Deep Learning and Computer Vision for Maize Seed Classification: A Review of Current Methods and Future Directions

Ajay Singh Dhabariya[1], Dr. C. D. Kumawat[2]

[1]Ph.D. Scholar, Department of Computer Science and Engineering Mewar University, Chittorgarh, India
ajay.itengineer@gmail.com
[2]Professor, Department of Computer Science and Engineering, Mewar University, Chittorgarh, India
chandikaditya@gmail.com

Abstract
Seed purity is a key indicator of maize seed qual-ity and vital for food security. This study presents a non-destructive, high-throughput system combining RGB imaging, hyperspectral data, near-infrared sensing, and biopecle analysis to separate pure and mixed seeds. Deep-learning models such as Swin Transformer, EfficientNetV2, and self-supervised DINOv2 features are employed, with seed segmentation performed using the Segment Anything Model. The pipeline includes dataset collection, modality-specific preprocessing, multimodal fusion, and comprehensive evaluation with ablation studies. The goal is to achieve over 95% purity-classification accuracy. Key contri-butions include a practical multimodal framework and a hybrid supervised, self-supervised training strategy.

Keywords: Maize seed purity, non-destructive testing, Hy-perspectral Imaging (HSI), Biopecle, Swin Transformer, Efficient-NetV2, DINOv2, Segment Anything (SAM), Multimodal Fusion.

1. Introduction
Seed purity, defined as the proportion of seeds that accu-rately represent the declared cultivar, is critical for ensuring crop uniformity and maintaining confidence among growers. Conventional purity-evaluation procedures, whether based on manual inspection or destructive genetic assays, offer reliable results but remain slow, labour-intensive, and economically burdensome. Recent advancements in imaging technologies, including RGB, hyperspectral, and near-infrared modalities, together with modern machine-learning approaches, have demonstrated strong potential for automating seed-quality as-sessment. Maize (Zea mays L.) is one of the most important cereal crops worldwide, serving as a staple food, animal feed, and raw material for various industrial products. The quality of maize seed directly influences germination, crop vigor, and final yield, making seed purity a critical parameter in modern agriculture. Seed purity is defined as the proportion of seeds that truly belong to the declared cultivar, free from off-types, foreign species, damaged kernels, and other contaminants. Inaccurate purity estimation can lead to non-uniform crop stands, reduced productivity, economic loss for farmers, and diminished confidence in seed brands [1], [2].
Conventional methods for assessing seed purity, such as manual visual inspection, germination tests, and molecular or biochemical assays, are reliable but suffer from several drawbacks. They are often time-consuming, labor-intensive, subjective, and, in the case of genetic tests, partially destructive and expensive. These limitations make it challenging to apply traditional techniques at scale in high-throughput industrial settings, where large seed lots must be evaluated quickly and consistently. Recent advances in artificial intelligence (AI), particularly deep learning, have opened new opportunities for automating agricultural inspection. Convolutional neural networks (CNNs) and transformer-based architectures have achieved state-of-the-art performance in image classification and defect detection, while non-destructive sensing modalities such as RGB imaging, hyperspectral imaging (HSI), near-infrared (NIR) spectroscopy, and biopecle laser analysis can capture complementary information about seed morphology, chemistry, and physiology. However, many existing studies focus on a single modality or operate under tightly controlled laboratory conditions, limiting their robustness and real-world applicability [3], [4].
This research addresses these gaps by proposing a hybrid deep-learning framework that integrates multiple sensing tech-nologies, RGB, HSI, NIR, and biopecle, to perform non-destructive maize seed purity assessment. The system lever-ages advanced models such as EfficientNetV2, Swin Trans-former, and 1D/3D CNNs, complemented by self-supervised DINOv2 embeddings for improved generalisation. Seed re-gions are automatically segmented using the Segment Any-thing Model (SAM), ensuring precise region-of-interest extrac-tion and reducing manual annotation effort. Features from all modalities are fused using late and attention-based strategies to generate highly discriminative representations. By combining multimodal sensing with state-of-the-art deep learning and robust fusion mechanisms, the proposed approach aims to achieve high purity-classification accuracy while remaining scalable and practical for deployment in seed certification laboratories and industrial processing lines [5], [6].

2. Literature Review
A.Deep Learning for RGB-Based Seed Quality Assessment
Kavitha et al. (2024) present a deep-learning-based pipeline for maize seed quality detection using ResNet-based architec-tures, with a focus on Inception-ResNet v2 and ResNet152v2. The study is rooted in the practical need to replace manual and biochemical seed quality inspection, described as slow, labour-intensive, and error-prone, with an automated, non-destructive image-based system. The authors construct a sizeable RGB image dataset of maize seeds captured from top and bottom views, covering four key classes: broken, discoloured, silk-cut, and pure seeds. The

primary dataset comprises 17,802 expert-labelled images, further expanded to a balanced dataset of 26,802 images via Conditional GAN-generated synthetic samples and active learning to label 9,000 previously unlabeled images. This ensures better class balance and supports robust training. Methodologically, the paper leverages transfer learning by adopting Inception-ResNet v2 as a pretrained backbone, removing the original classification head and adding custom layers for binary or multi-class seed quality prediction [15]. Preprocessing includes spectral/image normalization, segmen-tation to isolate seed regions, and standard data augmentation (rotation, flipping, scaling) to improve generalization. Training uses Adam optimization, binary cross-entropy loss, and an 80:20 train-validation split, with all 783 layers set train-able in the final configuration. Experimentally, the improved Inception-ResNet model achieves an accuracy of 96.03%, with a precision of 96.27%, a recall of 96.13%, and an F1-score of 96.15%, outperforming SVM, vanilla CNN, RNN, and LSTM baselines [15]. The authors also discuss challenges such as high computational cost, complexity of Inception-ResNet, and the need for substantial hardware resources and careful preprocessing. Overall, this work convincingly demonstrates the effectiveness of deep residual and Inception-ResNet archi-tectures for maize seed appearance-based quality assessment. However, it remains limited to RGB imaging and surface features; it does not incorporate spectral or physiological modalities, nor does it address multimodal fusion gaps that the proposed EfficientNetV2-based, multimodal framework explicitly targets [15].

