Opsi
Vol. 18, No. 2, December 2025
p-ISSN 1693-2102
e-ISSN 2686-2352

Feature-based classification of sugarcane quality using the K-nearest neighbor algorithm

Nur Indrianti [1]*, Muhammad Iqbal[1], Heru Cahya Rustamaji [2], Andrey Ferriyan [2], Panut Mulyono [3] , Moh. Ais Ananta[1]

[1]Department of Industrial Engineering, Faculty of Industrial Engineering, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[2]Department of Informatics, Faculty of Industrial Engineering, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[3]Department of Chemical Engineering, Faculty of Engineering, Universitas Gadjah Mada, Indonesia

***Corresponding Author**: n.indrianti@upnyk.ac.id

Article history:
Received: 24 November 2025
Revised: 16 December 2025
Accepted: 16 December 2025
Published: 30 December 2025

Keywords:
Sugarcane quality
Bululawang variety
K-Nearest Neighbor
Non-destructive classification
Sustainable agro-industry

ABSTRACT

The rapid advancement of artificial intelligence has enabled practical, data-driven approaches to agricultural quality assessment. However, many existing methods rely on complex sensor systems that are costly and difficult to deploy in the field. This study proposes a lightweight and interpretable K-Nearest Neighbor (KNN) model for non-destructive evaluation of sugarcane milling feasibility using five easily measurable physical attributes: relative distance ratio, internode length, mean diameter, circumference, and weight per centimeter. Samples with Brix less than 16 are categorized as not feasible for milling, while Brix equal to or greater than 16 are classified as possible. A dataset of 1,889 Bululawang samples collected in Malang, East Java, Indonesia, was evaluated across twenty-two scenarios that varied the train-test split, normalization method, distance metric, and neighborhood size. The optimal configuration, consisting of an 80:20 split, Standard normalization, the Minkowski distance metric, and k=75, achieved an accuracy of 78%. The findings confirm that physical measurements can serve as effective predictors of sugarcane quality and support data-driven inspection and sustainable resource utilization in line with SDGs 2, 9, and 12.

DOI:
https://doi.org/10.31315/opsi.v18i2.16000

1. INTRODUCTION

Sugarcane (Saccharum officinarum L.) is one of the most essential agro-industrial crops, serving as a significant source of sugar and bioethanol and supporting the livelihoods of millions of farmers worldwide [1]. Beyond its economic value, sugarcane plays a strategic role in food security and renewable energy initiatives in many countries, including Indonesia. Its commercial value depends mainly on sucrose

concentration, expressed as Brix, which influences sugar recovery and milling efficiency. However, sugarcane deteriorates rapidly after harvest due to respiration, moisture loss, and biochemical inversion of sucrose [1], [2]. In many production systems, quality evaluation continues to rely on destructive juice sampling and laboratory testing, which are time-consuming and unsuitable for making rapid decisions before milling [3]. Delays between harvesting and crushing further reduce cane purity and sugar yield [4], and storage beyond twenty-four to forty-eight hours significantly decreases pol, purity, and recoverable sugar, especially in burnt canes [3]. These conditions underscore the need for a rapid, non-destructive, and field-friendly method for assessing sugarcane quality.

Recent advancements in artificial intelligence have introduced image-based and deep learning approaches for predicting plant maturity and biochemical traits. Although these methods have demonstrated high accuracy in controlled environments, field deployment remains challenging because they require large datasets, substantial computational capacity, and complex sensor systems [5]. In practice, the South African Sugarcane Research Institute [6] reports that quality grading still depends on destructive sampling and subjective visual inspection. Onboard near-infrared spectroscopy has shown potential for real-time cane quality measurement [7], but the associated costs and calibration requirements limit its accessibility for small- and medium-sized growers. This gap between advanced sensing technologies and practical field-level methods underscores the need for accessible, scalable alternatives.

Feature-based classification using measurable morphometric attributes provides an accessible alternative. Physical parameters such as internode length, diameter, circumference, weight per unit length, and internode spacing can be collected without specialized laboratory instruments, making them suitable for rapid field-level screening. The KNN algorithm aligns well with this approach because it classifies samples based on distance similarity and does not require an explicit model training process [8]. KNN has demonstrated strong performance in agricultural quality assessment with low-dimensional continuous features [9] and several studies have confirmed its accuracy and ease of deployment under practical conditions [10]–[12]. Due to its interpretability and resilience to natural biological variability, KNN offers a realistic foundation for developing non-destructive sugarcane quality classifiers [13].

