A Database for Soybean Seed Classification

Gabriel M. L. Pereira*, Juliano H. Foleis*, Alceu de Souza Brito Jr†. and Diego Bertolini*
*Universidade Tecnológica Federal do Paraná - UTFPR-CM - Campo Mourão, Paraná, Brazil
Email: {gabrielmlobato, julianofoleiss, diegobertolini}@utfpr.edu.br
†Pontifícia Universidade Católica do Paraná - PUC-PR - Curitiba, Paraná, Brazil
Email: alceu@ppgia.pucpr.br

***Abstract*—This paper proposes a novel public database for soybean seed defect classification. The database is publicly available and features seven defect classes labeled based on visual characteristics by domain specialists. Each seed sample includes three images taken from random rotations to ensure a comprehensive representation. Additionally, seeds from three different locations, each labeled by different experts, are included in the database. No public currently available soybean defect classification dataset provides location information. This allows for the evaluation of the generalization capabilities of the model on seeds collected on previously unseen regions. Experiments were conducted and compared with the GB1352 database to validate our approach. We employed handcrafted texture descriptors and non-handcrafted features extracted from established convolutional neural network architectures, using the network's last layer activations as the feature set. The results show that partitioning folds by the regions where seeds were collected significantly impact classification performance. Employing the SVM classifier and the product prediction fusion rule, we achieved an F1-score of 85.06% with region-based cross-validation folds and 95.89% without region-based fold partitioning. In experiments involving two classes (intact vs. defective), we achieved an F1-score of 99.32% with region-based cross-validation folds. We expect this database to foster the development of more robust models capable of generalizing to previously unseen regions.**

I. Introduction

Soybeans are an essential crop in the agricultural sector worldwide. According to data provided by Embrapa [1], Brazil is the fourth-largest grain producer in the world. Due to various factors, soybean production has increased globally for several decades. Several factors, such as global demand, a solid market, and ongoing technological advances, contribute to this growth. In recent decades, there has been significant investment in all stages of grain cultivation. However, assessing the vigor of the seeds used for planting is crucial, as it is directly related to the field's production potential [1].

Seed quality analysis is mostly done via chemical reagents or visual assessment. Visual classification is commonly employed by agricultural companies and cooperatives for seed quality control [2]. This classification, conducted by human experts, visually evaluates the quality of the seeds. One of the difficulties in this process is the subjectivity of the human specialist. Despite frequent courses and training, the human classifier can make errors due to long hours of repetitive work and be more or less rigorous at different times. Such variance in human specialist knowledge leads to difficulty in standardizing the classification process, compromising the guarantee of germination potential. Therefore, researchers have employed machine learning techniques to assist in soybean seed defect classification [3]–[7].

Several studies in this field use private datasets [8], [9]. This hinders the verification of methods and the comparison of results. There are some public datasets, such as GB1352 [10]. However, the documentation for these datasets is superficial, omitting significant details such as the labeling process, the expertise level of the labelers, the seed collection locations, and the cultivars. The absence of these details makes it difficult to assess the generalization capability of classification models. This work presents a public dataset called SOYPR, labeled by various specialists with extensive experience in the field. The dataset contains information such as the cultivar planted in three different regions of Paraná State, Brazil. This work also presents a study evaluating classification models trained and tested with seeds from different regions. Additionally, we present results from experiments that assess the influence of the number of images per seed on classification rates.

II. Proposed Database

The database was created through a partnership between our university and a local agricultural cooperative. Seeds delivered by cooperative members at different times of the year were selected and labeled by three human experts with extensive experience and skills in visually classifying seeds. However, each seed was labeled by a single specialist. The seeds were delivered in December 2022 and digitalized in January 2023. Before digitalization, the seeds were kept in a dry place without exposure to sunlight. The seeds came from three locations in Paraná State, Brazil. This paper will discuss further how this allows cultivar and region invariance to be assessed in classification models. In this work, we will refer to this dataset as SOYPR.