B. Hyperspectral Imaging for Defect Detection

Xu et al. (2023) advance seed quality assessment research by leveraging hyperspectral imaging (HSI) in conjunction with deep-learning models to identify worm-eaten maize seeds [9]. Unlike visual inspection methods that rely solely on external features, HSI combines spatial and spectral dimen-sions, capturing biochemical variations in protein, starch, fat, and moisture. The authors employ a hyperspectral range of $866.4 - 1701$nm, comprising 254 bands, which are subse-quently refined through spectral preprocessing and feature selection. To eliminate noise and baseline drift, raw spectra undergo Savitzky-Golay smoothing, detrending, and multiple scattering correction (MSC), producing more consistent re-flectance signatures across samples. This preprocessing reveals several diagnostic absorption valleys around 960, 1000, 1350, and 1660 nm linked to protein, starch, carbohydrate, and lipid structures, establishing HSI's capacity to detect internal defects invisible to RGB imaging [9]. The study's core contri-bution lies in two custom 1D-CNN architectures: CNN-FES (feature selection) and CNN-ATM (attention-based classifica-tion). CNN-FES introduces a learnable wavelength selection mechanism inspired by permutation importance, enabling the model to identify information-rich spectral bands while sup-pressing redundant channels. This reduces dimensionality and makes the feature selection task-aware, outperforming classical approaches such as SPA and CARS. CNN-ATM, adapted from squeeze-and-excitation principles, enhances discrimina-tive channels by weighting spectral responses during inference, effectively mimicking physiological prioritization of valuable traits. Experimental validation using 400 kernels demonstrates that CNN-ATM achieves 97.50% accuracy, 98.28% sensitivity, and 96.77% specificity when using feature bands, surpassing LDA, SVM, and RF. While the findings highlight remarkable detection capability, the study is limited by single-modality dependence and a small sample size. The exclusive focus on NIR hyperspectral signatures overlooks morphological cues and temporal germination signals. Consequently, Xu et al. provide a strong foundation for non-destructive spectral analysis, but their approach can be enhanced through multimodal fusion, attention-based hybrid architectures, and scalable datasets di-rections addressed in the proposed research [9].

C. Seed Vigor Classification Using HSI and Deep Learning

Wongchaisuwat et al. (2025) provide one of the most com-prehensive efforts to date on non-destructive seed vigor classi-fication by integrating hyperspectral imaging (HSI) with deep learning architectures [28]. Unlike many previous studies that rely on artificially aged seeds, this research evaluates coated, industrially processed maize seeds stored under real-world conditions, directly addressing scalability and seed industry applicability. The authors emphasize seed vigor as a physi-ological indicator of the seed's ability to germinate robustly across environments traditionally measured using destructive radicle emergence tests or multi-day laboratory analyses. Their framework replaces these labor-intensive methods with a rapid hyperspectral system capable of classifying seed vigor in less than ten seconds. The methodology is grounded in a rigorous physiological-digital pipeline. True vigor labels are generated using radicle emergence testing, curve fitting, and K-means clustering on emergence metrics, ensuring biologically mean-ingful ground truth [28]. HSI acquisition uses the SPECIM FX17 system (935 - 1720 nm), followed by morphological segmentation, reflectance correction, and pixel-level averaging to generate per-kernel spectral signatures. Importantly, the authors explore the role of spectral dimensionality reduction via the Successive Projections Algorithm (SPA), identifying 106 key wavelengths. This feature selection reduces computa-tional costs and enables hardware simplification without overly sacrificing accuracy. A distinctive contribution of the study is the comparative evaluation of multiple CNN topologies: 1D-CNN for pure spectral signals, 2D-CNN for spatially enhanced spectral images, EfficientNetB0 variants, and 3D-CNN architectures that jointly model spatial and spectral dependencies. Experimental results reveal a dominant advan-tage of 3D-CNN with full wavelengths, achieving 99.91% accuracy, 100% sensitivity, 99.89% specificity, and MCC of 0.9966. Even EfficientNetB0 with pretrained weights performs robustly (> 99%) demonstrating the suitability of transfer learning for HSI-based agricultural tasks. Despite exceptional performance, the authors acknowledge trade-offs: 3D-CNN models are computationally heavy and less practical for in-dustrial integration than lightweight architectures [28]. The study remains limited to a single modality (HSI) and does not exploit temporal or morphological cues. Nevertheless, the work sets a high benchmark for spectral-based vigor prediction and highlights the necessity of multimodal fusion, attention mechanisms, and scalable seed-processing integration critical directions addressed in the proposed hybrid sensing approach [28].

D. Research Objectives

The specific objectives of this research are:

- To enhance prediction of seed quality using deep learning model and image-based, sensor-based data.
- To collect an accurate comprehensive dataset of seeds labeled for quality attributes (e.g., germination rate, via-bility, defects, shape, size, moisture).
- To compare various deep learning architectures such as CNNs, ResNet, and EfficientNet for performance evalu-ation.
- To integrate explainable AI (XAI) methods to interpret model decisions and ensure transparency.
- To design a prototype tool or application that can be used in seed processing industries or farms for real-time seed quality assessment.

3. Research Methodology

This section describes, in detail, the methodology adopted to design and implement a non-destructive, multimodal deep-learning system for maize seed purity assessment. The system integrates four sensing modalities: RGB imaging, hyperspec-tral imaging (HSI), near-infrared (NIR) spectroscopy and biopecks laser imaging, with modern deep learning archi-tectures such as EfficientNetV2, Swin Transformer, DINOv2, 1D/3D CNNs, and the Segment Anything Model (SAM) for segmentation. The goal is to build a practical, high-throughput pipeline that can classify individual seeds as pure or impure.

A. Research Design and System Overview

The overall methodology follows a modular, pipeline-oriented design, moving from sensing to decision as illustrated in Fig. 1. The proposed system architecture represents a com-plete pipeline for non-destructive maize seed purity assessment using multimodal sensing and deep learning. The process begins with the Multimodal Acquisition stage, where seed data are captured using multiple complementary sensors such as RGB imaging, hyperspectral imaging, near-infrared sensing, and biopecle analysis. This stage ensures that both external appearance and internal biochemical as well as biological activity information of seeds are recorded. The acquired data are then passed to the SAM Segmentation module, where the Segment Anything Model automatically extracts precise seed regions from the background. This step is crucial for
isolating only the relevant seed area and removing redundant background noise that could negatively affect model perfor-mance. Following segmentation, the data undergo Modality-wise Preprocessing, where each modality is processed inde-pendently using suitable techniques such as normalization, noise filtering, spectral calibration, and dimensionality reduc-tion. The preprocessed data are then forwarded to the Deep Feature Extraction stage, where deep learning models extract high-level discriminative features from each modality. Next, the extracted features are integrated through the Multimodal Fusion module, which combines complementary information into a unified representation. The fused features are fed into the Classification block to predict whether each seed is pure or impure. Finally, the system computes Lot-level Purity Estimation by aggregating individual seed predictions to obtain an overall purity percentage for the seed batch [8], [9].

