ISSN : 2455-9679
SJIF Impact Factor : 6.190

IJRTSM
INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"HEALTHY SOYBEAN SEEDS IMAGE DATASET FOR QUALITY ASSESSEMENT AND CLASSIFICATION "

Gayatri [1], Dr. Harish Patidar [2]
[1] *Department of Computer Science and Engineering, Mandsaur University, Mandsaur, MP, India*
[2] *Department Computer Science and Engineering, Mandsaur University, Mandsaur, MP, India*

ABSTRACT

Data preparation plays an important role in developing reliable machine learning and deep learning models for agricultural image analysis. In soybean seed research, the availability of well-prepared and consistent datasets is essential for building accurate classification and inspection systems. This study focuses on the development of a healthy soybean seed image dataset and a systematic data preparation process. A total of 1,197 RGB images of healthy soybean seeds were captured under controlled environmental conditions using a uniform background and consistent lighting setup. The controlled imaging environment helped reduce shadows and noise, resulting in clear and consistent images. Several preprocessing techniques were applied to improve the quality of the images, including color channel extraction, image enhancement, thresholding, edge detection, and contour-based cropping. These steps helped isolate the soybean seeds from the background and produce standardized images suitable for dataset creation. The final dataset was arranged through a 70:30 training and testing split to ensure the appropriate data distribution. In order to promote open research and reproducibility, the dataset was made publicly available on the Kaggle platform. Our dataset and processing pipeline represent a valuable resource for researchers, who can use them for soybean seed analysis, classification, and development of automated agricultural inspection systems.

Keywords: *Soybean seeds, Healthy seeds dataset, Real-time image acquisition, Deep learning, Convolutional Neural Networks (CNN), Seeds quality assessment, Image classification, Agricultural computer vision, Machine Learning, Soybean Classification, Dataset Labeling.*

I. INTRODUCTION

The integration of digital imaging and machine learning in agriculture has significantly developed in the past few years, especially in areas like seed classification, disease detection, and crop quality evaluation. Evaluating seed quality is highly important on the whole because the germination and growth of the plants in the field, as well as the final production, depend directly on the purity, physical condition, and standardization of seeds. [2] Manual inspection, on the one hand, is a time-consuming, subjective, and often inconsistent method, on the other hand, it raises the necessity of adopting automated and dependable techniques. Soybean is extensively grown and used in food, feed, and oil production. Due to its commercial importance, accurate assessment of seed quality is indispensable. Automatic seed classification systems depend on excellent image datasets for the training and testing phases. However, preparing datasets is not always easy as problems like uneven lighting shadows changing backgrounds, and low image resolution often arise. Thus, a carefully planned dataset is the key to creating effective and reliable AI-based seed classification models. The past research has mostly built their studies on the use of DSLR cameras and sophisticated imaging instruments. These types of equipment are capable of producing high-quality and detailed images but, at the same time, they make the access to such technology more difficult because of high cost and scarcity. Thanks to the recent

ISSN : 2455-9679
SJIF Impact Factor : 6.190

improvements in smartphone cameras, at present we can say that mobile phones are capable of taking images that are detailed enough. This is really a big plus as it brings the mobility, ease of use, and suitability for different situations including in the field conditions. For this project, we have devised a comprehensive protocol for dataset building by employing a smartphone equipped with a 50 MP camera on the back. The pictures were taken inside the light box that was used to create a consistent illumination and also to reduce the shadow completely. A blue EVA sheet was utilized as a background as it not only made the yellow soybean seeds stand out more vividly but also helped in the accurate segmentation of the seeds. Post-capture processing was performed using Python and OpenCV. Steps included RGB and CMYK channel separation, noise removal via mean-shift filtering, foreground-background separation with Otsu's thresholding, edge detection using Canny, contour identification, individual seed extraction, and resizing to 300×300 pixels. The resulting dataset is organized, segmented, and ready for use in machine learning or deep learning-based seed classification. This approach provides an economical, reproducible, and efficient method to generate high-quality soybean seed images using a smartphone, supporting automated classification systems and contributing to AI-driven digital agriculture research.

