Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2026, 2026:61
https://doi.org/10.28919/cmbn/9846
ISSN: 2052-2541

DRY BEAN SEED CLASSIFICATION USING DATA PREPROCESSING TECHNIQUES AND DISTANCE-BASED CLASSIFIERS

S. DAMIAN-MUNIZ[1], R. A. LIZARRAGA-MORALES[2,*], G. HERNANDEZ-GOMEZ[1], M. E. RAMIREZ-SILVA[1]

[1]DICIS, Department of Multidisciplinary Studies, University of Guanajuato, Av. Universidad s/n, suburb Yacatitas, Yuriria, Guanajuato, México

[2]DICIS, Department of Art and Enterprise, University of Guanajuato, Carr. Salamanca–Valle de Santiago 3.5 + 1.8 km, Comunidad de Palo Blanco, Salamanca 36885, Guanajuato, México

Abstract. The development of automatic systems for agricultural applications, particularly for seed classification, is a task of paramount importance to the agro-industrial sector. Accurate recognition of seed varieties supports varietal purity and scalable processing. As one of the world's most significant crops, dry beans underscore the essential need for precise seed classification to safeguard crop quality and enhance agricultural productivity. Traditional methods for dry bean seed classification involve manual and semiautomated approaches, that are inadequate for scale processing, suffering from slow throughput and reliance on human interpretation. In this work, we proposed a pipeline that is centered on Weighted K-Nearest Neighbors (WKNN) for classifying seven dry-bean varieties. Firstly, we remove outliers per class and features via the IQR rule, rebalance classes, and standardize features with Z-score normalization. The results of this methodology show an improvement in the accuracy, precision, recall, and F1-score metrics for the classification of 7 different dry bean seeds in comparison with other classical machine learning and state-of-the-art approaches.

Keywords: seed variety; dry bean seed classification; feature analysis; distance-based classifiers; outlier filtering

2020 AMS Subject Classification: 92B10, 68T10.

*Corresponding author

E-mail address: ra.lizarragamorales@ugto.mx

Received February 27, 2026

1. INTRODUCTION

The classification of seeds is a critical challenge in agriculture. Errors in seed selection can reduce crop purity and compromise productivity [1]. In particular, the quality and yield of crops are strongly influenced by the precise classification of the seeds, making accurate identification of seeds varieties essential for maximizing agricultural output [2]. Traditionally, this task has relied on manual visual inspection, a process frequently reported as error-prone, highly subjective, and inefficient for large-scale processing contexts [3, 4]. Among the world's most important crops requiring precise seed classification are common beans, which exemplify the critical need for accurate identification to ensure crop quality and productivity.

Beans (*Phaseolus vulgaris* L.) are an essential source of proteins, vitamins, and minerals for millions of people worldwide, particularly in developing regions [5, 6, 7]. Moreover, due to their nutritional importance and nitrogen-fixing capability, beans play a key role in sustainable agricultural systems. In addition, they contribute significantly to crop rotations that improve soil fertility [8]. However, the classification of beans is not limited to distinguishing viable from non-viable seeds; it also involves differentiating between species and varieties. Precise identification is crucial to maintain varietal purity, support breeding programs, and ensure consistent crop performance.

Traditional seed classification has relied on simple morphological traits such as shape and surface texture, often assessed through manual or visual inspection. For example, Nelson and Wang [9] proposed a manual system for classifying soybean seeds, using a caliper to measure basic ratios of height, length, and thickness. Their method helped describe differences between the varieties for manual inspection. Subsequently, with the introduction of technological tools, more detailed analyses became possible. In wildflowers of the genus *Collomia*, Hsiao and Chuang [10] used microscopes to study the patterns on seed surfaces. They classified the seeds into two main types based on overall texture and surface features. However, they found that surface patterns alone were insufficient for reliable species identification and concluded that combining them with other traits, such as flower morphology, was necessary. Moreover, beyond classification for species discrimination, seed classification can also be applied to assess seed quality. For instance, Hill et al.[11] used a density separation method to classify hydrated

seeds of lettuce, tomato, and onion, demonstrating that removing low-density samples improved germination rates. Despite these advancements, manual and semi-automated approaches are inadequate for industrial-scale processing, suffering from slow throughput and reliance on human interpretation.

Automated seed classification has become an active area of research, with recent studies focusing on methods that apply intelligence, such as machine learning techniques and image processing, to improve accuracy and efficiency. These approaches have achieved substantial advances in performance, demonstrating significant improvements over traditional techniques [12, 13]. Early work highlighted the potential of morphological features in a neural network-based classification of cereal grains. Majumdar and Jayas [14] found that incorporating measurements such as length, width, area, perimeter, aspect ratio, roundness, and compactness allowed these models to accurately distinguish between wheat, barley, oats, and rye. Later, Paliwal et al. [15] expanded the feature sets for these cereals by incorporating a larger number of morphological, color, and texture descriptors, and explored different feature subsets to evaluate their impact on classification performance and training time. Further evidence supports the effectiveness of morphological features combined with machine learning models for grain classification. Hussain and Ajaz [16] further confirmed the value of morphological features by using them to classify three wheat varieties Kama, Rosa, and Canadian, finding that the multilayer perceptron (MLP) achieved the best performance. These findings underscore that detailed morphological characteristics extracted through image processing pipelines are prominent for developing robust, objective, and efficient automated systems suitable for industrial-scale seed processing. Regarding dry bean seed classification, recent studies have explored combinations of morphological features with advanced classifiers to enhance performance. Among the most relevant contributions in bean classification, Koklu and Ozkan [17] proposed a system combining RGB images and morphological features. They tested different types of classifiers and found that the support vector machine (SVM) achieved the best performance metrics. Subsequently, Macuácua et al. [18] incorporated preprocessing techniques such as principal component analysis (PCA), hyperparameter optimization, SMOTE-based class balancing, and outlier

