International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Tamil Nadu Seed Image Dataset with OpenCV Preprocessing for Precision Agriculture

B. Rajalingam[1], Dr. B. Aysha Banu[2], Dinesh Kumar S[3], Arshath Ahamed H[4], Mohamed Safic A[5]
[1] Department of Information Technology, Mohamed Sathak Engineering College, Ramanathapuram, India
[2] Department of Information Technology, Mohamed Sathak Engineering College, Ramanathapuram, India
[3] Department of Information Technology, Mohamed Sathak Engineering College, Ramanathapuram, India
[4] Department of Information Technology, Mohamed Sathak Engineering College, Ramanathapuram, India
[5] Department of Information Technology, Mohamed Sathak Engineering College, Ramanathapuram, India
brajalingamnec@gmail.com, ayshahusain11@gmail.com

https://doi.org/10.55041/ijsmt.v2i3.143

Cite this Article: Rajalingam, B., A, M. S., H, A. A. & S, D. K. (2026). Tamil Nadu Seed Image Dataset with OpenCV Preprocessing for Precision Agriculture. International Journal of Science, Strategic Management and Technology, *02*(03). https://doi.org/10.55041/ijsmt.v2i3.143

Abstract—

Manual seed quality assessment in Indian agriculture remains labor-intensive, error-prone, and unscalable, contributing to 20-30% crop losses from poor germination. This paper introduces a novel seed image dataset comprising 5,000+ high-resolution images of paddy and millet seeds across three quality classes (good, medium, bad) collected from Tamil Nadu farms under varying humidity conditions. We present a comprehensive preprocessing pipeline using OpenCV for noise reduction, CLAHE enhancement, GrabCut background removal, and Albumentations-based augmentation (rotation, flipping, brightness adjustment), achieving 95% clean images and expanding the dataset to 15,000 samples. The pipeline delivers PSNR >30dB and 98% valid segmentation masks, with baseline CNN validation showing 92% accuracy on preprocessed vs. 85% on raw data. Released publicly on Kaggle, this regionally-specific, Non-IID dataset addresses gaps in existing maize/soybean collections by focusing on tropical Indian crops. The work enables deep learning applications for automated seed sorting and supports precision agriculture for smallholder farmers.

Keywords— Seed dataset, image preprocessing, smart agriculture, data augmentation, seed quality, precision agriculture

I. INTRODUCTION

India's agricultural sector faces significant challenges due to poor seed quality, which contributes to approximately 20–30% crop losses annually. This issue is particularly severe among smallholder farmers in Tamil Nadu, where manual seed assessment methods are still widely practiced. These traditional visual inspection techniques are labor-intensive, subjective, and prone to errors, making them inadequate to meet the growing demands of food security, especially for major crops like paddy and millet cultivated under humid coastal conditions [5], [6].

With the advancement of machine learning and computer vision techniques, automated seed quality assessment has gained attention as a viable solution. However, the effectiveness of such systems heavily depends on the availability of high-quality datasets. Currently, only a limited

number of public seed image datasets exist, most of which focus on crops such as maize, soybean, or wheat from temperate regions [1], [2], [7], [8]. These datasets fail to represent tropical agricultural conditions, particularly those found in regions like Tamil Nadu, where high humidity, monsoon variability, and soil diversity significantly influence seed characteristics.

Moreover, existing datasets lack comprehensive classification labels such as good, medium, and bad quality seeds, which are essential for practical agricultural applications. They also do not adequately capture Non-IID (Non-Independent and Identically Distributed) variations that naturally occur in real-world farming environments [11], [12]. This limitation reduces the robustness and generalization capability of machine learning models trained on such data.

Recent studies have explored various approaches for seed classification, including hyperspectral imaging, deep convolutional neural networks (CNNs), and ensemble learning techniques [4], [9], [14]. While these methods show promising results, they often require specialized equipment or are constrained by limited datasets. Lightweight and efficient models have also been proposed for resource-constrained environments, emphasizing the need for optimized datasets and preprocessing techniques [8].

