accuracies approaching 99% for several cereal types under controlled acquisition (Majumdar and Jayas 2000)). However, more realistic varietal classification settings show lower accuracy and stronger dependence on imaging variability; CNN-based barley varietal classification has been reported at 93%, indicating both the promise of learned features and the persistence of domain sensitivity (Kozłowski et al. 2019). This contrast motivates an evidence standard: authors should report not only in-distribution metrics but also performance under controlled perturbations (illumination, occlusion, motion blur) and, where possible, cross-domain validation.

Third, deployability requires reporting efficiency alongside accuracy. For online sorting, throughput constraints can be as decisive as recognition quality. Lightweight models demonstrate that competitive accuracy can be achieved under small memory footprints suitable for edge devices (e.g., 1.33 MB for LWheatNet at 98.59% accuracy), but such results must be paired with latency/throughput reporting under the intended hardware and image resolution to enable fair comparison and engineering decisions (Guo et al. 2025).

Finally, based on these observations, we recommend a minimum reporting standard for future grain inspection studies to support analytical comparability and industrial translation. At minimum, studies should specify acquisition setup (camera, lens, lighting, conveyor speed where relevant), dataset composition and split strategy (including cross-domain splits if deployment is claimed), primary and secondary metrics (e.g., F1 for defect classes, mAP for detection, calibration where decision thresholds matter), and efficiency indicators (model size, inference latency/throughput on stated hardware). Where novelty detection or safety-critical rejection is relevant, out-of-distribution evaluation should be included, consistent with emerging benchmark formulations in grain inspection datasets. To make the above selection logic actionable, we summarize the decision perspective in Table 5, which maps typical grain inspection conditions to the recommended model family and provides representative literature examples

Conclusions

This review consolidates evidence across classical machine vision, CNN-based deep learning, and transformer-family architectures to derive deployability-oriented conclusions rather than a purely narrative recap. First, method selection should be driven by the structure of the discriminative cue and the deployment envelope: when quality cues are explicit and measurable under controlled imaging (shape/size/color/texture), classical MV with handcrafted descriptors can remain competitive and efficient; in contrast, for real-world inspection and sorting

Table 5. Decision perspective for selecting classical MV, CNN, or Transformer for grain quality inspection.

Decision driver (what dominates your problem)	Classical MV (handcrafted features + shallow ML)	CNN (end-to-end conv models / detectors)	Transformer (ViT/Swin or hybrid CNN–Transformer)
Acquisition environment	Best when imaging is highly controlled (uniform background/lighting; stable camera geometry), enabling reliable segmentation and feature measurement	Tolerates moderate variability (lighting drift, mild occlusion/clutter) with augmentation and robust training	Most justified when global context is needed and acquisition variability is addressed via scale/attention designs (e.g., hierarchical windows)
Data regime	Works with small datasets because features are engineered	Works well from small-to-medium datasets, especially with transfer learning; scales with more data	Typically benefits from larger datasets; data-efficient training helps but still needs justification
Cue structure	Explicit, measurable cues (size/shape/color/texture)	Local-to-midrange cues (chips, cracks, localized mold/defects; detection on conveyor)	Diffuse/global cues (subtle distributed discoloration; long-range patch relations)
Compute/latency constraints	Lowest compute; easiest for embedded PLC-like integration	Good edge feasibility with lightweight backbones; strong throughput for sorting lines	Higher compute; efficient variants exist but still costlier; best when accuracy gains justify it
Interpretability needs	Highest (features correspond to physical traits)	Medium (saliency/Grad-CAM style)	Medium-to-high via attention visualization, but still model-complex
Representative grain-quality examples (from literature)	Morphology-based cereal grain classification under controlled acquisition achieves very high accuracy (Majumdar and Jayas 2000); controlled kernel imaging pipelines are widely discussed in agri-food CV reviews	CNN-based barley varietal classification reported strong performance, supporting CNNs as a practical default under realistic conditions	Improved Swin Transformer applied to maize seed variety recognition demonstrates transformer utility when multi-scale/global context is beneficial