removal. In their study, k-nearest neighbors (KNN) achieved the highest accuracy. Their approach surpassed the results obtained by Koklu and Ozkan. This highlights the importance of advanced data preparation steps for dry bean seed classification. Advanced oversampling and ensemble methods have also been applied. Khan et al. [19] introduced the use of ADASYN oversampling combined with extreme gradient boosting (XGB), demonstrating that sophisticated boosting algorithms can effectively handle class imbalance in morphological seed data. Nayak et al. [20] focused on SMOTE-boosted XGB, which provided further improvements in accuracy, underscoring the benefits of ensemble models integrated with data balancing methods. Koeshardianto et al. [21] reported competitive results using random oversampling with Random Forest and randomized search hyperparameter tuning. Recently, Lee et al. [22] proposed an innovative approach combining BLSMOTE, K-means clustering, and SVM. Their methodology demonstrates that integrating clustering methods with advanced oversampling strategies can significantly boost classification performance in highly imbalanced datasets. However, questions remain regarding the individual contribution of each morphological feature to classification performance. There are also uncertainties about the potential effects of variable reduction. In some scenarios, dimensionality reduction can simplify models or enhance interpretability. Nevertheless, it may also compromise the discriminative ability of classifiers by discarding relevant information [23]. In this paper, we propose an alternative approach for the automatic classification of seven varieties of dry bean seeds, integrating rigorous preprocessing techniques with a set of classical machine learning models. The analysis highlighted the simplicity and effectiveness of distance-based methods, with Weighted K-Nearest Neighbors (WKNN) consistently achieving the best results despite its lower computational complexity. The experimental design not only compared this model under standardized validation schemes but also benchmarked them against recent state-of-the-art contributions, showing that the proposed pipeline surpassed previously reported metrics. The remainder of this manuscript is organized as follows. Section 2 presents the materials and methods, including the dataset and the preprocessing pipeline. Section 3 reports the results and discussion, covering the experimental design, the evaluation of different machine learning classifiers, and the benchmarking against state-of-the-art approaches. Finally, Section 4 summarizes the conclusions of the work.

2. MATERIALS AND METHODS

This section describes the proposed methodology for classifying Dry Bean Seeds in 7 different classes. An overall description of the development of this methodology is presented in Fig. 1. This figure outlines the complete processing pipeline, which begins with the raw dataset and incorporates several key preprocessing stages. Initially, outliers were identified and removed using the Interquartile Range (IQR) method applied independently to each feature within each class. This filtering step ensures that extreme values do not distort the feature space or bias the training process. To address the evident class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This method generated synthetic instances for underrepresented classes, aligning all class sizes with that of the majority class. Following balancing, Z-score normalization was applied to all features to ensure consistent scaling across dimensions, which is critical for distance-based algorithms. For the classification stage, WKNN was selected due to its balance between simplicity and performance. This distance-based method extends the classical KNN by assigning higher influence to closer neighbors, which enhances classification accuracy while maintaining low computational cost.

To ensure a rigorous evaluation, two independent experiments were conducted. The first experiment focused on comparing the proposed methodology with other machine learning models under a unified preprocessing configuration, using both 10-fold cross-validation and an 80/20 hold-out strategy. In the second experiment, the proposed approach was compared with state-of-the-art published results.

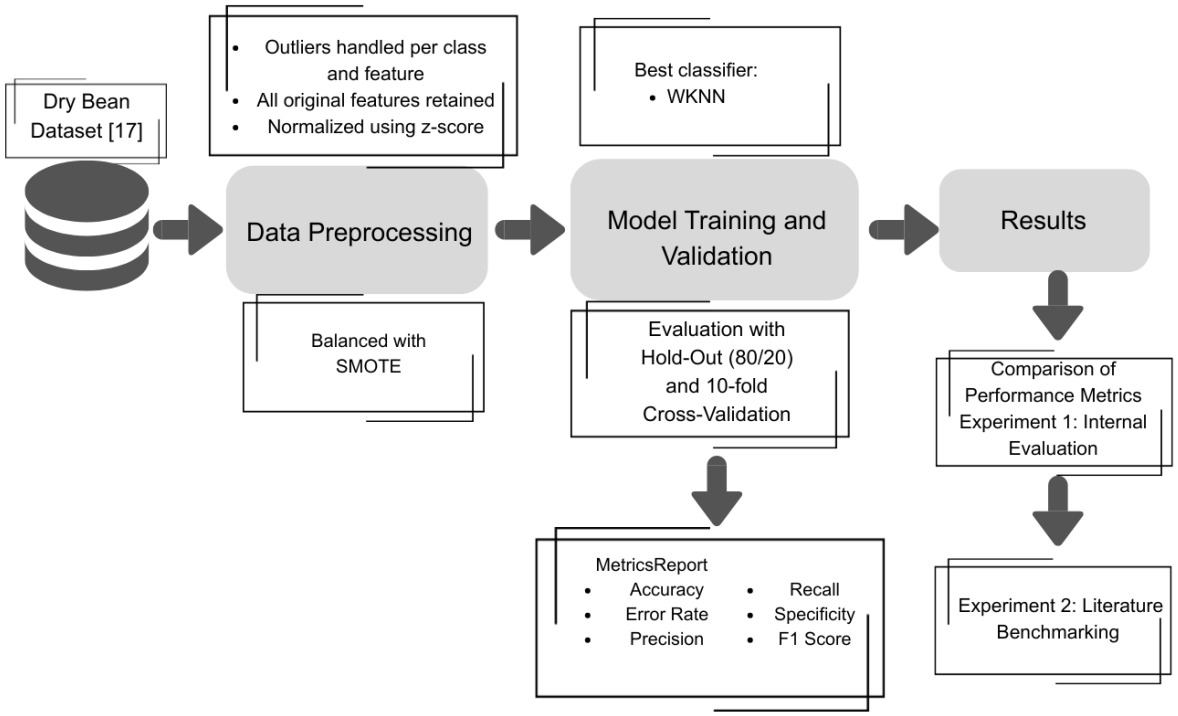


FIGURE 1. Overview of the classification pipeline, including preprocessing, model training, validation, and performance comparison.