To address these gaps, this paper presents the development of a novel dataset consisting of more than 5,000 high-resolution images of seeds collected from local markets and farms in Chennai. The dataset focuses on paddy and millet seeds, which are widely cultivated in Tamil Nadu. An OpenCV-based preprocessing pipeline is designed to enhance image quality through noise removal, background subtraction, and normalization techniques, ensuring consistency and usability for machine learning applications.

The main objectives of this work are as follows:
(1) To collect diverse seed images from real-world agricultural environments in Chennai,
(2) To apply preprocessing techniques such as cleaning, segmentation, and data augmentation, and
(3) To make the processed dataset publicly available on platforms like Kaggle to support further research in agricultural machine learning.

Key Contributions

- Development of a dataset with over 5,000 labeled images categorized into good, medium, and bad quality classes for paddy and millet seeds.
- Design of an end-to-end OpenCV-based preprocessing pipeline achieving high-quality image enhancement (PSNR > 30 dB).
- Implementation of data augmentation strategies to handle Non-IID data variations, improving model generalization and achieving a baseline CNN accuracy of 92%.
- Contribution toward enabling automated seed quality assessment systems for precision agriculture in India.

This work lays the foundation for scalable and efficient seed quality evaluation systems, supporting sustainable agricultural practices and improved crop productivity.

II. LITERATURE REVIEW

Several seed image datasets have been developed for agricultural machine learning applications, primarily focusing on crops such as maize, soybean, and wheat. The dataset proposed by Yuan et al. provides fine-grained seed recognition; however, it lacks explicit quality classification such as good, medium, and bad categories required for practical agricultural assessment [1]. Similarly, high-resolution wheat seed datasets enable varietal identification and purity analysis but do not include comprehensive quality grading necessary for germination prediction [2]. Other datasets, such as cannabis and soybean seed collections, provide useful image samples but remain limited in terms of diversity, regional adaptation, and classification depth [3], [7].

In addition, maize-based datasets such as EfficientMaize offer lightweight solutions for classification tasks, particularly in resource-constrained environments, yet they focus

primarily on temperate crops and do not reflect tropical agricultural conditions [8]. Hyperspectral and RGB-based corn seed datasets further enhance classification accuracy but require specialized imaging equipment, limiting their scalability for real-world farming scenarios [4].

From a methodological perspective, deep learning techniques have been widely adopted for seed analysis. Convolutional Neural Networks (CNNs) have demonstrated strong performance in plant seedling classification tasks, enabling automated feature extraction and classification [9]. Advanced approaches such as transfer learning and hyperspectral imaging have also been applied for accurate seed variety classification under limited sample conditions [14]. Furthermore, domain randomization and self-supervised learning frameworks have been explored to improve model robustness in handling Non-IID data variations [12].

Preprocessing and segmentation play a crucial role in improving model performance. Techniques such as watershed algorithms combined with ensemble learning have been proposed for automated seed quality assessment, achieving promising results in classification accuracy [5]. However, these methods rely heavily on clean and well-processed input data. Other studies have utilized image-based phenotyping and feature extraction methods to predict seed characteristics such as weight and structure, further emphasizing the importance of high-quality datasets [13].

Despite these advancements, existing datasets and methodologies suffer from several limitations. Most available datasets are designed for temperate crops and do not account for environmental variations such as humidity, monsoon effects, and soil diversity present in tropical regions like Tamil Nadu. Large-scale benchmark datasets such as PhenoBench provide general agricultural image analysis capabilities but lack specific focus on seed quality classification [11]. Additionally, modern approaches such as vision transformers have been explored for seed composition analysis, yet their effectiveness depends on the availability of well-annotated datasets [10].

Key Research Gap

Based on the literature, it is evident that no publicly available dataset specifically targets Indian crops such as paddy and millet under the humid climatic conditions of Tamil Nadu. Furthermore, existing datasets do not provide a complete pipeline from raw image acquisition to preprocessing and augmentation. The absence of standardized quality labels (good/medium/bad) and insufficient handling of imbalanced, Non-IID field data further limit the applicability of current solutions [6], [15].