Machine learning applications for agricultural quality assessment continue to grow, using Support Vector Machines, Random Forests, and KNN in diverse crop contexts. Table 1 summarizes previous research related to this study across three major categories: image-based classification, sensor-based modeling, and algorithmic developments to enhance KNN performance. Although these studies demonstrate the versatility of KNN, most rely on high-resolution imagery, specialized sensors, or complex preprocessing pipelines. Only a limited number of studies have applied KNN to directly measurable physical attributes of sugarcane, leaving a clear gap in the development of non-destructive and low-cost Brix-based quality classifiers.

Building on the reviewed literature, this study addresses the identified gap by developing a KNN-based classification model that uses readily measurable physical attributes of sugarcane internodes to classify stalks as milling-feasible or non-feasible. These field-measurable characteristics provide a practical foundation for rapid and non-destructive quality assessment without the need for complex sensors or laboratory instruments. The proposed approach enhances data-driven quality control and facilitates timely decision-making within agricultural supply chains by enabling rapid field-level screening and reducing reliance on destructive testing. In this way, the study positions simple morphometric data as a viable input for intelligent quality assessment, bridging the gap between laboratory-oriented analysis and operational field requirements. Furthermore, the framework aligns with global initiatives that promote sustainable and inclusive innovation, contributing to the achievement of Sustainable Development Goals related to food security, industrial innovation, and responsible production.

The remainder of this paper is organized as follows. Section 2 presents the materials and methods used in this study, including the dataset, measurement procedures, feature selection, and the KNN classification framework. Section 3 presents the experimental results obtained with the proposed approach, including scenario-based evaluation and classification performance. Section 4 discusses the findings with respect to methodological validity, industrial implications, and practical applicability in field conditions. Finally, Section 5 concludes the paper by summarizing the main contributions of the study and outlining recommendations for future research and potential industrial implementation. This structured presentation is intended to guide readers systematically from problem formulation through empirical evaluation to applied insights, ensuring clarity and coherence throughout the manuscript.

Table 1. Summary of previous KNN-based studies and identified research gaps

No	Authors (Year)	Research Objective	Research Object	Main Parameters	Methods	Weaknesses / Limitations	Research Gap
1	Sudipa et al. [11]	Develop a KNN-based model for fruit quality classification using physical attributes	Fruit datasets (Kaggle)	Size, weight, sweetness, acidity, ripeness	KNN with 5-fold CV	Generic dataset; not crop-specific	Lacks application to field-based crops and quantitative sugar quality (°Brix)
2	Subbarao et al. [27]	Classify dry bean varieties using morphometric features via KNN and NN	13,611 dry bean samples	Area, perimeter, ratio, compactness	KNN, Neural Network, MRMR, ReliefF	Tested only under lab conditions	No validation for perishable crops or quality classes based on chemical attributes
3	Gupta & Nahar [12]	Predict soil and crop quality using adaptive KNN and environmental data	Soil and sensor datasets	pH, NPK, temperature, and humidity	Adaptive KNN + Centroid + Feature selection	Focused on soil health, not crop product quality	Does not explore physical plant features for post-harvest quality evaluation
4	Mishra et al. [28]	Build a crop recommendation system using KNN and RF with agronomic parameters	Indian agricultural dataset (22 crops)	N, P, K, pH, rainfall, temperature	KNN, Random Forest	Focuses on recommendations, not classification of product quality	No connection to post-harvest quality or Brix-based classification
5	Li & Ercisli [20],	Propose data-efficient pest recognition using KNN Distance Entropy	Pest image dataset (6×1000 images)	Data quality, entropy, informative samples.	KNN Distance Entropy + ResNet-18	Relies on image data, not physical measures	Limited to pest recognition; lacks application to physical-quality data
6	Chao & Li [29]	Develop semi-supervised image classification using KNN Distance Entropy	Remote sensing dataset (NWPU-RS45)	Entropy, Euclidean distance, pseudo-labels.	KNN Distance Entropy + meta-learning	Restricted to the image domain; no agricultural variables	Does not address numeric, field-measured parameters relevant to crop quality
7	Zhang [30]	Analyze and optimize KNN parameters and distance metrics	Simulated datasets (non-agricultural)	K-value, distance metrics, weighting.	Comparative algorithmic analysis.	Theoretical; no agricultural application.	No empirical validation for crop datasets or physical features
8	Gurunathan et al. [22]	Detect leaf diseases using KNN with RGB and texture features	Images of healthy/diseased leaves	RGB, GLCM (entropy, contrast, energy)	Image preprocessing, GLCM, KNN	Dependent on color variation; non-physical features	Focuses on disease detection; not applicable to °Brix-based quality classification
9	Taner et al. [23]	Compare CNN and KNN for apple quality classification using deep features	Apple dataset (5808 images)	Texture, color, shape	Transfer learning (DenseNet201) + KNN	Image-based; lacks measurable field features	Does not examine lightweight, physical-feature models for field implementation
10	Current study	Classify sugarcane quality using physical parameters and KNN	Bululawang sugarcane, Malang	Length, diameter, circumference, weight/cm, internode spacing	KNN on normalized numeric dataset	–	Addresses the gap by applying KNN to physical, non-destructive, °Brix-based sugarcane quality classification