2.1. Data Set description and Preprocessing

The following sections describe the dataset and preprocessing steps in greater detail, as they form the basis for all subsequent modeling and evaluation procedures. In this study, the *Dry Bean Dataset*, introduced in [17] was used. This dataset comprises 13,611 instances corresponding to 7 distinct dry bean seed varieties: Seker, Barbunya, Bombay, Cali, Horoz, Sira, and Dermason. Each instance is described by 16 morphological attributes extracted from digital images through computer vision techniques. These features include measurements such as area, perimeter, eccentricity, compactness, Major Axis Length, Minor Axis Length, Aspect Ratio, among others. All 16 morphological variables were preserved throughout the process to maintain the full descriptive power of the data. No dimensionality reduction or variable selection techniques were applied, in order to avoid discarding potentially informative features. The following subsections analyze the contribution of each preprocessing stage in the classification workflow. The analysis is focused on how specific operations namely, outlier filtering, class balancing, and feature normalization, influenced the performance of classifiers.

2.1.1. *Outlier detection and removal*

Outliers in structured datasets can significantly degrade the performance of machine learning classifiers by distorting the feature space, inflating variance, and skewing class distributions. Their detection and removal are thus critical preprocessing steps, particularly in domains such as agriculture and biosystems engineering, where data often exhibit heterogeneity and imbalance [24, 25].

To address this issue, the Interquartile Range (IQR) method was selected due to its simplicity, robustness, and wide acceptance in identifying extreme values without assuming any specific data distribution. Mathematically, the IQR for a given feature X is defined in Eq. (1) as the difference between the third and first quartiles [26].

$$\text{IQR} = Q_3 - Q_1, \tag{1}$$

An instance $x_i \in X$ is considered an outlier if it lies outside the bounds established in Eq. (2).

$$x_i < Q_1 - 1.5 \cdot \text{IQR} \quad \text{or} \quad x_i > Q_3 + 1.5 \cdot \text{IQR}. \tag{2}$$

This filtering technique provides a simple yet effective means of detecting and removing extreme values, allowing for improved data quality and greater consistency across features and classes in subsequent modeling tasks.

In this study, the IQR method was employed to identify and remove extreme values from continuous features. Following established practices in the literature, the procedure was applied independently to each numerical attribute within each class. This dual-level filtering—by feature and by class—preserves the internal structure of each category while enabling context-aware outlier elimination. As a result, 1,617 instances were removed, reducing the dataset from 13,611 to 11,994 samples.

As shown in Table 1, the classes with the highest number of removed instances were *Dermason*, *Horoz*, and *Seker*, which also correspond to the most represented varieties in the original dataset. Interestingly, a non-negligible proportion of outliers was also found in the minority classes *Bombay* and *Barbunya*, which could have further impacted class imbalance if not corrected.

TABLE 1. Outlier removal by class.

Class	Before	After	Removed
Barbunya	1322	1159	163
Bombay	522	464	58
Cali	1630	1470	160
Dermason	3546	3147	399
Horoz	1928	1623	305
Seker	2027	1730	297
Sira	2636	2351	285
Total	**13,611**	**11,994**	**1,617**

2.1.2. *Class balancing*

Class imbalance is common in supervised learning. It becomes more critical in multiclass classification tasks. When one or more categories are underrepresented, models tend to focus on the dominant classes. As a result, predictions for minority groups are often poor [27, 28]. This imbalance can distort decision boundaries. It may reduce the model's sensitivity and make the output less reliable [29]. To reduce these effects, SMOTE (Synthetic Minority Over-sampling Technique) is often used. It does not duplicate existing data. Instead, it creates new synthetic instances by interpolating between a sample and its k-nearest neighbors [30]. This process is described in Eq. (3).

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{\text{knn}} - \mathbf{x}_i), \quad \lambda \in [0,1], \tag{3}$$

where $\mathbf{x}_i$ is an existing minority class sample, $\mathbf{x}_{\text{knn}}$ is one of its k nearest neighbors, and λ is a random number controlling the position of the synthetic instance between the two points [27]. Overall, this type of oversampling technique remains a fundamental approach for mitigating class imbalance. By enriching the minority class distribution without introducing data redundancy, it helps ensure that learning algorithms are trained on more balanced and representative datasets. To counter this, the SMOTE algorithm was applied to increase the number of samples in the minority classes. After removing anomalous points, synthetic samples were added for

the less represented categories. Each class was expanded to include 3,147 samples, matching the count of the most abundant class. As a result, the final dataset included 22,029 entries, equally distributed across all groups.This balancing step contributed to more reliable model performance, particularly in metrics that are sensitive to class prevalence. The resulting class distribution is shown in Fig. 2. As we can see in this figure, the Dry Bean dataset presented a clear imbalance in class representation. In particular, varieties such as *Bombay* and *Barbunya* appeared in much lower quantities compared to *Dermason*, the most frequent class.