where kernels are variably oriented, partially occluded, and acquired under drifting illumination, CNNs provide a strong default because their convolutional inductive bias supports robust learning under moderate data regimes and enables high-throughput detection on production lines (e.g., CNN-based barley varietal classification in realistic settings). Second, transformer-style models are most justified when the task depends on globally distributed or subtle, long-range patterns (diffuse discoloration, fine-grained signatures, or multi-scale context), where self-attention can capture dependencies that are difficult to aggregate with purely local receptive fields; recent grain/seed studies using Swin-style transformers illustrate the potential of attention mechanisms for such recognition problems. Third, the conclusion from the synthesis tables is that headline accuracy values are not, by themselves, evidence of deployability: claims intended for industrial translation should be supported by evaluation on shared benchmarks or harmonized protocols, and should include robustness checks under domain shift (device/region/lighting) because this variability can dominate error modes. Large-scale benchmarks such as GrainSpace explicitly formulate grain appearance inspection across fine-grained recognition, domain adaptation, and out-of-distribution recognition, reinforcing that cross-domain testing is an evidence requirement rather than an optional add-on. Fourth, engineering selection requires efficiency reporting alongside recognition metrics: future studies should report at least model size and inference throughput/latency on stated hardware and resolution, enabling readers to translate algorithmic performance into throughput budgets and actuation timing constraints. Finally, these takeaways collectively imply that the near-term trajectory is not a winner-take-all transition from CNNs to transformers, but a pragmatic move toward hybrid or task-specialized pipelines in which convolutional components provide efficient local feature extraction while attention modules contribute global context modeling when the signal structure warrants it, aligned with the broader transformer literature.

Availability of data and materials

Data sharing is not applicable as no data were generated or analyzed.

Author contributions

B.W. designed the research study, performed the research, and wrote the manuscript.

Ethics approval and consent to participate

Not applicable.

Conflict of Interest

The author has declared that no competing interests exist.

References

Ahad MT, Li Y, Song B, Bhuiyan T (2023) Comparison of CNN-based deep learning architectures for rice diseases classification. Artificial Intelligence in Agriculture 9: 22–35. https://doi.org/10.1016/j.aiia.2023.07.001

Ahmad R, Rehman M, Basit A, Basit MK, Ahmed H, Kajla NI (2025) Advanced rice grain classification using hybrid vision transformer models. Statistics, Computing and Interdisciplinary Research 7: 27–39. https://doi.org/10.52700/scir.v7i1.179

Ahmed SB, Ali SF, Khan AZ (2021) On the frontiers of rice grain analysis, classification and quality grading: A review. IEEE Access 9: 160779–160796. https://doi.org/10.1109/ACCESS.2021.3130472

Ayoub S, Baig I, Ashraf M, Okasha M (2025) Impact of dataset quality on deep learning models for dragon fruit and leaf health classification. Impact in Agriculture 1: 1–16. https://doi.org/10.65500/agriculture-2025-001

Bagherzadeh S, Maghooli K, Shalbaf A, Maghsoudi A (2022) Emotion recognition using effective connectivity and pre-trained convolutional neural networks in EEG signals. Cognitive Neurodynamics 16: 1087–1106. https://doi.org/10.1007/s11571-021-09756-0

Bao H, Luo J, Li J, Zhang H, Zhou H (2023) Automatic Detection of Stored Grain Fungal Spores Based on Deep Learning. 2021 7th IEEE International Conference on Network Intelligence and Digital Content, 127–131. https://doi.org/10.1109/icnidc59918.2023.10390801

Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D (2021) Machine learning in agriculture: a comprehensive updated review. Sensors 21: 3758. https://doi.org/10.3390/s21113758

Berkmanas R (2024) Food Grain Sorting With Computer Vision: Case Study. https://easyodm.tech/food-grain-sorting-with-computer-vision/ [August 14, 2025]

Bi C, Hu N, Zou Y, Zhang S, Xu S, Yu H (2022) Development of deep learning methodology for maize seed variety recognition based on improved swin transformer. Agronomy 12(8): 1843. https://doi.org/10.3390/agronomy12081843

Bishaw MW, Ejigu SM, Tiruneh SG, Dessalegn AA, Bishaw MW, Ejigu SM, Tiruneh SG, Workneh AA (2025) Automatic quality assessment of red haricot beans using deep learning. Cureus Journals 2: es44389-025-05423-1. https://doi.org/10.7759/s44389-025-05423-1

Chen X, Chen J, Han X, Zhao C, Zhang D, Zhu K, Su Y (2020) A light-weighted CNN model for wafer structural defect detection. IEEE Access 8: 24006–24018. https://doi.org/10.1109/ACCESS.2020.2970461