2. MATERIALS AND METHODS

2.1. Materials

The material used in this study is the Bululawang sugarcane variety, a high-yielding cultivar widely cultivated in East Java and known for its favorable sucrose accumulation characteristics [14], [15]. Sugarcane samples were obtained from farmers supplying a sugar mill in Malang City, East Java, Indonesia. A total of 1,889 internode samples were collected for analysis. To ensure consistency across samples, the cutting process began three internodes above the stalk base, which served as the zero point, and continued sequentially upward. Each internode was separated and measured individually to obtain its physical attributes and Brix value using a handheld refractometer.

Five measurable morphometric attributes were used as predictors of sugarcane quality, selected based on their established agronomic relevance to stalk growth, maturity, and sucrose accumulation. These attributes can be collected directly in the field without laboratory instruments, enabling rapid and non-destructive quality assessment. Previous studies have demonstrated that sugarcane quality varies along the stalk and that internode-level measurements, particularly internode length, are closely associated with Brix distribution and maturity gradients within individual canes [16]. Genetic and morphological analyses further indicate that internode length, stem diameter, and stalk girth are key contributors to stalk weight and yield performance [17]. In addition, correlation and path coefficient analyses reveal that stem diameter and Brix percentage influence yield indirectly through single cane weight, supporting the relevance of weight-related attributes as physical proxies for sucrose content [18]. Recent phenotypic studies also confirm that plant height and stem diameter are among the strongest morphological indicators associated with yield, reinforcing the importance of length- and diameter-based measurements in sugarcane quality assessment [19].

Internode spacing is defined as the linear distance between two consecutive nodes and reflects the distribution of maturity along the stalk. Internode length refers to the physical length of each internode segment measured using a ruler or digital caliper and typically increases with cane maturity. The mean internode diameter, measured at the midpoint of each segment, represents stalk robustness and is associated with juice yield potential. Internode circumference complements the diameter measurement by describing the cross-sectional structure and overall turgor of the stalk. Internode weight per centimeter is calculated by dividing the internode weight by its length and serves as an indicator of tissue compactness and moisture content, which are positively associated with sucrose accumulation.

To represent geometric position, the distance attribute was defined as the relative location of each internode along the stalk, calculated as the distance from the zero point to the internode midpoint and normalized by the total stalk length. Figure 1 illustrates the Bululawang sugarcane samples and the field data collection process carried out in this study.

Figure 1. Bululawang sugarcane and data collection activities

2.2. Methods

2.2.1 Dataset preparation

The classification of sugarcane quality using the KNN algorithm followed a structured process that included dataset preparation, algorithm configuration, classification processing, and performance evaluation. During dataset preparation, all numerical attributes were checked for completeness and consistency, and any missing or anomalous values were removed. The physical measurements of each internode were then standardized to ensure uniform scaling in the distance computation. Each feature was normalized to eliminate unit discrepancies and to prevent variables with larger numerical ranges from dominating the KNN calculations. The resulting feature matrix provided a consistent numerical representation of the physical morphology of the sugarcane samples.

For classification purposes, Brix values were grouped into two categories. Samples with Brix values less than sixteen were assigned to Class zero, indicating that they were not feasible for milling. In contrast, samples with Brix values of 16 or higher were assigned to Class 1, indicating they met the minimum sucrose requirement for efficient processing. Table 2 presents a subset of ten representative samples from the structured dataset, including their physical attributes, measured Brix values, and assigned quality classes.

Table 2. Sample dataset

Sequence No.	Sample No.	Distance Ratio	Length (cm)	Average Diameter (cm)	Circumference (cm)	Weight (g/cm)	Brix	Class
1	100101	0.031	17.0	3.90	11.0	9.529	16.2	1
2	100102	0.084	12.2	3.80	11.4	8.852	15.8	0
3	100103	0.142	19.2	3.80	11.1	8.438	16.4	1
4	100104	0.202	13.5	3.65	11.0	8.222	16.1	1
5	100105	0.259	18.0	3.55	10.8	8.222	16.4	1
6	100106	0.312	11.0	3.60	10.5	7.364	16.9	1
7	100107	0.361	15.6	3.55	10.4	7.179	15.5	0
8	100108	0.407	9.6	3.45	10.4	7.292	16.8	1
9	100109	0.451	14.7	3.45	10.4	7.143	16.5	1
10	100110	0.493	8.0	2.85	9.7	7.500	16.4	1

2.2.2 Scenario and evaluation metrics

To evaluate the robustness and generalizability of the proposed KNN model, a series of experimental scenarios was designed by systematically varying the dataset composition and key configuration parameters, as summarized in Table 3. Four experimental factors were examined: the proportion of training and testing data, the normalization method, the distance metric, and the number of nearest neighbors k. These factors were tested sequentially, with the optimal setting identified at each stage serving as the baseline for the subsequent stage.