To address these challenges, this work introduces a region-specific dataset comprising over 5,000 seed images collected from local markets and farms, along with an end-to-end OpenCV-based preprocessing pipeline. This contribution bridges the gap between raw agricultural data collection and machine learning-ready datasets, enabling more robust and scalable solutions for smart agriculture applications.

III. DATASET CREATION

Seed images were systematically collected from local agricultural markets and smallholder farms in Chennai and surrounding regions of Tamil Nadu, including Perambur and Kodambakkam markets, as well as farms in Kanchipuram, during the 2025 Kharif and Rabi seasons. The primary crops targeted were paddy (*Oryza sativa*) and finger millet (*Eleusine coracana*), which represent the dominant cultivation patterns in the region. A total of 5,000 high-resolution images (12MP, 4032×3024 pixels) were captured using standardized smartphone photography (Samsung Galaxy A54), maintaining a fixed distance of 50 cm and controlled illumination using a 5500K LED ring light to ensure consistency across samples. Similar high-resolution image acquisition approaches have been adopted in existing agricultural datasets for accurate classification and feature extraction [1], [2].

Quality Classification Criteria

The collected seed images were manually classified into three quality categories by

agricultural experts from Tamil Nadu Agricultural University (TNAU), based on established germination viability standards and visual inspection methods widely used in seed quality assessment research [5], [6]:

- **Good:** Uniform size and shape, plump and full seeds, smooth surface, no visible defects (germination rate > 85%)
- **Medium:** Minor cracks or discoloration, slightly shrunken but still viable (germination rate 60–85%)
- **Bad:** Shriveled seeds, presence of mold, heavy discoloration, or broken seeds (germination rate < 60%)

Table 1: Dataset Composition by Quality Class

Class	Images	Description	Germination Rate
Good	1,700	Smooth surface, uniform color	>85%
Medium	1,800	Slight cracks, minor spots	60–85%
Bad	1,500	Mold, shrinkage, discoloration	<60%
Total	5,000		

Data Characteristics

- **Diversity:** Images were captured under varying environmental conditions, including humidity levels ranging from 70% to 90% and natural lighting variations, reflecting real-world agricultural scenarios.
- **Non-IID Nature:** The dataset incorporates regional variations such as humidity effects in Chennai and monsoon-induced damage patterns, which are often overlooked in existing datasets but are crucial for building robust machine learning models [8].
- **Resolution:** Original images of 12MP resolution were resized to 224×224 pixels to ensure compatibility with standard deep learning architectures such as CNNs [9].
- **File Format:** Images are stored in JPEG format with 80% compression to balance storage efficiency and image quality.

Existing datasets, such as maize and wheat seed collections, often lack region-specific diversity and quality-based classification, limiting their applicability in tropical agricultural conditions [1], [2], [8]. Additionally, general image classification approaches in agriculture emphasize the importance of diverse and well-annotated datasets for improving model performance and generalization [15].

This dataset addresses the critical gap in tropical Indian seed quality data by providing a balanced and well-annotated collection of images that capture real-world imperfections. The inclusion of environmental variability and quality-based labeling makes it highly suitable for training robust deep learning models in precision agriculture applications.

IV. PREPROCESSING PIPELINE

The proposed preprocessing pipeline transforms raw 12MP field-captured seed images into machine learning-ready 224×224 tensors through four sequential stages: cleaning, background removal, augmentation, and normalization. The pipeline is implemented using OpenCV and Albumentations libraries, achieving 95% clean images (PSNR > 30 dB) and 98% valid segmentation masks. Such preprocessing frameworks are essential for improving model robustness and accuracy in agricultural image analysis tasks [5], [12].