Douik A, Abdellaoui M (2010) Cereal grain classification by optimal features and intelligent classifiers. International Journal of Computers Communications & Control 5: 506–516. https://doi.org/10.15837/ijccc.2010.4.2508

Du Y, Chang C-I, Ren H, Chang C-C, Jensen JO, D'Amico FM (2004) New hyperspectral discrimination measure for spectral characterization. Optical Engineering 43: 1777–1786. https://doi.org/10.1117/1.1766301

Evers T, Millar S (2002) Cereal grain structure and development: some implications for quality. Journal of Cereal Science 36: 261–284. https://doi.org/10.1006/jcrs.2002.0435

Fan L, Ding Y, Fan D, Di D, Pagnucco M, Song Y (2022) GrainSpace: A large-scale dataset for fine-grained and domain-adaptive recognition of cereal grains. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21116–21125. https://doi.org/10.1109/CVPR52688.2022.02044

Fidan M, Kaya U (2020) A novel color-based feature extraction method for svm based skin segmentation. Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering 21: 45–54. https://doi.org/10.18038/estubtda.818577

Fu Y, Yang G, Li Z, Song X, Li Z, Xu X, Wang P, Zhao C (2020) Winter wheat nitrogen status estimation using UAV-based RGB imagery and gaussian processes regression. Remote Sensing 12: 3778–3778. https://doi.org/10.3390/rs12223778

Guan-shi L, Tian S, Mo Y, Chen R, Zhao Q (2022) On the acquisition of high-quality digital images and extraction of effective color information for soil water content testing. Sensors 22: 3130–3130. https://doi.org/10.3390/s22093130

Gulzar Y (2025a) Applications of transfer learning in sunflower disease detection: advances, challenges, and future directions. Turkish Journal of Biology 49: 534–549. https://doi.org/10.55730/1300-0152.2763

Gulzar Y (2025b) Papaya leaf disease classification using pre-trained deep learning models: a comparative study. Applied Fruit Science 67: 287. https://doi.org/10.1007/s10341-025-01533-1

Gulzar Y (2025c) PapNet: an ai-driven approach for early detection and classification of papaya leaf diseases. Applied Fruit Science 67: 256. https://doi.org/10.1007/s10341-025-01466-9

Guo X, Wang J, Gao G, Cheng Z, Qiao Z, Zhang R, Ma Z, Wang X (2025) LWheatNet: a lightweight convolutional neural network with mixed attention mechanism for wheat seed classification. Frontiers in Plant Science 15: 1509656. https://doi.org/10.3389/fpls.2024.1509656

Gupta PK, Waoo AA (2024) Analyzing machine learning methods to enhance grain quality assessment and evolution. ShodhKosh: Journal of Visual and Performing Arts 5: 513–525. https://doi.org/10.29121/shodhkosh.v5.i1.2024.1876

Gurău L, Timar M, Coșereanu C, Coșniță M, Stanciu M (2023) Aging of wood for musical instruments: analysis of changes in color, surface morphology, chemical, and physical-acoustical properties during UV and thermal exposure. Polymers 15: 1794–1794. https://doi.org/10.3390/polym15071794

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 (2022) A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 45: 87–110. https://doi.org/10.1109/tpami.2022.3152247

Haque Z, Koklu M, Mirza MA, Omar M, Mukhayyokhon S (2025) Mitigating spatial scale loss in CNN-based fine-grained image classification: application to date fruit grading. Impact in Agriculture 1: 1–15. https://doi.org/10.65500/agriculture-2025-005

Hassanisaadi M, Bonjar G, Rahdar A, Pandey S, Hosseinipour A, Abdolshahi R (2021) Environmentally safe biosynthesis of gold nanoparticles using plant water extracts. Nanomaterials 11: 2033–2033. https://doi.org/10.3390/nano11082033

Hou Q, Lu C, Cheng M, Feng J (2024) Conv2Former: a simple transformer-style convnet for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 46: 8274–8283. https://doi.org/10.1109/tpami.2024.3401450

Islam MdM, Himel GMS, Uddin MS, Moazzam MdG (2024) A visual dataset for recognition of rice varieties. Data in Brief 54: 110442. https://doi.org/10.1016/j.dib.2024.110442