The experts labeled the seeds according to seven classes: (1) Intact seed, (2) Humidity damage, (3) Mechanical damage, (4) Greenish seeds, (5) Dirty seeds, (6) Cercospora Leaf Blight, and (7) Bug laceration. This work used the seven commonly adopted classes in the classification process. Other datasets employ five classes as described in [8], [9].

A distinguishing feature of this database is that the seeds came from different regions. More precisely, the seeds were sourced from three regions in Paraná State, Brazil, each with distinct soil, temperature, cultivars, and other characteristics. This contributes to a database resembling real-world scenarios,

where different locations produce seeds with varying characteristics. The seeds were also labeled by different experts, thus adding greater diversity to the classification stage. Seeds were received from farmers by Seed Processing Units (SPU). Each unit was responsible for collecting seeds from one of the three regions. Each SPU delivered to us approximately 100 seeds per class. They provided more seeds from the intact class (about 200). Thus, we received approximately 2400 labeled seeds. However, many of these seeds were discarded during digitalization as they were unsuitable. The SPUs receive seeds from nearby farms. However, we did not have access to the exact location where the seed was produced, only to the SPU where it was delivered. A single expert was responsible for a second labeling stage, in which he recommended the removal of specific samples due to subjectivity. We used a tray to organize the seeds in a grid pattern. The tray used has a capacity for 49 seeds. Figure 1 shows a tray filled with seeds.

Fig. 1: Tray containing 49 soybean seed samples.

The tray format simulates a real-world environment where it is possible to report which seed was predicted with which damage. In other words, from the seven rows and columns in the tray, we can show the row/column where each seed with a specific kind of damage is located. According to the cooperative, it is necessary to inform the farmer how many seeds exhibited defects and which ones exhibited them.

One of the issues encountered was that there was no guarantee that the damaged part of the seed would be visible. In some cases (mainly in the case of bug damage), the deterioration caused by the bug might be on the part of the seed that was not captured. Therefore, three images were captured per seed to increase the odds of capturing the damage. The seeds were initially placed in the tray and digitalized. Then, the seeds were randomly rotated and digitalized again. This process was repeated three times, resulting in three images per seed.

The digitalization process was conducted at the Plant Herbarium at our university in a controlled environment. The tray was placed inside a box with controlled white LED lighting for digitalization. A Canon DC 8.0v Digital Camera was used to capture the images. The images were saved in uncompressed TIFF format with dimensions of 3340×2588 pixels in RGB format. After digitalization, the images were segmented using image processing techniques. Segmentation and edge detection methods using the Sobel Operator were employed, along with seed color intensities for segmentation. We generated a dataset of 6,264 images of individual seeds labeled into seven classes at the end of this process. Subsequently, we improved the segmentation step using semantic segmentation with U-Net to reduce noise detected in some classes, such as Cercospora. Figure 2 shows the seeds after segmentation.

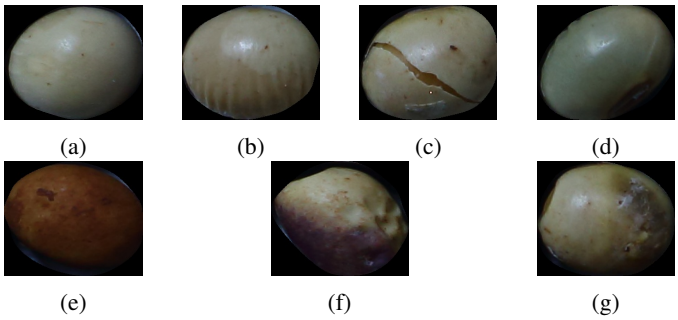

Fig. 2: Examples of segmented seeds: (a) Intact, (b) Humidity damage, (c) Mechanical Damage, (d) Greenish seed, (e) Dirty seed, (f) Cercospora Leaf Blight, and (g) Bug Damage.

TABLE I: Number of Seeds Collected from Each Region.