A. IMPORTANCE OF HEALTHY SOYBEAN SEEDS FOR RESEARCH

1. Essential for accurate identification

There is no publicly available, high-quality dataset focused specifically on healthy soybean seeds. Due to this gap, machine learning models struggle to accurately learn the visual characteristics of healthy seeds. The dataset created in this study addresses this problem by providing clean, well-curated, and high-resolution images of healthy soybean seeds

2. Makes seeds quality checking easier

In most places, seeds quality is checked manually, which takes time and may not always give consistent results. Test dataset will help computers identify healthy seeds quickly and more accurately.

3. Improves AI model performance

Most existing datasets focus on damaged or diseased seeds [9]. By providing images of healthy seeds, test dataset helps AI models understand the ideal characteristics of good-quality seeds, leading to better classification.

4. Useful for research and technology development

The dataset will support future research, model development, and automated systems for soybean seeds quality analysis using deep learning or machine learning models [13]-[15].

II. METHODOLOGY

The complete pipeline used to prepare a high-quality soybean seed healthy dataset. Each step has been systematically designed to ensure that the captured images are consistent, clear, and suitable for machine learning–based seed classification. The workflow begins with creating a controlled environment for image acquisition and ends with extracting individual seeds from processed images and saving them in multiple formats. [10] [8] [3]
Before capturing images, a controlled and stable imaging environment was created to eliminate variations that commonly affect image quality, such as poor lighting, shadows, uneven backgrounds, and reflections. This step ensures that every image has a uniform appearance, making image processing and classification easier and more accurate.

A. Environment Components

A controlled imaging environment was established to ensure high-quality image acquisition of soybean seeds.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

1) *Light Box Setup:* A light box was prepared to provide soft and diffused illumination. This removes shadows and gives uniform bright- ness across the entire image frame. This is essential because soybean seeds have subtle texture differences that must be captured clearly.

2) *Blue EVA Background Sheet:* A blue EVA foam sheet was used as the background for placing soybean seeds. This color was chosen specifically because: Soybean seeds mainly have light-yellow/brown tones Blue is their complementary color in CMYK color space This provides strong contrast between seeds and background As a result, even small seeds, damaged edges, or shriveled parts become more visible for segmentation.

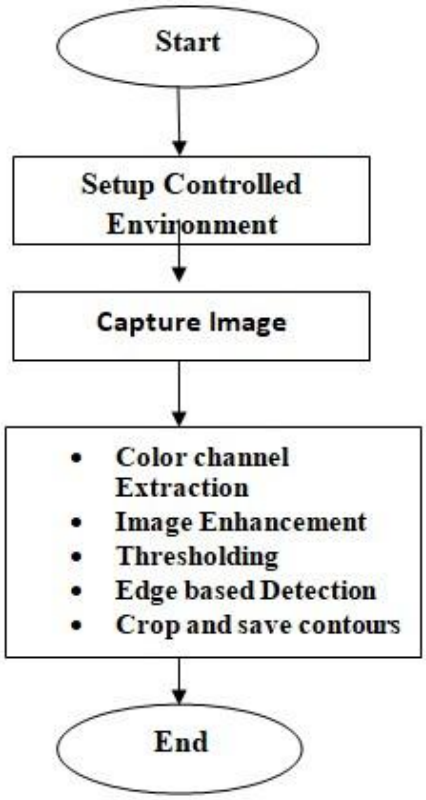

Fig. 1. Flow diagram of Dataset Prepration.

3) *Mobile Camera Configuration:* All images were captured using the rear dual camera setup (50MP + 8MP) of the smartphone. This camera provides:

- Ultra-high resolution
- Accurate color reproduction
- Better texture detail
- Enhanced clarity for feature extraction The front camera was not used because it offers lower resolution (16MP) and weaker color accuracy compared to the rear camera setup.

4) *Camera Positioning:* Distance: 25–30 cm from seeds Angle: 90 degrees (perpendicular) to avoid perspective distortion Focus: Locked manually on seed surface Flash: Off Natural LED white lighting used This controlled setup ensures that all images are captured under identical conditions.