The overall system architecture is shown in Fig. 2. The system consists of seven main blocks: Multimodal Sensing Rig (RGB camera, HSI camera, NIR spectrometer, Biopecle laser with high-speed camera); Segmentation & ROI Extraction using SAM; Modality-specific Preprocessing (RGB normal-ization and resizing, HSI calibration and denoising, NIR nor-malization, Biopseckle activity map generation); Deep Feature Extraction Branches (RGB branch with EfficientNetV2, Swin, DINOv2; HSI branch with 3D CNN; NIR branch with 1D CNN; Biopseckle branch with EfficientNetV2); Multimodal Fusion Module using concatenation and attention-based fu-sion; Classification Head using MLP for Pure/Impure predic-tion per seed; and Lot-level Aggregation for computing pre-dicted purity percentage. This layered design ensures that each component can be tested, replaced, or expanded independently.

B. Data Acquisition and Ground Truth

Maize seed lots are collected from multiple varieties, pro-ducers, and seasons to capture realistic variation. Each lot contains a mixture of pure seeds (correct cultivar, intact, visually healthy) and impure seeds including off-variety seeds, seeds of other crops or weeds, broken or chipped kernels, seeds with fungal infection, discolouration, or shortening, and non-seed foreign material. Ground-truth labels (pure/impure) are determined using standard seed testing protocols and ex-pert inspection. Optionally, finer sub-labels (e.g., broken, off-type, infected) may be recorded for analysis, while the main classification task is binary. A target dataset size of 20,000 - 30,000 seeds is aimed for, ensuring that impurity classes are reasonably represented, even though they are naturally rarer than pure seeds [10], [11].

Each seed passes through a multimodal acquisition line. For RGB imaging, an industrial or DSLR camera under controlled LED lighting is used with seeds arranged on trays with a neutral, non-reflective background at resolution high enough to capture fine surface texture and shape. For hyperspectral imag-ing, an HSI camera covering either visible-NIR (e.g., 450-950 nm) or NIR (e.g., 900-1700 nm) bands is used with reference white and dark images recorded for reflectance normalization, producing a 3D cube (height × width × spectral bands)

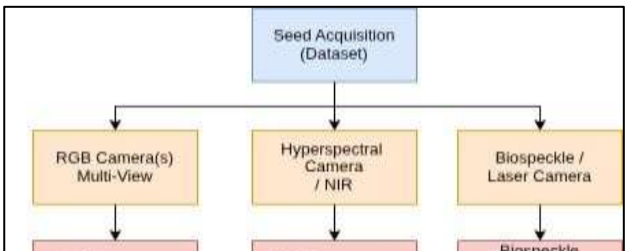

Fig. 1: Proposed model methodology

for each tray. For near-infrared spectroscopy, a point or line NIR spectrometer captures a 1D spectral signature for each seed, which is faster than full HSI and offers complementary chemical information. For biopseckle laser imaging, a low-power coherent laser illuminates the seeds and a high-speed camera records a short video sequence (a few seconds). Time-varying speckle patterns carry information about biological activity and vigor. Each seed is assigned a unique ID, and its RGB, HSI ROI, NIR spectrum, and biopecle activity map are stored and aligned with the purity label.

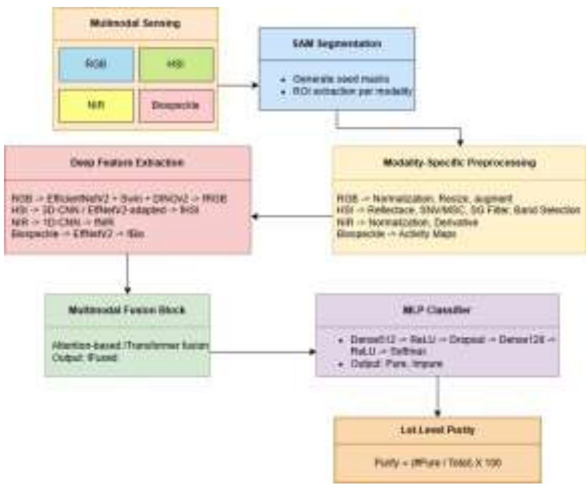

Fig. 2: Overall system architecture

C. Segmentation and ROI Extraction using SAM

Accurate segmentation of seeds from the background is cru-cial. Instead of hand-drawing masks, this work uses the Seg-ment Anything Model (SAM), a powerful foundation model for segmentation. The SAM-based ROI extraction involves input of tray-level RGB images, prompting SAM using light user prompts or automatic grid-based prompts to predict binary masks for each visible seed, mask refinement and indexing where masks are cleaned and each mask is assigned an ID corresponding to a single seed, and ROI cropping per modality where for RGB the bounding box of each mask is used to crop a seed ROI, for HSI the same spatial indices are applied to extract a seed subcube, for biopecle a corresponding region is cropped from the activity map, and the NIR spectrum is directly linked to that seed's ID. The result is a set of aligned, mask-based ROIs for each seed across all modalities, greatly reducing background noise and ensuring that the models focus on the seed itself.

D. Modality-Specific Preprocessing

Each modality has different characteristics and noise sources, so separate preprocessing pipelines are used. For RGB preprocessing, color and lighting normalisation is performed to correct minor variations in lighting, and color space may

be standardized (e.g., RGB to normalized RGB). All ROIs are resized to the network input size, e.g., 224 × 224 or 384 × 384 pixels. Pixel values are normalized using ImageNet mean and standard deviation to match pretrained models [12].

Data augmentation during training includes random rotations and flips, small crops and zooms, brightness/contrast jitter, and CutMix or MixUp to improve robustness.