Class	Region-01 (R1)	Region-02 (R2)	Region-03 (R3)	Sum
Intact	130	129	129	388
Dirty	95	98	93	286
Bug	92	96	94	282
Greenish	88	76	65	229
Cercospora	56	79	59	194
Mechanical	36	39	42	117
Humidity	54	33	27	114
Totals	551	550	509	1610

Table I shows each class's number of examples per region. The difference between the number of seeds collected and those included in the database is due to uncertainties among human classifiers or seed damage during digitalization. The database is available for download on Github **here**.

III. Related Works

Classifying seeds automatically using machine learning is not a new approach [11]. Several studies have already demonstrated contributions in this regard [2], [8]. Some works, such as [9], [12], use images from the tetrazolium test for classification.

In [8], the authors present a pipeline for segmenting and classifying soybean seeds through images. The segmentation was done with Mask R-CNN, while the classification was performed by SNet, which is computationally cheap. With only 1.29 million parameters, SNet achieves a classification accuracy of 96.2%, outperforming previous models. The dataset

consists of 336 images of soybean seeds divided into five classes. The database is not publicly available. The paper does not specify where or when the seeds were collected.

In the approach described in [2], the authors use machine learning to classify soybean seeds and seedlings based on their appearance and physiological potential. The results demonstrate a strong correlation between the appearance of the seeds and their physiological performance. A total of 700 seeds were used, divided into seven classes based on appearance (100 samples per class). Using a feature vector with 103 dimensions extracted from the software Ilastik and a 70/30 split for training and testing, the authors achieved a best-case accuracy of 94%. The dataset is not publicly available.

In [7], the authors used deep learning to classify soybean seeds with the dataset proposed in [10]. They modified the InceptionV3 architecture by adding five extra layers and applied transfer learning and adaptive learning rate adjustment techniques to enhance accuracy. The evaluation metrics showed an accuracy of 98.73% using an 80:10:10 split for training, validation, and testing.

In [6], a computational approach was described to quantify defects in soybean seeds through seed classification using deep learning techniques. The classification network, SSDINet, consists of a convolutional neural network with deep convolutional blocks and squeeze-and-excitation blocks. Experimental results demonstrate that SSDINet achieved an accuracy of 98.64%. The dataset includes 750 defective samples divided into seven classes and 250 good seed samples. The dataset is available upon request.

In [4], the authors used the dataset proposed by [10] for soybean seed classification. The training, validation, and test splits were not specified. The ResNet-50 model was trained over various numbers of epochs, and its accuracy and loss were evaluated. The results indicate that the ResNet-50 model is efficient for multi-class soybean seed classification, achieving an accuracy of 86.84% on the validation set.

In [13], the authors propose a dataset for identifying soybean cultivars. This database differs from our proposed work, which aims to classify seeds based on seed quality. The authors employ a segmentation approach using the Canny method and other techniques. The generated dataset contains 649 samples, with approximately 50 samples from 13 classes. Using various classifiers and feature extraction approaches based on pre-trained ConvNext architectures, they achieved an average accuracy rate of 86.78%.

According to our collaboration with specialists from the agricultural cooperative, there may be significant visual differences between soybean seeds from various cultivars and geographical locations. This suggests that a model trained on samples from specific cultivars or regions might not generalize well to previously unseen cultivars/location combinations. Currently, publicly available datasets lack information on location or cultivar, making such evaluations impossible. To address this, we curated a dataset meticulously gathering samples from three distinct regions, each with different cultivars, to evaluate the generalization capabilities of machine learning models.

IV. Experimental Settings

This section presents the texture descriptors used in the experiments, the metrics optimized by the classifier models, and the GB1352 dataset.

A. Texture Descriptors

General-purpose texture descriptors have been successfully employed in several image classification problems, including seed classification [9], [14]. In this work, we used some of the most well-known texture descriptors as a baseline, such as BSIF [15], LBP [16], LPQ [17], OBIF [18], and GLCM [19]. We also employed a 128-bin histogram of the hue channel of the image. These features are often referred to as *handcrafted features* [20]. Table II (top rows) shows the parameters used for feature extraction, as well as the dimensionality of the feature vector.