B. Capture Image

Once the environment was prepared, the soybean seeds were placed on the background sheet. Images were captured with the Smartphone rear 50MP camera.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Key objectives during image capture:

- Ensure seeds were placed with a small amount of space between them
- Avoid overlapping or touching seeds (if touched, later separated using edge detection)
- Capture images in high resolution to maintain details Keep lighting consistent for all image batches
- Each image contains multiple seeds, forming the "raw dataset." These raw images serve as the input for the image processing pipeline.

C. Color Channel Extraction

In this step, color information from the image is separated into individual channels [4] [5]. Classical computer vision techniques use specific color channels for segmentation. RGB Extraction: [12] The image is split into Red, Green, and Blue channels. How- ever, RGB alone may not provide ideal separation between yellow seeds and blue background. CMYK Conversion: To improve segmentation accuracy, the RGB image is converted into CMYK color space. Y

(Yellow) channel represents the soybean seeds more strongly C (Cyan) channel represents the blue background This strong contrast helps in achieving cleaner segmentation.

D. Image Enhancement

After channel extraction, images are further enhanced to improve segmentation quality. Mean-Shift Filtering:

This technique smoothens small variations in color while preserving edges.
Benefits:

- Removes noise
- Enhances seed boundaries
- Makes thresholding more accurate

This filtering is essential for images captured using smartphone sensors since minor lighting variations or shadows may appear.

E. Thresholding

Thresholding is the process of separating the foreground (soybean seeds) from the background. Otsu Thresholding:

- Otsu's method automatically determines the optimum threshold value. It converts the enhanced image into a binarymask. White pixels represent seeds Black pixels represent background This step creates the primary mask that is used to remove the background.

F. Edge-Based Detection

Even after thresholding, some seeds may be touching or may have unclear boundaries. To solve this, Canny edge detection is applied.

Canny helps in:

- Enhancing boundaries of seeds
- Detaching seeds that are touching each other
- Detecting cracks or defects on seed surfaces
- Improving contour extraction accuracy This ensures that each seed is processed individually.

G. Crop and Save Contours

This is the final and most important stage of dataset preparation. [5] [4]

ISSN : 2455-9679
SJIF Impact Factor : 6.190

Contour Extraction:

•Using Open CV, contours of each seed are detected from the binary mask. Contours outline the exact shape and boundary of each seed.

H. Cropping Individual Seeds

1) *Once contours are identified:* A bounding box is drawn around each seeds Each seed is cropped with slight padding Cropped images are resized to 300 × 300 pixels.

2) *Multiple Dataset Formats:* For each seeds, three images are saved: Original seeds image with background Seed-only image with transparent/white background Contour-outline image These versions help machine learning models learn color, texture, and shape features separately.

III. RESULT AND ANALYSIS

A. Dataset Overview

The healthy soybean seed dataset was purposely created using controlled image acquisition to guarantee consistency and dependability. To be precise 1197 RGB images displaying healthy soybean seeds were hand, picked one by one under the same lighting and background settings. The use of a controlled environment greatly reduced the occurrence of shadows and reflections thus the seed outlines were very clearly seen. The images were grouped and ready for the next stages of processing and analysis. To ensure a well, balanced distribution of the dataset, a split of 70% training and 30% testing was done, which yielded 838 training images and 359 testing images. Such a division not only provides ample data for the model to learn from but also keeps a separate test set for the assessment. The prepared dataset has been made publicly available on the Kaggle platform to support open research and encourage further studies in soybean seed classification and agricultural image analysis.

Table 1. Summaryofthe proposed healthy soybean dataset.

Parameter	Value
Total Images	1197
Image Type	RGB
Dataset Class	Healthy Soybean Seeds
Training Set (70%)	838
Testing Set (30%)	359
Dataset Platform	Kaggle

The dataset consists of 1,197 RGB images captured under controlled environmental conditions and is divided into training and testing subsets using a 70:30 ratio.