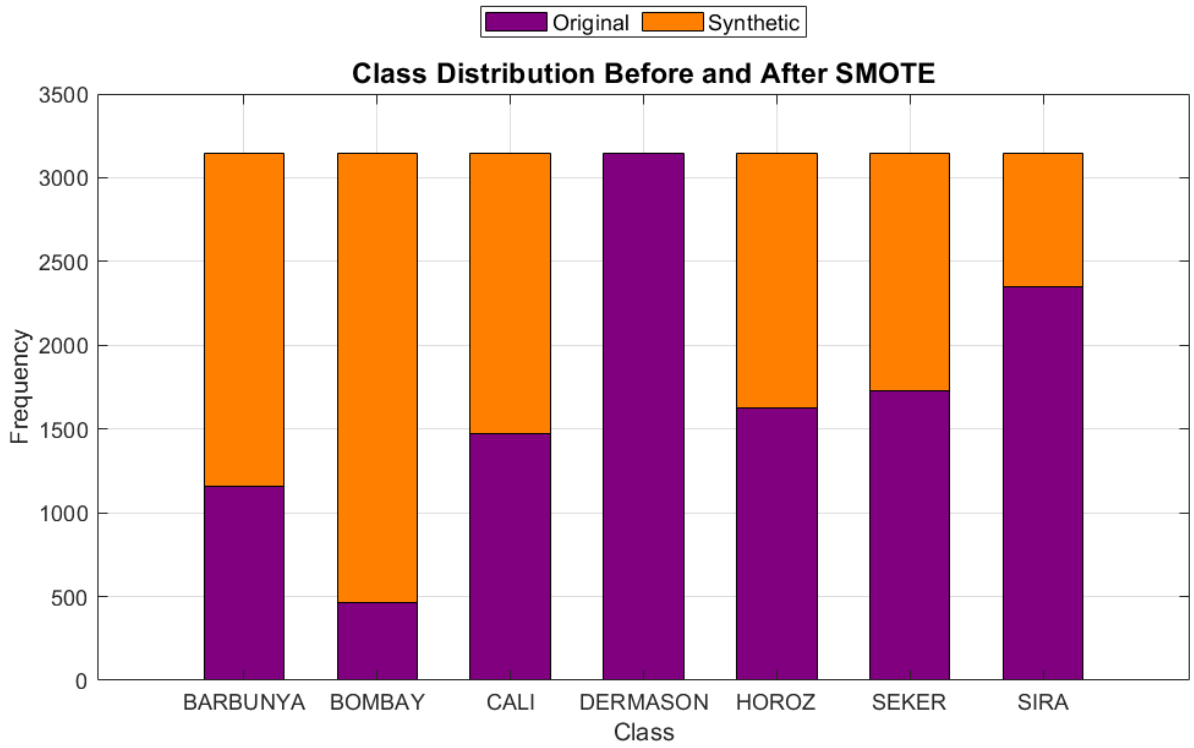


FIGURE 2. Sample distribution per class after applying SMOTE.

2.1.3. *Feature normalization*

Feature normalization is a fundamental preprocessing step in machine learning, particularly in classification tasks where input variables vary significantly in scale. Without proper normalization, features with larger numerical ranges can dominate the learning process, resulting in biased training and suboptimal model performance [31, 32]. To address this issue, Z-score normalization is commonly employed. This technique transforms each feature to have a mean of zero and a standard deviation of one, using the formula by Eq. (4).

$$X_{\text{new}} = \frac{X - \mu}{\sigma}, \tag{4}$$

where μ and σ represent the mean and standard deviation of the original feature, respectively. This transformation ensures that all variables contribute equally during model training, particularly in algorithms sensitive to distance metrics or gradient magnitudes, such as WKNN, SVM, and neural networks [31, 19]. In this study, Z-score normalization was applied to all continuous variables after the balancing step. This was necessary because the dataset contains features with diverse value ranges, and normalization ensures that all variables contribute equally during the learning process, improving model stability and performance.

2.2. Training and validation

Consistent and unbiased evaluation of classification models often relies on validation schemes that assess generalization performance on unseen data. Among the most widely adopted approaches are the Hold-Out method (HO) and k-Fold Cross-Validation (CV), which partition the dataset into training and testing subsets to estimate predictive accuracy. These strategies have demonstrated effectiveness across a variety of domains, including medical imaging, fault diagnosis, and imbalanced classification tasks [33, 34, 35].

In the Hold-Out approach (HO), the dataset is randomly partitioned into two disjoint subsets, typically allocating 80% for training and 20% for testing. This method is computationally efficient and well-suited for large-scale datasets or preliminary evaluations. However, its sensitivity to the specific data split may lead to high variance in performance estimates [34].

In contrast, k-Fold Cross-Validation (CV) divides the dataset into k parts of equal size, commonly with $k = 10$. At each iteration, one partition is held out for evaluation while the remaining $k-1$ subsets are used to train the model. The average performance across all folds provides a more stable and less biased estimate of generalization [33]. This strategy is especially advantageous when dealing with small sample sizes or heterogeneous feature distributions, as it maximizes data utilization and reduces the variance of performance estimates [35].

During the experiments performed in this work, these two evaluation strategies were implemented to ensure consistent model assessment: a standard hold-out split (80/20) and 10-fold

cross-validation. Given the multiclass structure and the original class imbalance present in the dataset, the evaluation prioritized macro averaged metrics and class wise precision.

2.3. Weighted K-Nearest Neighbors (WKNN)

The WKNN algorithm improves upon the standard KNN by giving greater influence to closer neighbors. This weighting is based on distance, allowing the model to better reflect the local structure of the data [36, 37]. It is particularly useful when the data contains irregular distributions or complex class boundaries. In general, this algorithm tends to perform well when class distributions are unbalanced or when some categories share similar features. In such cases, combining distance with contextual information can improve predictive accuracy [38].