A modern approach towards texture descriptors relies on features extracted by neural networks trained on large datasets dealing with many classes. Researchers have successfully used these descriptors in several pattern recognition problems, including seed classification [20]–[22]. They often surpass the results obtained with handcrafted texture descriptors [20], [21]. These descriptors are often known as *non-handcrafted features* [20], as they are learned by optimizing large neural networks. In this work, we employed the following commonly used neural networks for feature extraction: ResNet-50 [23], InceptionResNetV2 [24], EfficientNet [25], ViT-Small [26], and ViT-Base [26]. Table II (bottom rows) shows the parameters used for feature extraction.

TABLE II: Feature extraction parameters and dimensionality.

Feature	Parameters	Dimensionality
HSV	*Hue / 128 bins*	128
GLCM	*Cont, Homog, Corr, and Energy*	64
LBP	*P = 8 and R = 2*	59
LPQ	*winSize = 7*	256
OBIF	$\alpha = 2, 4$; $\varepsilon = 0.001$	138
BSIF	*filter = ICAtextureFilters, 11×11, 8bit*	256
ResNet-50	*Backbone* - Last feature layer	2048
InceptionResNet	*Backbone* - Last feature layer	1536
EfficientNet	*Backbone* - Last feature layer	1280
Vit-Small	*Backbone* - Last feature layer	384
Vit-Base	*Backbone* - Last feature layer	768

B. Metrics

Precision measures the quality of the prediction of the positive class and is calculated by:

$$Precision = \frac{TP}{TP + FP}$$

TP is the number of true positives, and FP is the number of false positives. *Recall* measures the proportion of positive samples that were identified out of all positive samples in the test dataset:

$$Recall = \frac{TP}{TP + FN}$$

FN is the number of false negatives. In this work, the models optimize the *F1-score* metric, defined as the harmonic mean between the *Precision* and *Recall* metrics:

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Optimizing the *F1-score* favors solutions where *Precision* and *Recall* are balanced and as high as possible. Consequently, a high *F1-score* is only achieved when both *Precision* and *Recall* are high.

C. GB1352 Dataset

We will conduct experiments with the database proposed by [10] to compare its complexity with the proposed database. In our experiments, we will refer to this dataset as GB1352. This database can be used for segmentation and classification tasks. The results described in [10] only cover the segmentation task. This database follows the GB1352-2009 standard and employs five classes for the classification stage. The classes are Intact, Immature, Skin-damaged, Spotted, and Broken soybeans. The database contains 5,513 images, ranging from 1002 to 1201 samples per class. Each image has a single seed and is 227×227 pixels in size. The article has no details about the different regions or cultivars from which the seeds were collected. There are also no details about the expertise of the human classifiers who generated the database. Another detail is that there is only one image per seed in this dataset, unlike the proposed database, which has three different images per seed. Other datasets mentioned in the literature are either non-public or do not have publicly available images, only extracted features [8], [9].

V. Experimental results and discussion

In all experiments in this study, 11 descriptors were evaluated: six handcrafted and five non-handcrafted, as detailed in Section IV-A. An SVM with an RBF kernel was selected as the classifier due to its robustness in handling high-dimensional datasets. Optimal SVM hyperparameters for C and γ were determined through an exhaustive search. In specific experiments with the SOYPR database, we used samples from three fields to evaluate the impact of the geographical regions where seeds were collected, as explained in Section II. Here, two fields were used for training and the third for testing. The reported results reflect the average Weighted F1-score across 3-fold cross-validation and the standard deviation. Additionally, we utilized the dataset described in Section IV-C, where each seed has a single image, in contrast to our proposed dataset, where each seed has three samples. All experiments were conducted using the *Scikit-learn*.