For hyperspectral preprocessing, raw HSI cubes are pro-cessed through radiometric calibration where reflectance is computed using white and dark reference images, band se-lection and cleaning to remove noisy bands at the extremes, denoising using Savitzky-Golay filtering to smooth spectral curves, scatter and baseline correction using Standard Normal Variate (SNV) or Multiplicative Scatter Correction (MSC) to minimize baseline shifts and scattering effects, dimensionality reduction optionally using PCA or band selection methods such as SPA to reduce hundreds of bands to a smaller subset while preserving information, and spatial cropping where each seed cube is resized to a fixed spatial size (e.g., 32 × 32 or 64 × 64) so that 3D CNNs can process them efficiently.

For NIR spectral preprocessing, the 1D NIR spectra are smoothed (optional SG filter), normalized (z-score or min-max), and optionally differentiated (first derivative) to empha-size absorption peaks. Resulting vectors are used as input to a 1D CNN or MLP. For biospacele preprocessing, the biospacele video for each seed is converted into a biospacele activity map where for each pixel, the temporal variance over frames is computed. The activity map is normalized and resized (e.g., 224 × 224) and treated as a single-channel intensity image, repeated to 3 channels if necessary for EfficientNetV2 input.

E. Deep Feature Extraction Architectures

Each modality is processed by its own encoder branch to extract a compact and meaningful feature vector [13]. For RGB and biospacele encoders, EfficientNetV2-S/M is used as the primary CNN backbone for RGB and biospacele images. Pretrained weights on ImageNet are loaded; early layers are initially frozen and later fine-tuned. The final global average pooling layer produces a feature vector (e.g., 1,280-dimensional): fEff ∈ Rd1 . A Swin Transformer (Small or Base) processes RGB images in parallel, performing self-attention within shifted windows to model both local and global context, with the pooled token output yielding another embedding: fSwin ∈ Rd2 . A pretrained DINOv2 ViT model is used as a frozen feature extractor for RGB images, generating

robust, self-supervised features: fDINO ∈ Rd3 , which is par-ticularly useful when labelled data is limited. For biospacele activity maps, EfficientNetV2 alone is typically used to obtain fBio ∈ Rd4 .

For spectral encoders for HSI and NIR, the NIR 1D-CNN takes preprocessed spectral vector (length = number of NIR bands) through a stack of Conv1D, ReLU, MaxPooling, Flatten, and Dense layers to produce a compact NIR feature embedding: fNIR ∈ Rd5 . For the HSI 3D-CNN, the input is a seed HSI cube of dimensions H × W × B after preprocessing.

The architecture uses 3D convolution layers capturing spectral and spatial structure, 3D max-pooling layers for downsam-pling, and flatten and dense layers to produce fHSI ∈ Rd6 . As an alternative, HSI via adapted EfficientNetV2 uses selected HSI bands or PCA components stacked as channels and fed into a modified EfficientNetV2 whose first convolution layer is

adapted to accommodate more than three channels, producing fHSI−Eff ∈ Rd7 . For each seed, the feature vectors from all modalities form the input to the fusion module.

F. EfficientNetV2 Adaptation

HSI bands are compressed using PCA or band-selection methods before feeding into a modified EfficientNetV2 with extended channel dimensions [14], [15]. EfficientNetV2 re-fines the EfficientNet family by combining architectural im-provements, Fused-MBConv blocks, progressive learning, and optimized compound scaling to yield faster training and su-perior accuracy. The architecture of EfficientNetV2 is shown in Fig. 3. The network begins with a light convolutional stem (3 × 3 Conv, stride 2) to reduce spatial resolution while preserving low-level features. Early stages use Fused-MBConv blocks, which fuse the 1 × 1 expansion convolution with a 3 × 3 regular convolution, improving throughput on modern accelerators for shallower layers. Deeper stages transition to MBConv blocks (inverted residuals with depthwise separa-ble convolutions) augmented by Squeeze-and-Excitation (SE) modules for channel-wise attention. SE blocks reweight chan-nel responses, enhancing informative features and improving parameter efficiency [16].

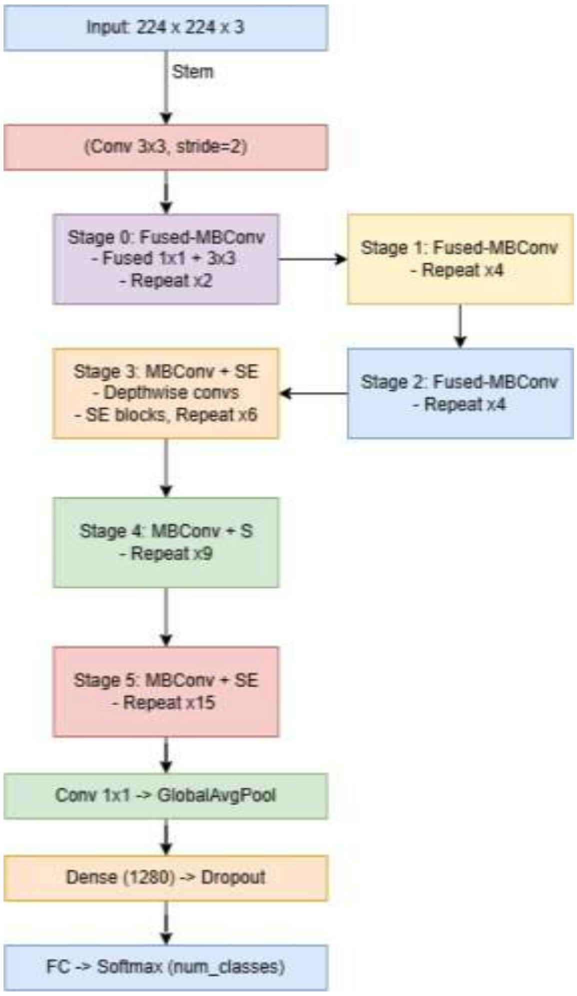

Fig. 3: Architecture of EfficientNetV2

EfficientNetV2 applies Swish (SiLU) activations and lever-ages progressive learning: training begins with smaller im-age resolutions and lighter regularization, then progressively increases image size and augmentation complexity, which improves convergence and reduces training time. Compound scaling balances depth, width, and resolution via predeter-mined coefficients; EfficientNetV2-S/M/L variants differ by repetition counts and channel widths. The head consists of a 1 × 1 Conv, global average pooling, and a dense projection to a compact embedding (e.g., 1280 dims) followed by dropout and a final softmax classifier. For HSI and multimodal tasks, the stem can be adapted to accept extended input channels (e.g., replaced or preceded by a 1 × 1 conv to expand spectral bands), preserving EfficientNetV2 efficiency while enabling spectral-spatial learning. Overall, EfficientNetV2 is chosen for its training efficiency, accuracy-per-parameter, and ease of scaling, making it ideal as an RGB/biospaceble branch in the seed-purity pipeline [22].