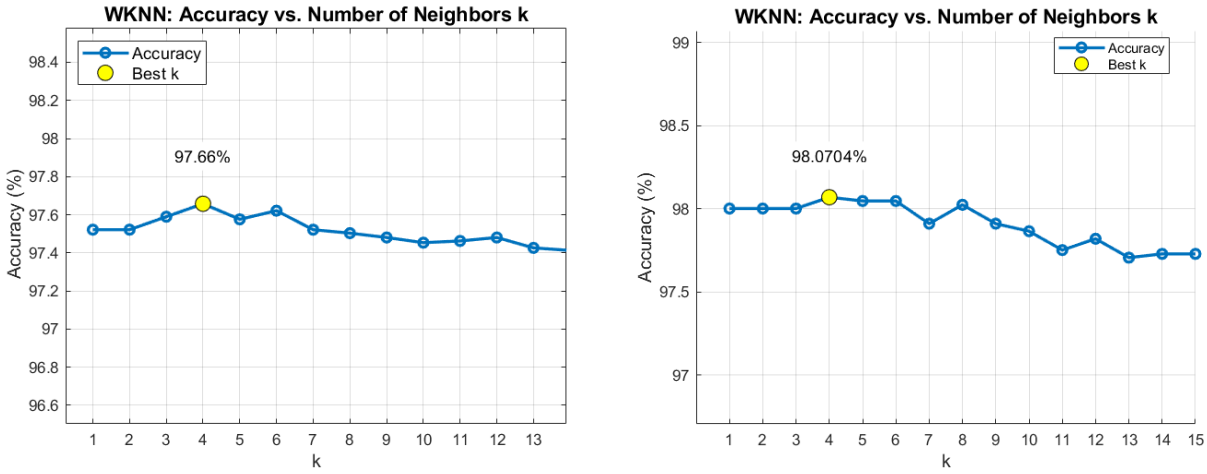
One of the main motivations behind WKNN is to overcome the equal-contribution limitation of traditional KNN. To address this, several weighting schemes have been proposed, including inverse distance, squared inverse distance, and exponential decay [39, 40]. In this work, Eq. 5 was applied.

$$w_j = \frac{1}{d_j^2 + \varepsilon}, \tag{5}$$

where d_j is the distance to the j-th neighbor, and ε is a small constant added to prevent numerical instability.

As previously mentioned, for this study, two validation schemes were used: 10-fold cross-validation and an 80/20 hold-out split. Addionally, Manhattan distance was selected due to its robustness in high-dimensional spaces. The same weighting function was applied in both settings.

To find the optimal number of neighbors, a search was conducted with k values from 1 to 15. As shown in Figure 3, the model reached its highest performance with $k = 4$, achieving 97.65% accuracy in cross-validation and 98.07% in the hold-out evaluation. This classifier was implemented in MATLAB R2024a, ensuring full integration with the proposed preprocessing pipeline.



(a) Cross Validation (10-fold) (b) Hold-Out (80/20 split)

FIGURE 3. Accuracy comparison for $k = 1$ to 15 using WKNN (Manhattan distance).

2.4. Evaluation metrics

Assessing a classifier's performance requires more than a single metric. Different measures highlight how the model handles imbalanced data, shifting distributions, and error types. Among the most widely used are accuracy, precision, recall, and F1-score, which provide insight into the model's ability to correctly identify each class. Table 2 provides definitions for these metrics.

In multiclass problems, relying solely on standard metrics can be misleading especially when class frequencies vary [41]. To account for this, macro-averaging was adopted in this study. Macro-averaging evaluates each class separately and then takes the average across all classes. This approach prevents majority classes from dominating the results and ensures that every category contributes equally, regardless of how often it appears in the data. It is especially useful when underrepresented classes carry meaningful information.

Table 3 reports the macro-averaged metrics utilized in this work, enabling a more balanced comparison across all target categories [41, 42, 43]. In all formulas, the following symbols are used: TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives). For macro-averaged metrics, K denotes the total number of classes, and the subscript i refers to the values corresponding to class i. The terms p_i and r_i denote the precision and recall for class i, respectively.

TABLE 2. Binary classification metrics and their interpretation.

Measure	**Formula**	**Evaluation focus**
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Overall correct predictions
Error Rate	$\frac{FP+FN}{TP+FP+TN+FN}$	Overall misclassification rate
Recall (r)	$\frac{TP}{TP+FN}$	Sensitivity; true positive rate
Specificity	$\frac{TN}{TN+FP}$	True negative rate
Precision (p)	$\frac{TP}{TP+FP}$	Positive predictive value
F1-Score	$\frac{2\cdot p\cdot r}{p+r}$	Harmonic mean of precision and recall

TABLE 3. Macro-averaged metrics for multi-class classification.

Measure	**Formula**	**Evaluation focus**
Macro Accuracy	$\frac{1}{K}\sum_{i=1}^{K}\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}$	Mean correct classification across classes
Macro Error Rate	$\frac{1}{K}\sum_{i=1}^{K}\frac{FP_i+FN_i}{TP_i+TN_i+FP_i+FN_i}$	Mean misclassification across classes
Macro Precision	$\frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i+FP_i}$	Mean precision per class
Macro Recall	$\frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i+FN_i}$	Mean recall per class
Macro F1-Score	$\frac{1}{K}\sum_{i=1}^{K}\frac{2\cdot p_i\cdot r_i}{p_i+r_i}$	Mean F1-Score per class

3. RESULTS

Two complementary experiments were rigorously designed to evaluate the proposed classification pipeline. All models were trained and tested under the same preprocessing sequence—class-wise outlier removal, Z-score normalization, and class balancing via SMOTE. Robustness was assessed with two validation schemes: 10-fold cross-validation and an 80/20 hold-out split. The first experiment compares a set of classical machine learning supervised classifiers under both schemes, with particular attention to the WKNN, adopted as the primary classifier in this study. The second experiment benchmarks the proposed approach against recent state-of-the-art approaches reported in the literature.