In the first experiment, we evaluated the performance of the SOYPR and GB1352 datasets using different descriptors. The results are presented in Table III. Preliminary experiments demonstrated the superiority of SVM over MLP in this task, leading us to use the SVM classifier in all subsequent experiments. A 3-fold cross-validation protocol was used. In this experiment, we did not consider the region where the seeds were collected for the SOYPR dataset to partition the three folds. Thus, seeds from the same region can be in both the training and test sets. This simulates the situation in the GB1352 dataset. Our primary objective was to determine if the results from the SOYPR dataset were comparable to those from the GB1352 dataset. In the SOYPR dataset, we also evaluated the impact of having different numbers of images per seed in training (one and three). We evaluated configurations with a single image per seed for training and testing (SOYPR.1-1) and three images per seed for training against one image per seed for testing (SOYPR.3-1). The GB1352 dataset, containing only one image per seed, did not allow a corresponding GB1352.3-1 evaluation.

TABLE III: Average F1-scores for all descriptors, varying the number of images per seed for training. Random partitioning.

Descriptors	SOYPR.1-1	SOYPR.3-1	GB1352
HSV	73.00 ± 0.9	67.56 ± 1.8	67.32 ± 0.7
GLCM	66.52 ± 1.7	54.19 ± 0.6	60.80 ± 1.3
LBP	66.34 ± 0.7	60.38 ± 2.5	57.36 ± 0.2
LPQ	68.57 ± 0.9	67.94 ± 1.0	60.13 ± 1.1
OBIF	70.87 ± 1.7	67.50 ± 1.3	67.81 ± 0.2
BSIF	70.38 ± 0.6	68.40 ± 1.6	61.04 ± 1.1
ResNet50	81.16 ± 1.2	83.72 ± 1.0	89.96 ± 1.4
IncepResnet	79.27 ± 0.2	80.56 ± 1.5	88.01 ± 1.2
EfficientNet	82.05 ± 1.3	85.87 ± 0.3	90.16 ± 1.3
Vit-Small	83.82 ± 1.6	86.44 ± 0.8	90.55 ± 0.6
Vit-Base	**83.97** ± 1.5	**88.60** ± 0.3	**91.06** ± 1.0

Table III shows that the results for both datasets exhibited similar behavior when random partitioning is used. Non-handcrafted descriptors outperformed the handcrafted ones for both datasets. OBIF, BSIF, and the HSV Hue histogram achieved the best results among handcrafted descriptors. Vit-Base and Vit-Small performed best among non-handcrafted descriptors. An increase in the number of samples per seed in the training set negatively impacted models trained with handcrafted descriptors. Handcrafted features are more sensitive to intra-class variation that arises when the number of samples per seed increases, as there is no guarantee that the damage will be apparent. In contrast, models trained with non-handcrafted features were not.

A second experiment evaluated the impact of partitioning the SOYPR dataset regarding the regions where the seeds were collected. We performed cross-validation across the regions R1, R2, and R3, ensuring that the test set did not contain samples from the regions in the training set. We evaluated using one and three images per seed in the training set as in the previous experiment. Table IV presents the results.

The results in Table IV show a significant decrease in performance compared to random partitioning (Table III). There was a performance drop of 10 to 15 percentage points for both handcrafted and non-handcrafted descriptors. The performance drop was expected, considering that, according to our collaborating specialists, there are noticeable visual differences among seeds from different cultivars and geographical locations. These differences make the underlying classification problem more difficult because certain variations in the test set are not present in the training set. Meanwhile, it more accurately represents a real-world situation where we

TABLE IV: Cross-validation with region-aware partitioning.

Descriptors	SOYPR.1-1	SOYPR.3-1
HSV	46.82 ± 3.4	47.02 ± 2.0
GLCM	48.12 ± 1.5	54.19 ± 0.6
LBP	41.87 ± 1.8	47.09 ± 2.4
LPQ	51.16 ± 1.6	55.93 ± 1.6
OBIF	50.10 ± 1.1	54.39 ± 0.5
BSIF	52.95 ± 3.5	56.76 ± 4.5
ResNet50	67.60 ± 1.8	70.13 ± 2.5
IncepResnet	68.02 ± 0.5	68.02 ± 0.5
EfficientNet	64.25 ± 4.0	68.12 ± 4.4
Vit-Small	73.03 ± 0.6	73.03 ± 0.6
Vit-Base	**73.83** ± 0.8	**74.74** ± 0.8

expect the model to generalize to previously unseen cultivars or regions.