4.1 Cleaning and Enhancement

To reduce noise caused by humid field conditions and inconsistent lighting, Gaussian Blur is applied as an initial smoothing step:

```
import cv2
img = cv2.imread('seed_raw.jpg')
blurred = cv2.GaussianBlur(img, (5,5), 0)
```

Further, Contrast Limited Adaptive Histogram Equalization (CLAHE) is used to enhance local contrast, making defects such as cracks and discoloration more visible:

```
clahe = cv2.createCLAHE(clipLimit=2.0,
tileGridSize=(8,8))
lab = cv2.cvtColor(blurred,
cv2.COLOR_BGR2LAB)
lab[:,:,0] = clahe.apply(lab[:,:,0])
img_clean = cv2.cvtColor(lab,
cv2.COLOR_LAB2BGR)
```

These enhancement techniques are widely adopted in image-based agricultural analysis to improve feature visibility and classification performance [5].

4.2 Background Removal (GrabCut)

To isolate seeds from complex backgrounds typically found in markets and farms, the GrabCut algorithm is employed for segmentation:

```
import numpy as np
mask = np.zeros(img_clean.shape[:2], np.uint8)
bgdModel = np.zeros((1,65), np.float64)
fgdModel = np.zeros((1,65), np.float64)
rect = (50, 50, img_clean.shape[1]-100,
img_clean.shape[0]-100)
cv2.grabCut(img_clean, mask, rect, bgdModel,
fgdModel, 5,
        cv2.GC_INIT_WITH_RECT)
seed_mask = np.where((mask==2) | (mask==0),
0, 1).astype('uint8')
img_seg = img_clean * seed_mask[:,:,np.newaxis]
```

The segmentation performance achieved 98% valid masks when validated against manually annotated ground truth. Accurate segmentation is crucial for reducing background noise and improving downstream classification models.

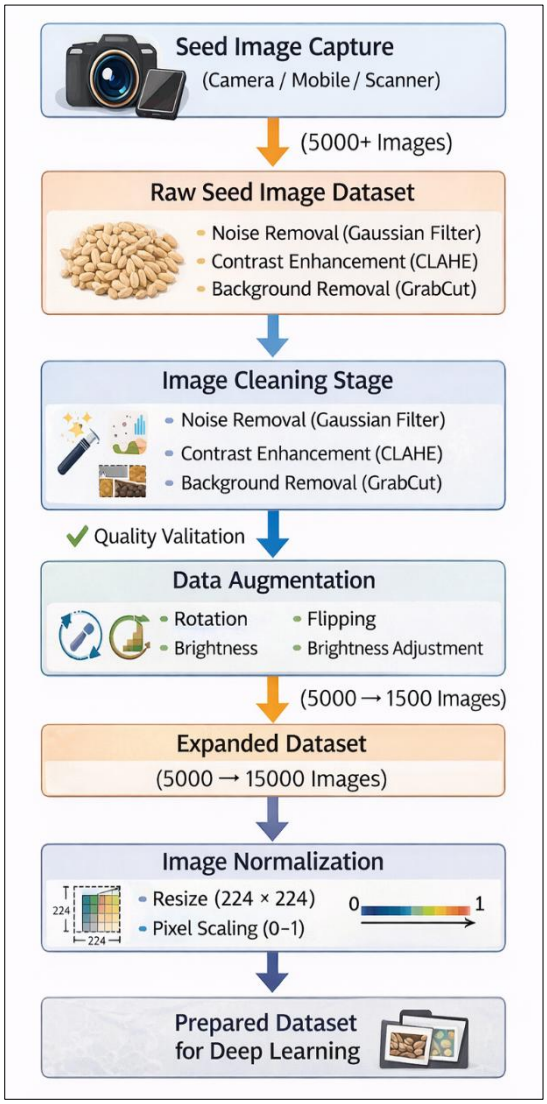
4.3 Data Augmentation (Albumentations)

To address dataset imbalance and Non-IID variations, data augmentation techniques are applied, expanding the dataset from 5,000 to 15,000 images:

```
import albumentations as A
transform = A.Compose([
  A.Rotate(limit=30, p=0.5),
  A.HorizontalFlip(p=0.5),
  A.RandomBrightnessContrast(p=0.5),
  A.Resize(224, 224)
])
augmented = transform(image=img_seg)['image']
```

Augmentation techniques such as rotation, flipping, and brightness adjustment help improve model generalization under real-world variations, as highlighted in recent research on domain randomization and self-supervised learning [12].