3.1. Evaluation of Classical Classifiers

This experiment evaluated five supervised classifiers: K-Nearest Neighbors (KNN), Weighted K-Nearest Neighbors (WKNN), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). Table 4 summarizes their global performance metrics under both hold-out and cross-validation schemes. As it can be seen in this table, the distance-based approaches achieved the highest overall performance, with both KNN and WKNN showing strong results. The MLP classifier yielded stable results across both validation strategies, with 96.96% accuracy in the hold-out evaluation and 96.66% under cross-validation. To determine optimal configurations, both one and two hidden-layer architectures were tested. The best setups were two hidden-layers with 128 and 64 neurons, respectively for cross-validation, and 16 and 8 neurons, for the hold-out evaluation. Training used a learning rate of 0.3, 1000 epochs, cross-entropy loss, and early stopping criteria. XGBoost achieved 97.89% accuracy in the hold-out setting and 97.48% under cross-validation. The best configuration was identified through grid search, with a learning rate of 0.2, tree depth of 8, and 200 estimators. The model employed the multi:softmax objective to support multiclass classification. By contrast, SVM obtained the lowest accuracy scores among the tested methods—96.75% in the hold-out evaluation and 96.64% in cross-validation. This model was evaluated with default parameters.

From this table, we can highlight that the best-performing model was the WKNN reaching a 98.07% of accuracy in the hold-out setting and 97.65% in cross-validation. These results confirm the effectiveness of instance-based methods in leveraging the geometric structure of the dataset.

TABLE 4. Global performance metrics under Hold-Out (HO) and Cross-Validation (CV).

Validation	Model	Acc.	Err.	Prec.	Rec.	Spe.	F1
Hold-Out (HO)	KNN	98.00	2.00	98.00	98.00	99.67	98.00
	WKNN	**98.07**	**1.93**	**98.07**	**98.07**	**99.67**	**98.06**
	MLP	96.96	3.04	96.96	96.96	99.49	96.96
	XGBoost	97.89	2.11	97.89	97.89	99.65	97.89
	SVM	96.75	3.25	96.76	96.75	99.46	96.75
Cross-Validation (CV)	KNN	97.57	2.43	97.58	97.57	99.60	97.57
	WKNN	**97.65**	**2.35**	**97.66**	**97.65**	**99.61**	**97.65**
	MLP	96.66	3.34	96.68	96.69	99.45	96.68
	XGBoost	97.48	2.52	97.48	97.48	99.58	97.48
	SVM	96.64	3.36	96.64	96.64	99.46	96.64

Overall, the results confirm the consistent advantage of distance-based classifiers in this context, followed closely by XGBoost, with MLP and SVM showing reliable but comparatively lower performance.

3.2. Discussion

Table 5 benchmarks the best-performing WKNN against recent state-of-the-art contributions that also employed the Dry Bean dataset. These methods were selected for comparison because they represent recent advances reported in the literature while relying on the same dataset, ensuring a fair and consistent evaluation. The symbol "—" in this table indicates that the study did not report results for the given validation scheme. In our case, both validation protocols were applied, showing that the proposed instance-based models achieved superior results: WKNN reached 98.07% accuracy under the hold-out scheme and 97.65% with cross-validation. These outcomes surpass the results reported by Koklu et al. (2020) and Macuácua et al. (2023), as

well as more recent strategies such as ADASYN+XGBoost (Khan et al., 2023), Random Oversampling + Random Forest with randomized search tuning (Koeshardianto et al., 2023), and BLSMOTE+SVM (Lee et al., 2024). These findings highlight that competitive performance can be achieved without the need for more complex architectures, offering a practical alternative for scenarios with limited computational resources.

TABLE 5. Benchmark comparison between recent studies and the proposed models. The symbol "—" indicates that the study did not report results for that validation scheme.

Reference	**Best Classifier**	**Acc. (HO)**	**Acc. (CV)**
Koklu et al. (2020) [17]	SVM	—	93.13
Macuácua et al. (2023) [18]	SMOTE + KNN	95.90	—
Khan et al. (2023) [19]	ADASYN + XGB	95.40	—
Nayak et al. (2023) [20]	SMOTE + XGB	97.32	—
Koeshardianto et al. (2023) [21]	Random OS + RF	96.58	—
Lee et al. (2024) [22]	BLSMOTE + K-means + SVM	—	97.54
Proposed	**SMOTE + WKNN**	**98.07**	**97.65**

4. CONCLUSION

This study developed and evaluated a classification pipeline for dry bean varieties that integrates class-wise and feature-wise outlier filtering, Z-score normalization, SMOTE-based class balancing, and distance-based classifiers. The methodology was tested under two widely used validation schemes—10-fold cross-validation and 80/20 hold-out—ensuring a comprehensive evaluation of generalization performance.

The experimental results demonstrate that distance-based classifiers, particularly WKNN, consistently achieved the highest performance across both validation settings. Among them, WKNN emerged as the best-performing model, reaching 98.07% accuracy in the hold-out setting and 97.65% in cross-validation, surpassing recent state-of-the-art approaches reported in the literature. These outcomes confirm that when appropriate preprocessing strategies—such

as normalization and targeted outlier removal—are applied, simpler classifiers can rival or even outperform more complex ensemble or neural network models.

A detailed comparison of the evaluated models shows that distance-based approaches are especially effective in structured morphological datasets, where class boundaries are well defined and the feature space benefits from geometric proximity metrics. This observation aligns with prior studies in agricultural classification, reinforcing the suitability of instance-based methods for this type of problem. Moreover, the proposed pipeline maintained model simplicity and interpretability, important features for practical deployment in agricultural systems.