Our proposed dataset contains three images per seed, as described in Section II. In a third experiment, we evaluated the effect of combining the predictions for each image to reach a final decision. We employed static combination methods proposed by Kittler [27], specifically voting, sum, max, and product. The sum and product rules produced better results than voting and max in all experiments. Therefore, we present the results obtained using these two approaches. Table V shows the average weighted F1-score from our experiments. In this context, the SOYPR-A results refer to the SOYPR dataset with three images per seed with 3-fold cross-validation partitioning without considering the location where the seeds were collected, as in the first experiment. As in the second experiment, the SOYPR-B results refer to the SOYPR dataset with three images per seed with the folds based on the regions where the seeds were collected.

TABLE V: F1-scores using fusion rules to combine the three samples per seed in SOYPR database for both random and region-aware partitioning.

Descriptors	SOYPR-A		SOYPR-B	
	Sum	*Prod*	*Sum*	*Prod*
HSV	74.79 ± 1.0	76.06 ± 0.7	60.16 ± 1.3	62.95 ± 1.1
GLCM	62.46 ± 1.1	65.50 ± 0.1	47.77 ± 1.8	50.41 ± 2.1
LBP	66.94 ± 0.4	68.74 ± 0.1	59.31 ± 3.1	61.43 ± 1.4
LPQ	76.82 ± 0.8	78.14 ± 0.7	66.51 ± 2.7	67.06 ± 1.0
OBIF	77.02 ± 0.4	79.03 ± 0.8	67.91 ± 3.5	69.71 ± 2.5
BSIF	78.36 ± 1.0	80.28 ± 0.5	68.45 ± 5.9	70.27 ± 4.0
ResNet50	93.52 ± 0.2	94.00 ± 0.1	80.73 ± 4.1	81.87 ± 4.6
IncepResnet	90.29 ± 0.2	90.91 ± 0.3	79.20 ± 1.6	80.14 ± 1.5
EfficientNet	93.52 ± 0.2	94.12 ± 0.6	77.47 ± 8.2	78.50 ± 7.9
Vit-Small	95.20 ± 0.7	95.63 ± 0.4	82.70 ± 3.6	83.00 ± 4.8
Vit-Base	95.77 ± 0.5	**95.89** ± 0.3	84.78 ± 4.5	**85.06** ± 5.0

Table V shows that combining the predictions from images from the same seed can improve classification accuracy. We observed an increase of about seven percentage points in SOYPR-A and 11 percentage points in SOYPR-B. This improvement is likely because the combination methods enhance the ability to detect regions with visible damage on the seeds. For instance, among the three samples, one may show more apparent damage than the others. In this context, the product rule demonstrated superior performance to other methods, likely due to variations in the presence and absence of damage in the samples. The sum rule also proved stable, yielding results close to those of the product rule.

Figure 3 shows the confusion matrix per class for SOYPR-B, using the Vit-Base backbone and Product Rule. Humidity damage exhibits the lowest classification accuracy, reflecting real-world challenges, as these characteristics are often not easily visible. Experts have noted these difficulties, and while techniques such as immersing seeds in water can make damage more apparent, these methods are time-consuming and not used in this protocol to avoid affecting the performance of other classes. The Intact and Cercospora classes achieve the highest performance. Farmers often reject dirty seeds, although not a defect, which can lower the selling price if their percentage is too high. While the classification rates for this class should technically be higher, we observe significant variation due to the subjectivity of human classifiers' perception.