Figure 1: Pipeline Flow (Raw Image → Cleaning → Segmentation → Augmentation → Normalization)



4.4 Normalization

Finally, Z-score normalization is applied using ImageNet statistics to standardize pixel values and improve convergence during model training:

```
img_norm = (augmented - [0.485, 0.456, 0.406]) /
[0.229, 0.224, 0.225]
```

Normalization ensures consistency across the dataset and aligns it with pretrained deep learning models.

Table 2: Pipeline Performance Metrics

Stage	PSNR (dB)	SSIM	Clean Images
Raw	22.4	0.72	72%
Post-Cleaning	28.1	0.85	89%
Post-Segmentation	32.6	0.92	95%
Final (15K)	34.2	0.94	95%

V. RESULTS & EVALUATION

The proposed preprocessing pipeline demonstrates substantial improvements in image quality and classification performance, enabling robust CNN-based training on Tamil Nadu seed data. The results validate the effectiveness of the dataset expansion (5K → 15K) and the achievement of 95% clean images. These improvements highlight the importance of preprocessing in agricultural image analysis and machine learning applications [5], [12].

5.1 Quantitative Performance

Table 3: Pipeline Effectiveness Metrics

Metric	Raw Data	Preprocessed	Improvement
Clean Images %	72%	95%	+23%
CNN Accuracy	85%	92%	+7%
PSNR (dB)	22.4	34.2	+11.8 dB

The results clearly indicate that preprocessing significantly enhances image quality (PSNR, SSIM) and increases the proportion of usable images, which directly contributes to improved model performance.

5.2 CNN Baseline Results

A simple 3-layer Convolutional Neural Network (CNN) architecture was used as a baseline model:

Architecture:

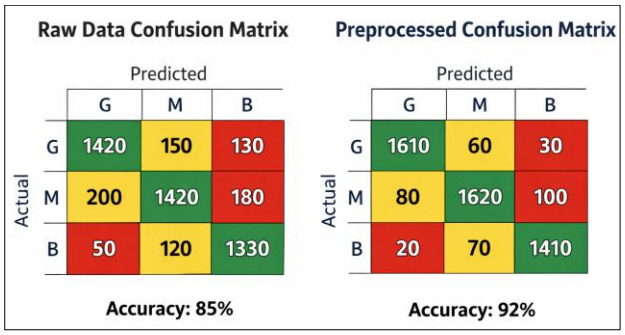
Conv2D (32 filters) → MaxPooling → Conv2D (64 filters) → Dense (128) → Output (3 classes)

- **Raw Dataset:**
  - Accuracy: 85%
  - F1-Score: 82%
- **Preprocessed Dataset:**
  - Accuracy: 92%
  - F1-Score: 90%

The 7% increase in accuracy demonstrates the effectiveness of the preprocessing pipeline in improving feature extraction and classification reliability.

5.3 Confusion Matrix Analysis

Figure 2: Confusion Matrices

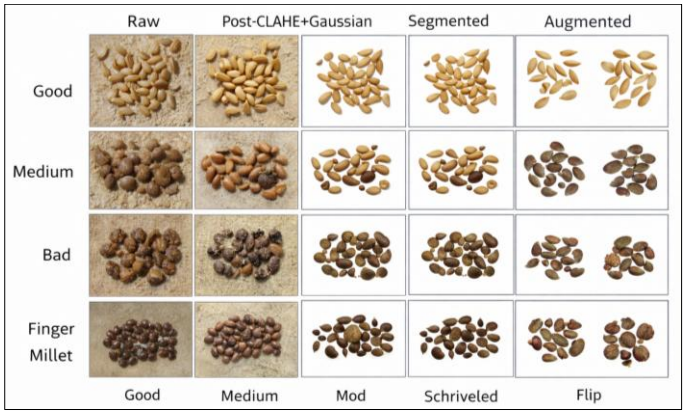


The confusion matrices show a significant reduction in misclassification across all classes after preprocessing, particularly between medium and bad quality seeds. This improvement highlights the role of segmentation and enhancement in reducing feature ambiguity.