Fig. 2. Sample healthy soybean seed image from the proposed dataset captured under controlled imaging conditions.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

B. Image Acquisition Analysis

The image acquisition process was conducted in a controlled environment to maintain consistency across all captured images. A uniform background was used to simplify object segmentation and reduce background noise. Additionally, stable illumination conditions helped minimize shadows and variations in brightness. As a result, the soybean seeds are clearly distinguishable from the background, which simplifies subsequent pre-processing and segmentation steps. The controlled acquisition process also ensures that the dataset maintains high visual quality and uniformity, making it suitable for machine learning and deep learning applications.

C. Image Pre-processing Results

After image acquisition, several preprocessing techniques were applied to enhance the quality of the captured soybean images and prepare them for dataset generation. The preprocessing pipeline includes color channel extraction, image enhancement, thresholding, edge detection, and contour extraction. Initially, color channel extraction was performed to analyze the color distribution of the soybean seeds. Image enhancement techniques were then applied to improve the contrast and visibility of seed features. These operations help highlight important visual characteristics of the soybean seeds. Thresholding was done to cut the soybean seed out of its background. That process basically turns the image into a black, and, white version, which simplifies the seed object isolation task. Then the edge detection was carried out to get the edges of the soybean seed; that step gives a neat outline of both its shape and the internal structure.

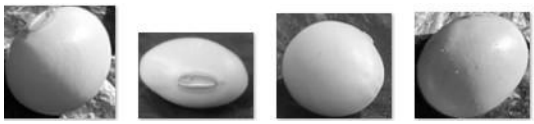

Fig. 3. The original RGB image is converted into grayscale to simplify the image processing task. This step reduces computational complexity while preserving the structural information of the soybean seed.

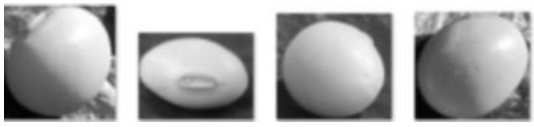

Fig. 4. Image enhancement techniques are applied to improve the contrast and visibility of the soybean seed. This step highlights important visual features of the seed and helps in improving segmentation accuracy.

Fig. 5. Thresholding converts the grayscale image into a binary image where the soybean seed is clearly separated from the background. This step helps in identifying the seed region more effectively.

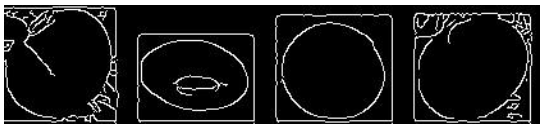

Fig. 6. Edge detection is applied to identify the boundaries of the soybean seed. The detected edges provide structural information about the shape of the seed and assist in further contour extraction.

I. Contour Extraction and Final Dataset Generation

After detecting the seed boundaries, contour detection was applied to extract the soybean seed region from the image.

ISSN : 2455-9679
SJIF Impact Factor : 6.190

This step enables the system to accurately isolate the seed from the surrounding background. The extracted seed contours were then cropped and saved as individual dataset images. The final processed images contain clearly segmented soybean seeds with minimal background noise. These standardized images form the final dataset used for further research and analysis.

J. Dataset Quality Analysis

The developed healthy soybean dataset demonstrates high visual consistency and image quality due to the controlled acquisition environment and systematic pre-processing steps. The pre-processing pipeline effectively removes noise, enhances image clarity, and highlights seed boundaries. The resulting dataset provides a reliable and structured image resource that can support the development of machine learning and deep learning models for soybean seed classification and agricultural automation. The public availability of the dataset further contributes to reproducible research and encourages future studies in agricultural image analysis.