From an application standpoint, the combination of low computational cost and high accuracy makes the proposed method suitable for resource-constrained environments. Potential implementations include embedded sorting systems and on-field inspection devices, where efficient operation with limited hardware resources is essential. These characteristics position the approach as a practical solution for real-time classification tasks in agricultural production chains.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] W. Liu, S. Zeng, G. Wu, H. Li, F. Chen, Rice Seed Purity Identification Technology Using Hyperspectral Image with LASSO Logistic Regression Model, Sensors 21 (2021), 4384. https://doi.org/10.3390/s21134384.

[2] M. Dogan, Y.S. Taspinar, I. Cinar, R. Kursun, I.A. Ozkan, et al., Dry Bean Cultivars Classification Using Deep CNN Features and Salp Swarm Algorithm Based Extreme Learning Machine, Comput. Electron. Agric. 204 (2023), 107575. https://doi.org/10.1016/j.compag.2022.107575.

[3] C.H.R. Mendigoria, H.L. Aquino, O.J.Y. Alajas, R.S. Concepcion II, E.P. Dadios, et al., Varietal Classification of Lactuca Sativa Seeds Using an Adaptive Neuro-Fuzzy Inference System Based on Morphological Phenes, J. Adv. Comput. Intell. Intell. Inform. 25 (2021), 618–624. https://doi.org/10.20965/jaciii.2021.p0618.

[4] S. Ermiş, U. Ercan, A. Kabaş, Ö. Kabaş, G. Moiceanu, Machine Learning-Based Morphological Classification and Diversity Analysis of Ornamental Pumpkin Seeds, Foods 14 (2025), 1498. https://doi.org/10.3390/foods14091498.

[5] S. Barrera, J.C. Berny Mier y Teran, J.D. Lobaton, R. Escobar, P. Gepts, et al., Large Genomic Introgression Blocks of Phaseolus parvifolius Freytag Bean into the Common Bean Enhance the Crossability between Tepary and Common Beans, Plant Direct 6 (2022), e470. https://doi.org/10.1002/pld3.470.

[6] M. El-Fouly, E.-Z. Abou El-Nour, Foliar Feeding with Micronutrients to Overcome Adverse Salinity Effects on Growth and Nutrients Uptake of Bean (Phaseolus vulgaris), Egypt. J. Agron. 43 (2021), 1–12. https://doi.org/10.21608/agro.2021.49359.1238.

[7] L. Sinkovič, V. Blažica, B. Blažica, V. Meglič, B. Pipan, How Nutritious Are French Beans (Phaseolus vulgaris L.) from the Citizen Science Experiment?, Plants 13 (2024), 314. https://doi.org/10.3390/plants13020314.

[8] I. Borromeo, F. Domenici, C. Giordani, M. Del Gallo, C. Forni, Enhancing Bean (Phaseolus vulgaris L.) Resilience: Unveiling the Role of Halopriming against Saltwater Stress, Seeds 3 (2024), 228–250. https://doi.org/10.3390/seeds3020018.

[9] R.L. Nelson, P. Wang, Variation and Evaluation of Seed Shape in Soybean, Crop Sci. 29 (1989), 147–150. https://doi.org/10.2135/cropsci1989.0011183X002900010033x.

[10] Y. Hsiao, T.I. Chuang, Seed-Coat Morphology and Anatomy in Collomia (Polemoniaceae), Am. J. Bot. 68 (1981), 1155–1164. https://doi.org/10.1002/j.1537-2197.1981.tb07821.x.

[11] H.J. Hill, A.G. Taylor, T.G. Min, Density Separation of Imbibed and Primed Vegetable Seeds, J. Am. Soc. Hortic. Sci. 114 (1989), 661–665. https://doi.org/10.21273/JASHS.114.4.661.

[12] İ. Kılıç, N. Yalçın, A Novel Hybrid Methodology Based on Transfer Learning, Machine Learning, and ReliefF for Chickpea Seed Variety Classification, Appl. Sci. 15 (2025), 1334. https://doi.org/10.3390/app15031334.

[13] X. Jin, Y. Zhao, H. Wu, T. Sun, Sunflower Seeds Classification Based on Sparse Convolutional Neural Networks in Multi-Objective Scene, Sci. Rep. 12 (2022), 19890. https://doi.org/10.1038/s41598-022-23869-4.

[14] S. Majumdar, D.S. Jayas, Classification of Cereal Grains Using Machine Vision: I. Morphology Models, Trans. ASAE 43 (2000), 1669–1675. https://doi.org/10.13031/2013.3107.

[15] J. Paliwal, N.S. Visen, D.S. Jayas, N.D.G. White, Cereal Grain and Dockage Identification Using Machine Vision, Biosyst. Eng. 85 (2003), 51–57. https://doi.org/10.1016/S1537-5110(03)00034-5.

[16] R.H. Ajaz, L. Hussain, Seed Classification Using Machine Learning Techniques, J. Multidiscip. Eng. Sci. Technol. 2 (2015), 1098–1102. www.jmest.org.

[17] M. Koklu, I.A. Ozkan, Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques, Comput. Electron. Agric. 174 (2020), 105507. https://doi.org/10.1016/j.compag.2020.105507.

[18] J.C. Macuácua, J.A.S. Centeno, C. Amisse, Data Mining Approach for Dry Bean Seeds Classification, Smart Agric. Technol. 5 (2023), 100240. https://doi.org/10.1016/j.atech.2023.100240.

[19] M.S. Khan, T.D. Nath, M.M. Hossain, A. Mukherjee, H.B. Hasnath, et al., Comparison of Multiclass Classification Techniques Using Dry Bean Dataset, Int. J. Cogn. Comput. Eng. 4 (2023), 6–20. https://doi.org/10.1016/j.ijcce.2023.01.002.