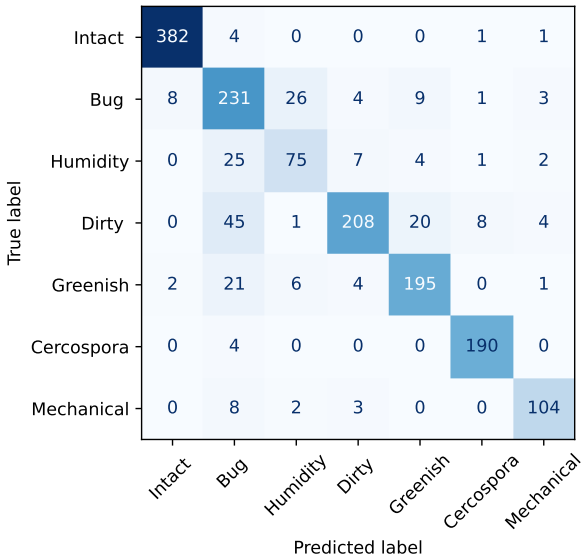

Fig. 3: Confusion Matrix of the best model using Vit-Base and Prod Rule - SOYPR-B.

According to our collaborating experts in the field, evaluating the type of defect is not always necessary. Instead, assessing the percentage of defective seeds can be a significant contribution. This acceptable percentage can vary depending on production and quality standards. Therefore, we conducted a fourth experiment to distinguish between intact and damaged seeds. For the SOYPR dataset, six classes were grouped as the defective class. For the GB1352 dataset, four classes were merged as the defective class, maintaining the same number of intact samples as in previous experiments. Table VI presents the results for the SOYPR and GB1352 datasets. We evaluated performance both without and with a fusion rule using the product rule.

Table VI shows that achieving accuracy rates above 95% in both datasets is possible. Using combination techniques with three samples per seed and applying product rules, we achieved an accuracy rate of 99.3% for SOYPR-B. In all cases,

TABLE VI: F1-scores for Intact vs. Damaged classification.

Descriptors	**SOYPR-B** *Without Fusion*	**SOYPR-B** *Prod*	**GB1352**
HSV	93.53 ± 0.4	97.10 ± 0.2	85.86 ± 2.0
GLCM	88.45 ± 1.7	85.99 ± 5.3	83.24 ± 1.4
LBP	90.32 ± 0.3	93.68 ± 1.1	81.52 ± 2.1
LPQ	93.44 ± 0.2	96.94 ± 1.5	86.69 ± 1.8
OBIF	94.21 ± 0.3	97.03 ± 2.0	89.30 ± 3.1
BSIF	90.39 ± 0.7	92.44 ± 2.6	88.19 ± 3.2
Resnet50	96.03 ± 0.6	97.73 ± 1.7	94.68 ± 1.4
IncepResnet	95.73 ± 0.2	98.83 ± 0.5	94.39 ± 1.0
EfficientNet	97.03 ± 0.0	98.94 ± 0.0	95.22 ± 0.7
Vit-Small	96.74 ± 0.2	97.74 ± 1.3	94.39 ± 1.0
Vit-Base	**97.00** ± 0.8	**99.32** ± 0.2	**95.36** ± 0.6

the Vit-Base feature achieved the best performance.

VI. Concluding Remarks

In this study, our primary goal was to evaluate our newly proposed soybean seed defect classification database in a real-world scenario. The database includes seeds collected from three geographical locations, as regional differences can affect seed visual aspects. Our new dataset is publicly available. We investigated the robustness of handcrafted and non-handcrafted feature descriptors in this context.

Our experiments led to the following observations: 1) Non-handcrafted extraction methods outperformed traditional handcrafted methods. 2) Dividing folds by the regions where seeds were collected significantly impacts classification accuracy. 3) Combining predictions across the three samples of each seed improves classification performance. 4) F1-scores above 99% are achievable even in realistic scenarios when classifying between intact vs. damaged seeds. The results from both evaluated datasets support these conclusions.

Based on our experiments, achieving performance comparable to that reported in the literature is feasible using simple prediction fusion rules with three samples per seed. In future studies, we plan to assess an updated version of this database, applying a multi-label approach to seed evaluation.

Acknowledgment

This study was financed in part by the Universidade Tecnológica Federal do Paraná - UTFPR and COAMO Agroindustrial Cooperativa.

References

[1] A. Aragao and E. Contini, "O agro no brasil e no mundo: uma síntese do período de 2000 a 2020," Embrapa SIRE, Tech. Rep., 2021.