IV. CONCLUSION

In this study, a dataset of images of healthy soybean seeds was created. Also a method of preparing data for the analysis of agricultural images was presented. A total of 1, 197 images of healthy soybean seeds were captured with an RGB camera in a controlled environment with a uniform background and lighting conditions. This kind of setting greatly helped to avoid shadows and other unwanted elements, results are very clear and uniform images of soybean seeds. Several preprocessing steps were applied to improve the quality of the captured images. These steps included color channel extraction, image enhancement, thresholding, edge detection, and contour-based cropping. The pre-processing process helped in clearly separating the soybean seed from the background and in obtaining well-defined seed boundaries. As a result, the final images became more suitable for further analysis and machine learning-based applications. To ensure proper data distribution, the prepared dataset was divided into 70% for training and 30% for testing. Besides that, the dataset has been published on the Kaggle platform for easy access and usage by other researchers. In summary, the dataset creation and data preparation procedures have resulted in a valuable research tool for the analysis of soybean seeds. This work lays a foundation for further research on the characterization of different soybean seed types, identification of seed defects, and the creation of automated systems for agricultural inspection.

REFERENCES

[1] F. J. Rodríguez-Pulido, D. F. Barbin, D. W. Sun, B. Gordillo, M. L. Gonzalez-Miret, and F. J. Heredia, “Grape seed characterization by NIR hyperspectral imaging,” Postharvest Biology and Technology, vol. 76, pp. 74–82, 2013. doi: 10.1016/j.postharvbio.2012.09.007.

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, “Machine learning in agriculture: A review,” Sensors. [Online]. Available: www.mdpi.com/journal/sensors

[3] G. J. Junior, A. Cardoso, L. Marques, I. Peretta, and P. Grider, “Proposed approach for creating soybean grain image dataset,” in Proc. Latinoware, 2024, pp. 222–228. doi: 10.5753/latinoware.2024.245770.

[4] W. Lin, Y. Lin, J. Chen, Z. Gao, and C. Huang, “Soybean image segmentation based on multiscale Retinex with color restoration,” Journal of Physics: Conference Series, vol. 2284, p. 012010, 2022.

[5] K. Kiratiratanapruk and W. Sinthupinyo, “Color and texture for corn seed classification by machine vision,” in Proc. Int. Symp. Intelligent Signal Processing and Communications Systems (ISPACS), 2011, pp. 7–11.

[6] Y. Li, J. Jia, L. Zhang, A. M. Khattak, S. Sun, W. Gao, and M. Wang, “Soybean seed counting based on pod image using two-column convolutional neural network,” IEEE Access, vol. 7, pp. 64177–64185, 2019.

[7] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

[8] W. Lin et al., “Soybean image dataset for classification,” Data in Brief, vol. 48, 2023.

[9] W. Lin et al., “Soybean seeds dataset,” Mendeley Data, vol. 6, 2023. doi: 10.17632/v6vzvfszj6.6.

[10] G. Dhakad, “Healthy soybean seed image dataset for ML,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/drgayatridhakad/soybean-seeds-healthy-dataset

ISSN : 2455-9679
SJIF Impact Factor : 6.190

[11] G. Zhao et al., "Real-time recognition system of soybean seed full-surface defects based on deep learning," Computers and Electronics in Agriculture, vol. 187, p. 106230, 2021. doi: 10.1016/j.compag.2021.106230.

[12] K. Kiratiratanapruk and W. Sinthupinyo, "Color and texture for corn seed classification by machine vision," in Proc. ISPACS, Chiang Mai, Thailand, 2011, pp. 1–5. doi: 10.1109/ISPACS.2011.6146100.

[13] A. Sable, P. Singh, A. Kaur, M. Driss, and W. Boulila, "Quantifying soybean defects: A computational approach using deep learning techniques," Agronomy, vol. 14, no. 6, p. 1098, 2024. doi: 10.3390/agronomy14061098.

[14] Y. Gulzar, "Enhancing soybean classification with modified inception model: A transfer learning approach," Emirates Journal of Food and Agriculture, vol. 36, pp. 1–9, 2024. doi: 10.3897/ejfa.2024.122928.

[15] G. F. D. Wendling et al., "Soybean seed classification using NIR and machine learning," Brazilian Archives of Biology and Technology, vol. 67, 2024.

[16] H. Hang and K. Ogasawara, "Grad-CAM-based explainable artificial intelligence related to medical text processing," Bioengineering, vol. 10, no. 9, p. 1070, 2023. doi: 10.3390/bioengineering10091070.