[20] J. Nayak, P.B. Dash, B. Naik, An Advance Boosting Approach for Multiclass Dry Bean Classification, J. Eng. Sci. Technol. Rev. 12 (2023), 107–115. https://doi.org/10.25103/jestr.162.14.

[21] M. Koeshardianto, K.E. Permana, D.S.Y. Kartika, W. Setiawan, Beans Classification Using Decision Tree and Random Forest with Randomized Search Hyperparameter Tuning, Commun. Math. Biol. Neurosci. 2023 (2023), 118. https://doi.org/10.28919/cmbn/8225.

[22] C.Y. Lee, W. Wang, J.Q. Huang, Clustering and Classification for Dry Bean Feature Imbalanced Data, Sci. Rep. 14 (2024), 31058. https://doi.org/10.1038/s41598-024-82253-6.

[23] T.T.H. Phan, L.H.B. Nguyen, Enhancing Rice Seed Purity Recognition Accuracy Based on Optimal Feature Selection, Ecol. Inform. 86 (2025), 103044. https://doi.org/10.1016/j.ecoinf.2025.103044.

[24] C.S. Dash, A.K. Behera, S. Dehuri, A. Ghosh, An Outliers Detection and Elimination Framework in Classification Task of Data Mining, Decis. Anal. J. 6 (2023), 100164. https://doi.org/10.1016/j.dajour.2023.100164.

[25] H. Torkey, E. Ibrahim, E. Hemdan, A. El-Sayed, M.A. Shouman, Diabetes Classification Application with Efficient Missing and Outliers Data Handling Algorithms, Complex Intell. Syst. 8 (2021), 237–253. https://doi.org/10.1007/s40747-021-00349-2.

[26] J. Jung, D. Kim, H. Cho, M. Lee, J. Jeong, et al., Multi-Algorithmic Approach for Detecting Outliers in Cattle Intake Data, J. Agric. Food Res. 15 (2024), 101021. https://doi.org/10.1016/j.jafr.2024.101021.

[27] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Intell. Res. 16 (2002), 321–357. https://doi.org/10.1613/jair.953.

[28] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, in: D.S. Huang, X.P. Zhang, G.B. Huang, (eds) Advances in Intelligent Computing, Lecture Notes in Computer Science, Vol 3644, Springer, Berlin, Heidelberg, (2005). https://doi.org/10.1007/11538059_91.

[29] G. Husain, D. Nasef, R. Jose, J. Mayer, M. Bekbolatova, et al., SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models, Algorithms 18 (2025), 37. https://doi.org/10.3390/a18010037.

[30] M. Kivrak, U. Avci, H. Uzun, C. Ardic, The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients, Diagnostics 14 (2024), 2634. https://doi.org/10.3390/diagnostics14232634.

[31] A. Tukhtaev, D. Turimov, J. Kim, W. Kim, Feature Selection and Machine Learning Approaches for Detecting Sarcopenia Through Predictive Modeling, Mathematics 13 (2024), 98. https://doi.org/10.3390/math13010098.

[32] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, Z. Siddique, Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance, Technologies 9 (2021), 52. https://doi.org/10.3390/technologies9030052.

[33] T.J. Bradshaw, Z. Huemann, J. Hu, A. Rahmim, A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging, Radiol. Artif. Intell. 5 (2023), e220232. https://doi.org/10.1148/ryai.220232.

[34] H. Bragança, J.G. Colonna, H.A.B.F. Oliveira, E. Souto, How Validation Methodology Influences Human Activity Recognition Mobile Systems, Sensors 22 (2022), 2360. https://doi.org/10.3390/s22062360.

[35] J. Allgaier, R. Pryss, Cross-Validation Visualized: A Narrative Guide to Advanced Methods, Mach. Learn. Knowl. Extr. 6 (2024), 1378–1388. https://doi.org/10.3390/make6020065.

[36] S.A. Dudani, The Distance-Weighted k-Nearest-Neighbor Rule, IEEE Trans. Syst. Man Cybern. SMC-6 (1976), 325–327. https://doi.org/10.1109/TSMC.1976.5408784.

[37] S. Li, K. Zhang, Q. Chen, S. Wang, S. Zhang, Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm, IEEE Access 8 (2020), 139512–139528. https://doi.org/10.1109/ACCESS.2020.3012768.

[38] B. Karabulut, G. Arslan, H.M. Ünver, A Weighted Similarity Measure for k-Nearest Neighbors Algorithm, Celal Bayar Üniv. Fen Bilim. Derg. 15 (2019), 393–400. https://doi.org/10.18466/cbayarfbe.618964.

[39] L. Xiang, Y. Xu, J. Cui, Y. Liu, R. Wang, et al., GM(1,1)-Based Weighted K-Nearest Neighbor Algorithm for Indoor Localization, Remote Sens. 15 (2023), 3706. https://doi.org/10.3390/rs15153706.

[40] E. Hofer, M. v. Mohrenschildt, Locally-Scaled Kernels and Confidence Voting, Mach. Learn. Knowl. Extr. 6 (2024), 1126–1144. https://doi.org/10.3390/make6020052.

[41] M. Sokolova, G. Lapalme, A Systematic Analysis of Performance Measures for Classification Tasks, Inf. Process. Manag. 45 (2009), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

[42] S. Farhadpour, T.A. Warner, A.E. Maxwell, Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices, Remote Sens. 16 (2024), 533. https://doi.org/10.3390/rs16030533.

[43] M.C. Hinojosa Lee, J. Braet, J. Springael, Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores, Appl. Sci. 14 (2024), 9863. https://doi.org/10.3390/app14219863.