[2] A. D. de Medeiros, N. P. Capobiango, J. M. da Silva, L. J. da Silva, C. B. da Silva, and D. C. F. dos Santos Dias, "Interactive machine learning for soybean seed and seedling quality classification," *Scientific reports*, vol. 10, no. 1, p. 11267, 2020.

[3] E. R. de Oliveira, P. H. Bugatti, and P. T. M. Saito, "Assessment of clustering techniques to support the analyses of soybean seed vigor," *PLOS ONE*, vol. 18, no. 8, pp. 1–20, 08 2023.

[4] A. Saini, K. Guleria, and S. Sharma, "A deep learning-based fine-tuned ResNet50 model for soybean seeds multiclass classification," in *2023 Global Conference on Information Technologies and Communications (GCITC)*. IEEE, 2023, pp. 1–6.

[5] N. Zhang, E. Zhang, and F. Li, "A soybean classification method based on data balance and deep learning," *Applied Sciences*, vol. 13, no. 11, 2023.

[6] A. Sable, P. Singh, A. Kaur, M. Driss, and W. Boulila, "Quantifying soybean defects: A computational approach to seed classification using deep learning techniques," *Agronomy*, vol. 14, no. 6, 2024.

[7] 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.

[8] Z. Huang, R. Wang, Y. Cao, S. Zheng, Y. Teng, F. Wang, L. Wang, and J. Du, "Deep learning based soybean seed classification," *Computers and Electronics in Agriculture*, vol. 202, p. 107393, 2022.

[9] D. F. Pereira, P. H. Bugatti, F. M. Lopes, A. L. Souza, and P. T. Saito, "Contributing to agriculture by using soybean seed data from the tetrazolium test," *Data in Brief*, vol. 23, p. 103652, 2019.

[10] W. Lin, Y. Fu, P. Xu, S. Liu, D. Ma, Z. Jiang, S. Zang, H. Yao, and Q. Su, "Soybean image dataset for classification," *Data in Brief*, vol. 48, p. 109300, 2023.

[11] P. Shatadal, D. Jayas, J. Hehn, and N. Bulley, "Seed classification using machine vision," *Canadian Agricultural Engineering*, vol. 37, no. 3, pp. 163–168, 1995.

[12] D. M. Rocha, L. H. P. Nóbrega, D. Bernardi, G. Conti, E. A. Nakajima, M. F. Ziech, and C. L. Bazzi, "Random forests in the supervised classification of multidimensional images of the tetrazolium test," *Journal of Agricultural Science*, vol. 11, no. 15, p. 115, 2019.

[13] E. S. Flores, M. R. Thielo, and F. R. Rodrigues Padilha, "Towards automatic soybean cultivar identification: Soycult dataset and deep learning baselines," in *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2023, pp. 151–156.

[14] A. EL IDRSSI, Y. Ruichek *et al.*, "Palmprint recognition using state-of-the-art local texture descriptors: A comparative study," *IET Biometrics*, 2020.

[15] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proceedings of the 21st Int. Conf. on Pattern Recognition (ICPR2012)*, 2012, pp. 1363–1366.

[16] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[17] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 236–243.

[18] M. Crosier and L. D. Griffin, "Using basic image features for texture classification," *International journal of computer vision*, vol. 88, no. 3, pp. 447–460, 2010.

[19] M. Hall-Beyer, "Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales," *IJSR*, vol. 38, pp. 1312–1338, 03 2017.

[20] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition*, vol. 71, pp. 158–172, Nov. 2017.

[21] P. Napoletano, *Hand-Crafted vs Learned Descriptors for Color Texture Classification*. Cham: Springer, 2017, vol. 10213.

[22] M. Souza Jr, W. C. Horikoshi, P. T. Saito, and P. H. Bugatti, "Soybean seed vigor classification through an effective image learning-based approach," *Multimedia Tools and Applications*, pp. 1–24, 2023.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 4278–4284.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[27] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, p. 226–239, Mar. 1998.