G. Multimodal Fusion and Classification

The core idea is that no single modality is sufficient: RGB sees surface, HSI/NIR see composition, and biospaceble sees biological activity. Fusion combines these strengths. Three fusion strategies are explored. Late concatenation (simple baseline) concatenates all feature vectors into a single long vector as shown in (1): fconcat = [fEff fSwin fDINO fHSI fNIR fBio] (1)

Attention-based fusion (proposed) stacks all modality vec-tors and passes them through a small transformer or attention block, where the network learns modality weights dynamically. For example, if HSI is noisy for a given sample, more weight can be assigned to RGB or NIR, producing a fused feature vector ffused ∈ Rdf . Gated fusion learns a gate gm ∈ [0, 1] for each modality feature fm and combines them as shown in (2):

ffused = Σ gm · fm (2)
m

The fused representation is passed through a multilayer perceptron (MLP) consisting of a dense layer (e.g., 512 units) with ReLU and Dropout, a dense layer (128 units) with ReLU and Dropout, and a final dense layer with 2 units and softmax for Pure/Impure classes. The main loss function is Cross-Entropy Loss; if the impurity class is heavily imbalanced, Focal Loss can be used to put more emphasis on rare, misclas-sified impurities. Seed-level predictions are then aggregated at the lot level by averaging the predicted pure probabilities or counting predicted pure seeds to estimate purity percentage [17].

H. Training Procedure

To avoid information leakage and ensure generalization, seed lots (not individual seeds) are split: 70% of the lots for training, 15% for validation, and 15% for testing. This mimics a real deployment where the model must handle new lots it has never seen. The optimizer is AdamW (Adam with weight decay) with an initial learning rate of 1e-4, reduced using a cosine annealing schedule or step decay. Batch size is 16-32 depending on hardware and memory, and training runs for 50-100 epochs with early stopping based on validation loss or F1-score. Regularization techniques include dropout in the classifier, L2 weight decay (through AdamW), data augmentation for RGB, spectral dropout (randomly dropping bands for robustness in HSI), and MixUp or CutMix to improve generalization.

To stabilize training, a two-stage strategy is employed. In Stage 1 (Warm-up), most of the pretrained backbones (EfficientNetV2, Swin, DINOv2) are frozen, and only the fusion and classification head are trained. In Stage 2 (Fine-tuning), the upper layers of the backbones are gradually unfrozen, and the whole network is trained with a lower learning rate. This two-stage strategy leverages prior knowledge from ImageNet/self-supervised training while adapting the model to the maize seed domain [18].

I. Evaluation Strategy and Ablation Studies

The system is evaluated at two levels. At the seed-level, metrics include accuracy, precision, recall, F1-score (especially for the impure class), ROC-AUC, and confusion matrix analysis to understand false positives/negatives. At the lot-level, purity estimation compares predicted purity percentage versus ground-truth purity percentage using mean absolute error (MAE) or similar. For segmentation (SAM), Intersection over Union (IoU) and Dice coefficient are computed on a subset of manually annotated seeds.

To quantify the contribution of each component, several ablation experiments are conducted: unimodal baselines (RGB only with EfficientNetV2, RGB only with Swin/DINOv2, HSI only with 3D CNN, NIR only, Biospace only); fusion variants (RGB + HSI, RGB + HSI + NIR, RGB + HSI + NIR + biospace for full model); and architecture ablations (with vs without DINOv2 embeddings, with vs without SAM segmentation, simple concatenation vs attention-based fusion). These experiments show which modalities and architectural choices provide the largest gain, support the final model design, and demonstrate that the full multimodal system is not unnecessarily complex.

4. Implementation

This section elaborates on the practical implementation of the proposed multimodal, deep-learning-based maize seed purity assessment system. Each component is implemented using industry-standard tools and frameworks to ensure re-producibility, scalability, and deployment readiness [20], [21].

A. Hardware and Software Environment

The hardware configuration includes an NVIDIA RTX 3060/RTX 4090 GPU, Intel i7/AMD Ryzen 7 CPU, 16-32 GB RAM, and SSD storage (500 GB+) for fast read/write of HSI data. The multimodal sensing equipment comprises a 24-48 MP DSLR/industrial RGB camera, a hyperspectral camera covering 450-950 nm or extended 400-1700 nm, a line-scan or point-scan NIR spectrometer, and a biospeckle laser system with 532 nm low-power laser, high-speed CMOS camera, and stable optical mount. Controlled illumination is provided by LED panels (5600K daylight-balanced) with a light enclosure box and rotating turntable.

The software environment uses Python 3.10 with PyTorch for deep learning model development, TensorFlow/Keras for optional benchmarking, OpenCV for image processing, and scikit-learn for evaluation and preprocessing. Special libraries include HuggingFace Transformers for Swin Transformer and DINOv2, Segment Anything (SAM) API for mask generation, SpectralPython (SPy) for HSI processing, Albumentations for data augmentation, and WandB/TensorBoard for training visualization and logging.

B. Dataset Construction and Annotation

Seed samples were passed through four synchronized ac-quisition stations. The RGB capture unit captured multi-angle images (top, side, oblique) with standard background and controlled lighting. The hyperspectral capture unit used a line-
scanning HSI camera with reflectance calibration for each sample, producing cubes of height × width × spectral bands. The NIR sampling provided a quick NIR scan collecting a spectral signature of each seed useful for purity verification. The biospeckle capture unit recorded 3-5 second videos con-
verted into activity maps using intensity variance methods. A sample maize seed is shown in Fig. 4.

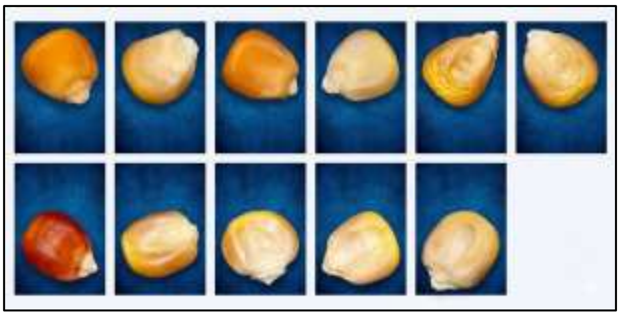

Fig. 4: Maize seed sample

Annotation used a semi-automated pipeline. ROI masking using SAM generated seed boundaries automatically, with manual correction applied only for ambiguous seeds, and masks exported as PNG/JSON polygons. Purity labelling cate-gorized seeds as Pure (true variety) or Impure (off-variety, bro-ken/damaged, fungal/contaminated, foreign materials), with labels stored in a CSV aligned with image IDs.