Jiang K, Wang Z, Chen C, Wang Z, Cui L, Lin C (2022) Magic ELF: Image Deraining Meets Association Learning and Transformer. Proceedings of the 30th ACM International Conference on Multimedia, 827–836. https://doi.org/10.1145/3503161.3547760

Khan S, Naseer M, Hayat M, Zamir S, Khan F, Shah M (2022) Transformers in vision: a survey. ACM Computing Surveys 54: 1–41. https://doi.org/10.1145/3505244

Koklu M, Cinar I, Taspinar YS (2021) Classification of rice varieties with deep learning methods. Computers and Electronics in Agriculture 187: 106285. https://doi.org/10.1016/j.compag.2021.106285

Kozłowski M, Górecki P, Szczypiński PM (2019) Varietal classification of barley by convolutional neural networks. Biosystems Engineering 184: 155–165. https://doi.org/10.1016/j.biosystemseng.2019.06.012

Krichen M (2023) Convolutional neural networks: a survey. Computers 12: 151–151. https://doi.org/10.3390/computers12080151

Lingwal S, Bhatia KK, Tomer MS (2021) Image-based wheat grain classification using convolutional neural network. Multimedia Tools and Applications 80: 35441–35465. https://doi.org/10.1007/s11042-020-10174-3

Liu Q, Zhang X, Jiang X, Zhang C, Li J, Zhang X, Yang J, Yu N, Zhu Y, Liu J, Xie F, Li Y, Hao Y, Feng Y, Wang Q, Gao Q, Zhang W, Zhang T, Dong T, Cui B (2024) A histopathologic image analysis for the classification of endocervical adenocarcinoma silva patterns depend on weakly supervised deep learning. American Journal Of Pathology 194: 735–746. https://doi.org/10.1016/j.ajpath.2024.01.016

Loddo A, Di Ruberto C (2021) On the efficacy of handcrafted and deep features for seed image classification. Journal of Imaging 7: 171. https://doi.org/10.3390/jimaging7090171

Ma J, Sun D-W, Qu J-H, Liu D, Pu H, Gao W-H, Zeng X-A (2016) Applications of computer vision for assessing quality of agri-food products: a review of recent research advances. Critical Reviews in Food Science and Nutrition 56: 113–127. https://doi.org/10.1080/10408398.2013.873885

Ma N, Su Y, Yang L, Li Z, Yan H (2024) Wheat seed detection and counting method based on improved YOLOv8 model. Sensors 24: 1654. https://doi.org/10.3390/s24051654

Majumdar S, Jayas D (2000) Classification of cereal grains using machine vision: I. Morphology models. Transactions of the ASAE 43: 1669–1675. https://doi.org/10.13031/2013.3107

Meghraoui K, Sebari I, Pilz J, Ait El Kadi K, Bensiali S (2024) Applied deep learning-based crop yield prediction: a systematic analysis of current developments and potential challenges. Technologies 12: 43. https://doi.org/10.3390/technologies12040043

Pabamalie LAI, Premaratne HL (2010) A Grain Quality Classification System. 2010 International Conference on Information Society, 56–61. https://doi.org/10.1109/i-Society16502.2010.6018794

Patra AK, Varshney A, Sahoo L (2025) An explainable vision transformer with transfer learning based efficient drought stress identification. Plant Molecular Biology 115: 98. https://doi.org/10.1007/s11103-025-01620-7

Qi X, Zhang S, Wang L, Hu X, Zhang H, Feng W, Wang C, Guo T, He L (2025) Classification of different gluten wheat varieties based on hyperspectral preprocessing, feature screening, and machine learning. Food Chemistry: X 26: 102329. https://doi.org/10.1016/j.fochx.2025.102329

Ramesh S, Sasikala S, Paramanandham N (2021) Segmentation and classification of brain tumors using modified median noise filter and deep learning approaches. Multimedia Tools and Applications 80: 11789–11813. https://doi.org/10.1007/s11042-020-10351-4

Rezaei M, Diepeveen D, Laga H, Jones M, Sohel F (2024) Plant disease recognition in a low data scenario using few-shot learning. Computers and Electronics in Agriculture 219: 108812–108812. https://doi.org/10.1016/j.compag.2024.108812