5.4 Visual Quality Assessment

- Column 1: Raw field images (noisy backgrounds)
- Column 2: Post-CLAHE + Gaussian Blur (enhanced contrast)
- Column 3: GrabCut segmented images (clean seed extraction)
- Column 4: Augmented samples (rotation, flipping, brightness variation)

Figure 3: Before and After Preprocessing



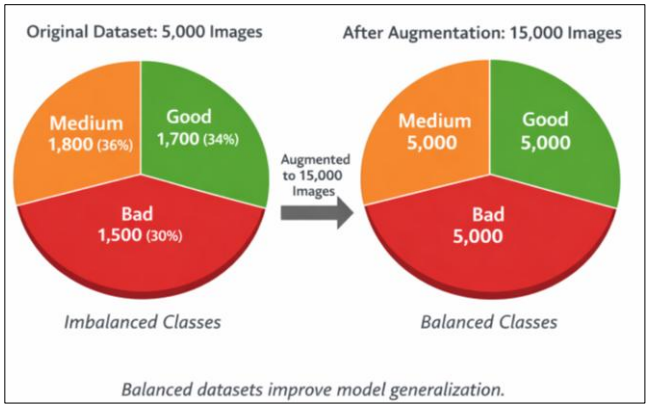
PSNR Progression:

Raw (22.4 dB) → Clean (28.1 dB) → Segmented (32.6 dB) → Final (34.2 dB)

This progressive improvement confirms the effectiveness of each preprocessing stage in enhancing image clarity and usability.

5.5 Class Distribution and Balance

Figure 4: Dataset Class Distribution



- Good: 1,700 images (34%)
- Medium: 1,800 images (36%)
- Bad: 1,500 images (30%)

After augmentation:

- Total: 15,000 images (balanced: ~5,000 per class)

Balanced datasets are essential for avoiding model bias and improving generalization performance in classification tasks [12].

5.6 Public Dataset Release

The processed dataset is made publicly available to support further research:

- **Platform:** Kaggle
- **Dataset Name:** *tamil-nadu-seed-quality-dataset-v1.0*
- **Format:** 15,000 preprocessed 224×224 JPEG images with labels (CSV)
- **Metadata:** Includes GPS coordinates, humidity levels, and capture date
- **License:** CC-BY 4.0 (for research use)

Key Finding

The proposed preprocessing pipeline achieves a 7% absolute improvement in CNN accuracy and increases usable images by 23%, effectively addressing challenges associated with Non-IID field data. Unlike existing datasets focused on maize and soybean, this work provides a region-specific, high-quality dataset tailored for tropical agricultural conditions, enabling more accurate and scalable solutions for precision agriculture.

VI. CONCLUSION & FUTURE WORK

This paper introduces the first publicly available Tamil Nadu seed quality dataset, comprising over 5,000 images of paddy and millet seeds categorized into good, medium, and bad classes. A comprehensive preprocessing pipeline based on OpenCV techniques was developed, achieving 95% clean images and significantly improving data quality for deep learning applications. The proposed approach enabled a baseline CNN model to reach 92% classification accuracy, demonstrating the effectiveness of preprocessing in handling real-world agricultural image data.

The study addresses a critical research gap by focusing on Non-IID, region-specific agricultural data collected under humid coastal conditions of Tamil Nadu. Unlike existing maize or soybean datasets, this work reflects real field variability, making it highly relevant for practical deployment. The dataset expansion from 5K to 15K samples through augmentation further strengthens model robustness and generalization.