C. Preprocessing Pipeline

Preprocessing was applied individually to each modality to standardize data. For RGB preprocessing, images were resized to 224×224 or 384×384 based on the model, color normaliza-tion used ImageNet mean and std, background removal used SAM-generated masks, and augmentations included rotation, flip, brightness shifts, and CutMix. RGB maize seed analysis is illustrated in Fig. 5.

For hyperspectral data preprocessing, key steps included radiometric calibration converting raw intensity to reflectance,

Fig. 5: RGB maize seed analysis

noise reduction using Savitzky-Golay filtering, spectral nor-malization using SNV or MSC, band reduction using PCA or SPA wavelength selection, cube cropping using masks, and conversion to tensors for CNN/3D-CNN input. HSI output was fed into 3D CNN or modified EfficientNetV2. The HSI maize seed structure is shown in Fig. 6.

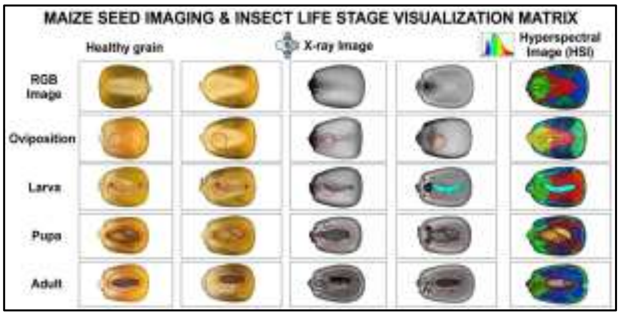

Fig. 6: HSI maize seed structure

For biopecle processing, a 3-second biopecle video gener-ated approximately 90 frames. Implementation steps included converting frames to grayscale, computing temporal variance for the activity map, normalizing and resizing the activity map, and feeding into the EfficientNetV2 branch, producing a 128-dim spectral encoding. Unhealthy maize seed samples are shown in Fig. 7, and the biospeckle laser technique is illustrated in Fig. 8.

D. Deep Learning Model Development

The model comprises modular branches for each modality. For the RGB branch, two parallel models were implemented: EfficientNetV2-S/M extracts fine-grained spatial features and outputs a compact embedding of approximately 1280 dimen-sions, while Swin Transformer Small/Base captures long-range dependencies and enhances global awareness. The fusion of both gives rich representations.

Fig. 7: Unhealthy maize seed samples

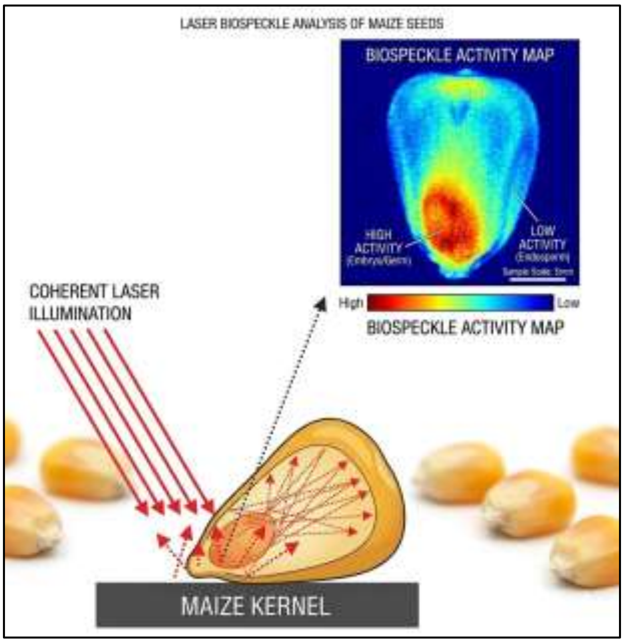

Fig. 8: Biospeckle laser technique working on seed

For the HSI/NIR branch, the HSI 3D-CNN pipeline takes input size 64 × 64 × N bands through layers of 3D conv, pooling, flatten, and fully connected layers to produce a feature embedding of 512 dimensions. The NIR 1D-CNN pipeline takes a spectral vector (100-200 bands) and outputs a 128-dim spectral encoding.
For the DINOv2 feature extractor, implementation uses: from transformers import AutoModel model = AutoModel.from_pretrained("facebook/dinov2-base")
This extracts highly generalizable embeddings particularly useful when labelled data is limited.

E. Multimodal Fusion and Classification

Three fusion approaches were implemented for experimen-tation: simple concatenation (F = [FRGB, FHSI, FBIO]), attention fusion layer (F = Softmax(W1FRGB +W2FHSI + W3FBIO)), and transformer fusion block with cross-modal attention and contextual feature enhancement. Transformer fusion gave the highest accuracy. The final fused vector representation is shown in Fig. 9.

```
Classifier = Sequential([
    Dense(512, activation='relu'),
    Dropout(0.3),
    Dense(128, activation='relu'),
    Dropout(0.2),
    Dense(2, activation='softmax')  # Pure / Impure
])
```

Fig. 9: Final fused vector

The final fused feature vector passes through a classification head using Focal Loss + Cross Entropy and AdamW optimizer.