Supreetha S, Premalathamma R, Manjula SH (2024) Deep Learning Techniques to Detect Nutrient Deficiency in Rice Plants. 2024 International Conference on Inventive Computation Technologies (ICICT), 699–705. https://doi.org/10.1109/ICICT60155.2024.10544924

Shedole S, Sowmya BJ, Neha Arthi VP (2022) A convolution neural network-based wheat grain classification system. Journal of Scientific Research 66: 22–29. https://doi.org/10.37398/JSR.2022.660204

Shen L, Wang Q, Zhang Y, Qin F, Jin H, Zhao W (2023) DSKCA-UNet: Dynamic selective kernel channel attention for medical image segmentation. Medicine 102: e35328–e35328. https://doi.org/10.1097/md.0000000000035328

Shi P, Duan M, Yang L, Feng W, Ding L, Jiang L (2022) An improved U-net image segmentation method and its application for metallic grain size statistics. Materials 15: 4417–4417. https://doi.org/10.3390/ma15134417

Shuai L, Li Z, Chen Z, Luo D, Mu J (2024) A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing. Computers and Electronics in Agriculture 217: 108577. https://doi.org/10.1016/j.compag.2023.108577

Szczypiński PM, Zapotoczny P (2012) Computer vision algorithm for barley kernel identification, orientation estimation and surface structure assessment. Computers and Electronics in Agriculture 87: 32–38. https://doi.org/10.1016/j.compag.2012.05.014

Tan Y, Chang X, Zhong J, Feng W, Yang M, Tian T, Gong L, Wu W (2022) Chemical linkage and passivation at buried interface for thermally stable inverted perovskite solar cells with efficiency over 22%. CCS Chemistry 5: 1802–1814. https://doi.org/10.31635/ccschem.022.202202154

Tian Y, Zhang J, Zhang Z, Wu J (2024) Research on super-resolution enhancement technology using improved transformer network and 3D reconstruction of wheat grains. IEEE Access 12: 62882–62898. https://doi.org/10.1109/ACCESS.2024.3396148

van Vliet LJJ, Atkins K, Kurup S, Siles L, Hepworth J, Corke FMK, Doonan JH, Lu C (2025) DeepCanola: Phenotyping brassica pods using semi-synthetic data and active learning. Computers and Electronics in Agriculture 237: 110470. https://doi.org/10.1016/j.compag.2025.110470

Wang L, Liu D, Pu H, Sun D-W, Gao W, Xiong Z (2015) Use of hyperspectral imaging to discriminate the variety and quality of rice. Food Analytical Methods 8: 515–523. https://doi.org/10.1007/s12161-014-9916-5

Wang N, Dowell F, Zhang N (2003) Determining wheat vitreousness using image processing and a neural network. Transactions of the ASAE 46: 1143. https://doi.org/10.13031/2013.13937

Wang W, Dai J, Chen Z, Huang Z, Li Z, Zhu X, Hu X, Lü T, Lu L, Li H, Wang X, Qiao Y (2023) InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14408–14419. https://doi.org/10.1109/cvpr52729.2023.01385

Xu M, Park J-E, Lee J, Yang J, Yoon S (2024) Plant disease recognition datasets in the age of deep learning: challenges and opportunities.

Frontiers in Plant Science 15: 1452551. https://doi.org/10.3389/fpls.2024.1452551

Yang Z-Y, Xia W-K, Chu H-Q, Su W-H, Wang R-F, Wang H (2025) A comprehensive review of deep learning applications in cotton industry: from field monitoring to smart processing. Plants 14: 1481. https://doi.org/10.3390/plants14101481

Yuan L, Hou Q, Jiang Z, Feng J, Yan S (2022) VOLO: Vision Outlooker for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–13. https://doi.org/10.1109/tpami.2022.3206108

Zangana HM, Li S, Wani S (2025) Diffusion models for agricultural imaging: a systematic review of methods, applications and future prospects. Impact in Agriculture 1: 1–11. https://doi.org/10.65500/agriculture-2025-003

Zaw K, Myo Z, Thoung D (2023) Support vector machine based classification of leaf diseases. International Journal of Science and Engineering Applications: 143–147. https://doi.org/10.7753/ijsea0708.1003

Zhang N, Chen Y, Zhang E, Liu Z, Yue J (2025) Maize quality detection based on MConv-SwinT high-precision model. PLoS ONE 20: e0312363. https://doi.org/10.1371/journal.pone.0312363