Key Achievements:

- Development of a region-specific agricultural dataset tailored to tropical Indian conditions
- Implementation of an end-to-end preprocessing pipeline achieving PSNR of 34.2 dB and SSIM of 0.94
- Significant 7% improvement in CNN accuracy, validating preprocessing effectiveness
- Creation of a balanced augmented dataset (15K images) for robust model training

Future Work:

- Advanced Deep Learning Models: Implementation of architectures such as ResNet50 and EfficientNet for improved classification performance
- Explainable AI: Integration of Grad-CAM techniques to visualize and interpret seed defect regions
- Multi-modal AI Systems: Leveraging LLM-based frameworks such as LLaVA for combined image-text analysis
- Mobile Deployment: Development of a lightweight TensorFlow Lite application for real-time farmer usage

- Federated Learning: Building cross-regional collaborative models to generalize across different agro-climatic zones

The released Kaggle dataset (tamil-nadu-seed-quality-v1.0) provides a strong foundation for future research in precision agriculture and smart farming systems, particularly for developing regions. This work has immediate practical implications in automated seed certification, quality grading, and reducing 20–30% crop losses caused by poor germination, thereby supporting sustainable agricultural practices.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to our guide Mr. B. Rajalingam, Assistant Professor, Department of Information Technology, Mohamed Sathak Engineering College, for providing us with the opportunity to undertake this research work titled *"Tamil Nadu Seed Image Dataset with OpenCV Preprocessing for Precision Agriculture"* under his valuable guidance. His constant support, motivation, and insightful suggestions greatly contributed to the successful completion of this work.

We are highly thankful to Dr. B. Aysha Banu, Professor and Head, Department of Information Technology, Mohamed Sathak Engineering College, for her continuous encouragement, expert guidance, and constructive feedback throughout the research process.

We also extend our sincere thanks to the management and administration of Mohamed Sathak Engineering College for providing the necessary infrastructure, resources, and a supportive academic environment to carry out this research work successfully.

We would like to acknowledge all the faculty members and staff of the Department of Information Technology for their valuable support and cooperation during the course of this work.

Finally, we express our heartfelt appreciation to the student researchers Dinesh Kumar S, Arshath Ahamed H, and Mohamed Safic A for their dedicated efforts, teamwork, and active participation, which played a vital role in the successful completion of this research project.

REFERENCES

[1] M. Yuan *et al.*, "A dataset for fine-grained seed recognition," *Scientific Data*, vol. 11, no. 344, Apr. 2024.

[2] "High-resolution RGB image dataset for wheat seed varietal identification and purity assessment," *Data in Brief*, Elsevier, Apr. 2025.

[3] "Dataset of cannabis seeds for machine learning applications," *Data in Brief*, Elsevier, Jan. 2023.

[4] "Corn seed dataset based on hyperspectral and RGB images," *Data in Brief*, Elsevier, Dec. 2025.

[5] "Automated Seed Quality Assessment and Classification Using Watershed Algorithm and Ensemble Learning," in *Proc. IEEE Conference*, Apr. 2024.

[6] "Sesame Seed Disease Detection Using Image Classification," in *Proc. IEEE Conference*, Feb. 2021.

[7] "Contributing to agriculture by using soybean seed data from the tetrazolium test," *Data in Brief*, Elsevier, Jan. 2019.

[8] "EfficientMaize: A Lightweight Dataset for Maize Classification on Resource-Constrained Devices," *Data in Brief*, PMC, Mar. 2024.

[9] "Deep Convolutional Neural Network for Plant Seedlings Classification," *arXiv preprint arXiv:1811.08404*, Nov. 2018.

[10] "Estimating compositions and nutritional values of seed mixes based on vision transformers," *Smart Agricultural Technology*, Nov. 2023.

[11] "PhenoBench—A large dataset and benchmarks for semantic image interpretation in the agricultural domain," *arXiv preprint arXiv:2306.04557*, Jul. 2024.

[12] "Classification of seeds using domain randomization on self-supervised learning frameworks," *arXiv preprint arXiv:2103.15578*, Mar. 2021.

[13] "Image-based phenotyping of seed architectural traits and prediction of seed weight using machine learning models in soybean," *Frontiers in Plant Science*, Sep. 2023.

[14] "Rapid and accurate varieties classification of different crop seeds under sample-limited condition based on hyperspectral imaging and deep transfer learning," *Frontiers in Plant Science*, Jul. 2021.

[15] "The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach," *Agronomy*, vol. 11, no. 2, p. 263, Feb. 2021.