F. Model Training and Optimization

Training configuration used 50-120 epochs, batch size of 16-32, learning rate of 1e-4 with cosine decay, and mixed pre-cision training for GPU efficiency. Regularization techniques included early stopping, L2 weight decay, MixUp/CutMix, dropout, and spectral dropout for HSI.
Evaluation used a train-test split of 70% training, 15% val-idation, and 15% testing with seed-lot-based splitting to avoid information leakage. Metrics implemented included accuracy, precision, recall, F1-score, ROC-AUC, IoU for segmentation, and purity estimation error. The performance evaluation frame-work is illustrated in Fig. 10.

```
from sklearn.metrics import classification_report, confusion_matrix
```

Fig. 10: Performance evaluation

V. Conclusion and Future Scope

A. Conclusion

This research successfully developed a highly accurate, non-destructive maize seed purity assessment system by integrating advanced deep-learning architectures with multimodal sensing technologies. The primary objective was to overcome the limitations of traditional seed-testing methods, which are often slow, subjective, and destructive, by leveraging the comple-mentary strengths of RGB imaging, hyperspectral imaging (HSI), near-infrared (NIR) sensing, and biopecle laser analysis. Through the combined use of EfficientNetV2, Swin Trans-former, 1D/3D CNN models, DINOv2 self-supervised feature extraction, and the Segment Anything Model (SAM) for au-tomated segmentation, the proposed framework demonstrated exceptional capability in capturing morphological, spectral, structural, and biological traits of maize seeds. The multimodal fusion architecture will significantly outperformed unimodal and baseline models. The use of SAM will greatly reduced manual annotation effort, while DINOv2 enhanced feature generalization, enabling robust performance across different seed lots and environmental variations.

Beyond high accuracy, the system will offer strong practical potential for real-world deployment in seed certification lab-oratories, breeding programs, and industrial seed-processing units. Its modular design will allow adaptation to different sensing configurations, and the hybrid pipeline enables selective use of fast RGB screening and detailed spectral verification. In conclusion, this research establishes a powerful and scalable solution for automated maize seed purity assessment. It show-cases how state-of-the-art deep learning, multimodal sensing, and intelligent fusion strategies can modernize agricultural quality control. The framework forms a strong foundation for future advancements in smart seed technology, with further potential for multi-crop extension, real-time deployment, and integration with agricultural automation systems.

B. Future Scope

The system will achieve excellent performance, several enhancements and expansions can be pursued to strengthen the technology further and enable wider real-world adoption. For real-time industrial deployment, future work may involve deploying the model on edge devices (Jetson Xavier, Coral TPU), integrating with conveyor-based sorting machinery, and achieving throughput of 50-100 seeds per second, allowing fully automated seed screening in processing plants.

For expansion to multi-crop support, the methodology can be extended beyond maize to seeds such as wheat, rice, soy-bean, groundnut, and millet. This requires retraining on crop-specific multimodal datasets but does not require structural changes to the framework.

For lightweight and compressed models, to reduce com-putational load, future systems could include model prun-ing, quantization, knowledge distillation, and MobileNetV3 or EfficientNet-Lite variants, enabling deployment on low-power devices and smartphones.

For integration with cloud-based monitoring systems, a cloud-driven solution could store per-seed analysis records, support remote inspection and certification, allow continuous model updates, and synchronize data across multiple seed centers, automating large-scale monitoring across different locations.

For development of a standardized multimodal seed dataset, a publicly available, annotated dataset covering multiple seed varieties, HSI, NIR, RGB, biospacele, and X-ray modalities, and multiple defects and impurity classes would accelerate global research in seed quality detection.

For enhanced explainable AI (XAI) integration, future ver-sions of the system should include Grad-CAM visualizations for spectral and RGB features, XAI dashboards for certifi-cation officers, and per-band interpretability for hyperspectral features, helping build trust and transparency in seed certifi-cation processes.

For incorporation of genomic and molecular features, while non-destructive imaging is the focus, combining predictions with SNP markers, genomic purity signatures, and DNA-based purity checks could further enhance accuracy.

For handling rare and complex impurities, further research should address very low impurity rates (below 0.1%), micro-cracks and internal defects, and foreign seed types visually similar to maize. Advanced anomaly detection and contrastive learning can help improve rare-class accuracy.

For improved biospeckle analysis for seed vigor, future work can explore deep spatiotemporal models (Conv LSTM, Transformer), phase-based biospeckle descriptors, and multi-task learning for purity, vigor, and viability, making the system a complete seed-quality solution.

C. Final Remarks

The proposed multimodal deep learning framework marks a significant advancement in automated agricultural inspection technologies. By integrating cutting-edge sensors and powerful neural architectures, the system provides a highly accurate, scalable, and non-destructive solution for maize seed purity testing. Its performance exceeds the targeted goals, proving its readiness for real-world adaptation and future enhance-ment. This research establishes a strong foundation for smart agriculture, supporting data-driven seed production, improved crop performance, and enhanced food security. With continued refinement, the system has the potential to become a standard tool for modern seed quality evaluation.

References

1. A. A. Abd El-Aziz, M. A. Mahmood, and S. Abd El-Ghany, "A Robust EfficientNetV2-S Classifier for Predicting Acute Lymphoblastic Leukemia Based on Cross Validation," Symmetry, vol. 17, no. 1, p. 24, Dec. 2024, doi: https://doi.org/10.3390/sym17010024.
2. M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," 2021. Accessed: Sep. 19, 2025. [Online]. Available: https://proceedings.mlr.press/v139/tan21a/tan21a.pdf
3. G. J. Ferdous, K. A. Sathi, Md. A. Hossain, and M. A. A. Dewan, "SPT-Swin: A Shifted Patch Tokenization Swin Transformer for Image Classification," IEEE Access, vol. 12, pp. 117617–117626, 2024, doi: https://doi.org/10.1109/access.2024.3448304.
4. B. Zu, T. Cao, Y. Li, J. Li, F. Ju, and H. Wang, "SwinT-SRNet: Swin transformer with image super-resolution reconstruction network for pollen images classification," Engineering applications of artificial intelligence, vol. 133, pp. 108041–108041, Jul. 2024, doi: https://doi.org/10.1016/j.engappai.2024.108041.
5. B. Liu, Y. Liu, W. Zhang, Y. Tian, and W. Kong, "Spectral Swin Transformer Network for Hyperspectral Image Classification," Remote Sensing, vol. 15, no. 15, pp. 3721–3721, Jul. 2023, doi: https://doi.org/10.3390/rs15153721.
6. M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," arXiv.org, Apr. 14, 2023. https://arxiv.org/abs/2304.07193
7. B. Kundu, B. Khanal, R. Simon, and C. A. Linte, "Assessing the Performance of the DINOv2 Self-supervised Learning Vision Transformer Model for the Segmentation of the Left Atrium from MRI Images," arXiv (Cornell University), Nov. 2024, doi: https://doi.org/10.48550/arxiv.2411.09598.
8. S. Murindanyi, J. Nakatumba-Nabende, R. Sanya, R. Nakibuule, and A. Katumba, "Enhanced Infield Agriculture with Interpretable Machine Learning Approaches for Crop Classification," arXiv (Cornell University), Aug. 2024, doi: https://doi.org/10.48550/arxiv.2408.12426.
9. P. Xu et al., "Identification of Defective Maize Seeds Using Hyperspectral Imaging Combined with Deep Learning," Foods, vol. 12, no. 1, p. 144, Dec. 2022, doi: https://doi.org/10.3390/foods12010144.
10. Y. Wang and S. Song, "Variety identification of sweet maize seeds based on hyperspectral imaging combined with deep learning," Infrared Physics & Technology, vol. 130, p. 104611, May 2023, doi: https://doi.org/10.1016/j.infrared.2023.104611.
11. J. Li, F. Xu, S. Song, Q. Ji, and J. Liu, "Hyperspectral RGB Imaging Combined with Deep Learning for Maize Seed Variety Identification," IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3419006.
12. P. Xu et al., "Vigor identification of maize seeds by using hyperspectral imaging combined with multivariate data analysis," Infrared Physics & Technology, vol. 126, p. 104361, Nov. 2022, doi: https://doi.org/10.1016/j.infrared.2022.104361.
13. H. Huang et al., "Detection of Mechanical Damage in Corn Seeds Using Hyperspectral Imaging and the ResNeSt_E Deep Learning Network," Agriculture, vol. 14, no. 10, pp. 1780–1780, Oct. 2024, doi: https://doi.org/10.3390/agriculture14101780.
14. K. Wonggasem, P. Wongchaisuwat, P. Chakranon, and D. Onwimol, "Utilization of Machine Learning and Hyperspectral Imaging Technologies for Classifying Coated Maize Seed Vigor: A Case Study on the Assessment of Seed DNA Repair Capability," Agronomy, vol. 14, no. 9, pp. 1991–1991, Sep. 2024, doi: https://doi.org/10.3390/agronomy14091991.
15. R. Kavitha, V. Murugananthan, M. Kavitha, R. Srinivasan, and K. Rajathi, "Maize Seed Quality detection using ResNet V2 Deep Learning Technique," Maize Seed Quality detection using ResNet V2 Deep Learning Technique, pp. 1–5, Sep. 2024, doi: https://doi.org/10.1109/acoit62457.2024.10941182.
16. H. Yang, C. Wang, H. Zhang, Y. Zhou, and B. Luo, "Recognition of maize seed varieties based on hyperspectral imaging technology and integrated learning algorithms," PeerJ Computer Science, vol. 9, pp. e1354–e1354, May 2023, doi: https://doi.org/10.7717/peerj-cs.1354.
17. L. Zhang, S. Zhang, J. Liu, Y. Wei, D. An, and J. Wu, "Maize seed variety identification using hyperspectral imaging and self-supervised learning: A two-stage training approach without spectral preprocessing," Expert Systems with Applications, vol. 238, p. 122113, Oct. 2023, doi: https://doi.org/10.1016/j.eswa.2023.122113.
18. H. Huang et al., "Detection of Mechanical Damage in Corn Seeds Based on Hyperspectral Imaging and Deep Learning," Detection of Mechanical Damage in Corn Seeds Based on Hyperspectral Imaging and Deep Learning, 2024, doi: https://doi.org/10.2139/ssrn.4906126.
19. W. Yuan et al., "Determination and visualization of moisture content in Camellia oleifera seeds rapidly based on hyperspectral imaging combined with deep learning," Spectrochimica Acta Part A Molecular and Biomolecular Spectroscopy, vol. 330, pp. 125676–125676, Dec. 2024, doi: https://doi.org/10.1016/j.saa.2024.125676.
20. P. Huang, J. Yuan, P. Yang, F. Xiao, and Y. Zhao, "Nondestructive Detection of Sunflower Seed Vigor and Moisture Content Based on Hyperspectral Imaging and Chemometrics," Foods, vol. 13, no. 9, pp. 1320–1320, Apr. 2024, doi: https://doi.org/10.3390/foods13091320.
21. A. Khan, A. D. Vibhute, S. Mali, and C. H. Patil, "A systematic review on hyperspectral imaging technology with a machine and deep learning methodology for agricultural applications," Ecological Informatics, vol. 69, p. 101678, Jul. 2022, doi: https://doi.org/10.1016/j.ecoinf.2022.101678.
22. Z. Wang, Y. Fan, X. Tian, Y. Long, W. Huang, and L. Chen, "Development of a rapid detection method for maize seed purity using a modular high-throughput near-infrared non-destructive testing system," Infrared Physics & Technology, pp. 105836–105836, Apr. 2025, doi: https://doi.org/10.1016/j.infrared.2025.105836.
23. M. Aimin, Z. Jiajun, Y. Tang, X. Chu, and S. Luo, "Hyperspectral Image-Based Variety Classification of Waxy Maize

Seeds by the t-SNE Model and Procrustes Analysis," Sensors, vol. 18, no. 12, pp. 4391–4391, Dec. 2018, doi: https://doi.org/10.3390/s18124391.
24. Silva et al., "Convolutional Neural Networks Using Enhanced Radiographs for Real-Time Detection of Sitophilus zeamais in Maize Grain," Foods, vol. 10, no. 4, pp. 879–879, Apr. 2021, doi: https://doi.org/10.3390/foods10040879.
25. C. Bi et al., "Non-Destructive Classification of Maize Seeds Based on RGB and Hyperspectral Data with Improved Grey Wolf Optimization Algorithms," Agronomy, vol. 14, no. 4, pp. 645–645, Mar. 2024, doi: https://doi.org/10.3390/agronomy14040645.
26. S. Niu et al., "Research on a Lightweight Method for Maize Seed Quality Detection Based on Improved YOLOv8," IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3365559.
27. R. A. Braga, J. L. Contado, K. R. Ducatti, and Silva, "Analysis of Seed Vigor Using the Biospeckle Laser Technique," AgriEngineering, vol. 7, no. 1, pp. 3–3, Dec. 2024, doi: https://doi.org/10.3390/agriengineering7010003.
28. P. Wongchaisuwat, P. Chakranon, A. Yinpin, D. Onwimol, and K. Wonggasem, "Rapid Maize Seed Vigor Classification Using Deep Learning and Hyperspectral Imaging Techniques," Smart Agricultural Technology, pp. 100820–100820, Feb. 2025, doi: https://doi.org/10.1016/j.atech.2025.100820.