Table 3. Scenarios used in this study

Group	Basis of Scenario	Scenario		
1	Train-test data split	KNN-1: 90:10	KNN-2: 85:15	KNN-3: 80:20
2	Normalization formula	KNN-4: Min max		
		KNN-5: Standard		
		KNN-6: Robust		
3	Distance formula	KNN-4: Euclidean		
		KNN-5: Cityblock (or Manhattan)		
		KNN-7: Minkowski ($p = 3$)		

Group	Basis of Scenario	Scenario		
4	Variation of *k* Values	KNN-8: *k*=1	KNN-13: *k*=15	KNN-18: *k*=300
		KNN-9: *k*=2	KNN-14: *k*=25	KNN-19: *k*=500
		KNN-10: *k*=3	KNN-15: *k*=50	KNN-20: *k*=750
		KNN-11: *k*=5	KNN-16: *k*=75	KNN-21: *k*=900
		KNN-12: *k*=10	KNN-17: *k*=100	KNN-22: *k*=1000

The evaluation began with Scenario Group 1, which assessed three train-test data splits. For example, under the 90:10 split, 1,700 of the 1,889 samples were allocated to training, and 189 to testing. The split that produced the highest accuracy was then selected as the baseline configuration for Scenario Group 2. The split that yielded the highest accuracy was selected as the baseline for the next scenario group.

Scenario Group 2 evaluated the effect of different normalization methods. The optimal data split from Group 1 was fixed, and each normalization method was applied to identify the approach that yielded the most consistent classification performance. The optimal normalization method was then used as the fixed input for Scenario Group 3, which identified the distance formulation that best represented feature similarity within the dataset.

Finally, Scenario Group 4 examined the influence of varying *k* values on model performance. The number of neighbors (*k*) was varied from 1 to 1000 in order to observe how local versus broad neighborhood structures affected classification stability.

This sequential evaluation framework ensured that the model configuration was progressively refined at each stage. The optimal parameter from each group was retained and carried forward, enabling efficient evaluation of the combined effects of data composition, feature scaling, distance formulation, and neighborhood size. This approach enabled the identification of the most accurate and stable configuration for sugarcane quality classification using the KNN method.

To assess the performance of the KNN classification model, a confusion matrix was used to summarize the relationship between predicted and actual class labels, as illustrated in Figure 2. In this matrix, True Positive (TP) refers to the number of samples correctly predicted as belonging to the milling feasible class (°Brix ≥ 16%), and True Negative (TN) refers to samples correctly identified as belonging to the non-feasible class (°Brix < 16%). Conversely, False Positive (FP) represents samples incorrectly predicted as feasible, while False Negative (FN) refers to samples that were feasible but incorrectly classified as non-feasible. These four components formed the basis for calculating the key performance metrics used in this study, including accuracy and error rate.

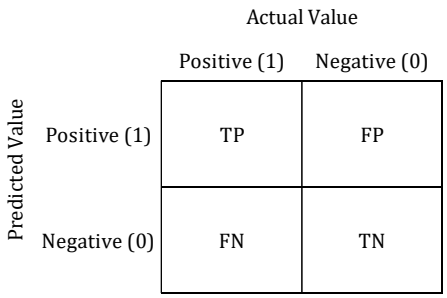

Figure 2. Confusion matrix

The classification model's accuracy was calculated as the proportion of correctly predicted samples, given by (TP + TN)/(TP + FP + FN + TN). Meanwhile, the error rate was defined as the proportion of incorrect predictions, calculated as (FP + FN) / (TP + FP + FN + TN). These two metrics were used to evaluate the overall predictive performance of the K-Nearest Neighbors model in this study.

2.2.3 Algorithm configuration

The KNN model and algorithm were implemented in Google Colab using Python. The program was designed to automate the workflow, including data preprocessing, normalization, distance computation, and

classification prediction based on the selected parameter settings. The complete modeling process is illustrated in Figure 3.

The modeling procedure presented in this study corresponds to the configuration used in the KNN-3 scenario described in the experimental setup. This configuration applies the min max normalization, the Minkowski distance metric with $p = 3$, and a neighborhood size of k =175.

The dataset was divided into training and testing subsets. Table 3 displays the first five records of the training data generated in Step 3, and Table 4 presents the first five records of the testing data produced in Step 5. These examples illustrate the structure of the normalized input used for model development and validation, in which Class 1 denotes milling-feasible sugarcane (°Brix ≥ 16) and Class 0 denotes non-feasible sugarcane (°Brix < 16).

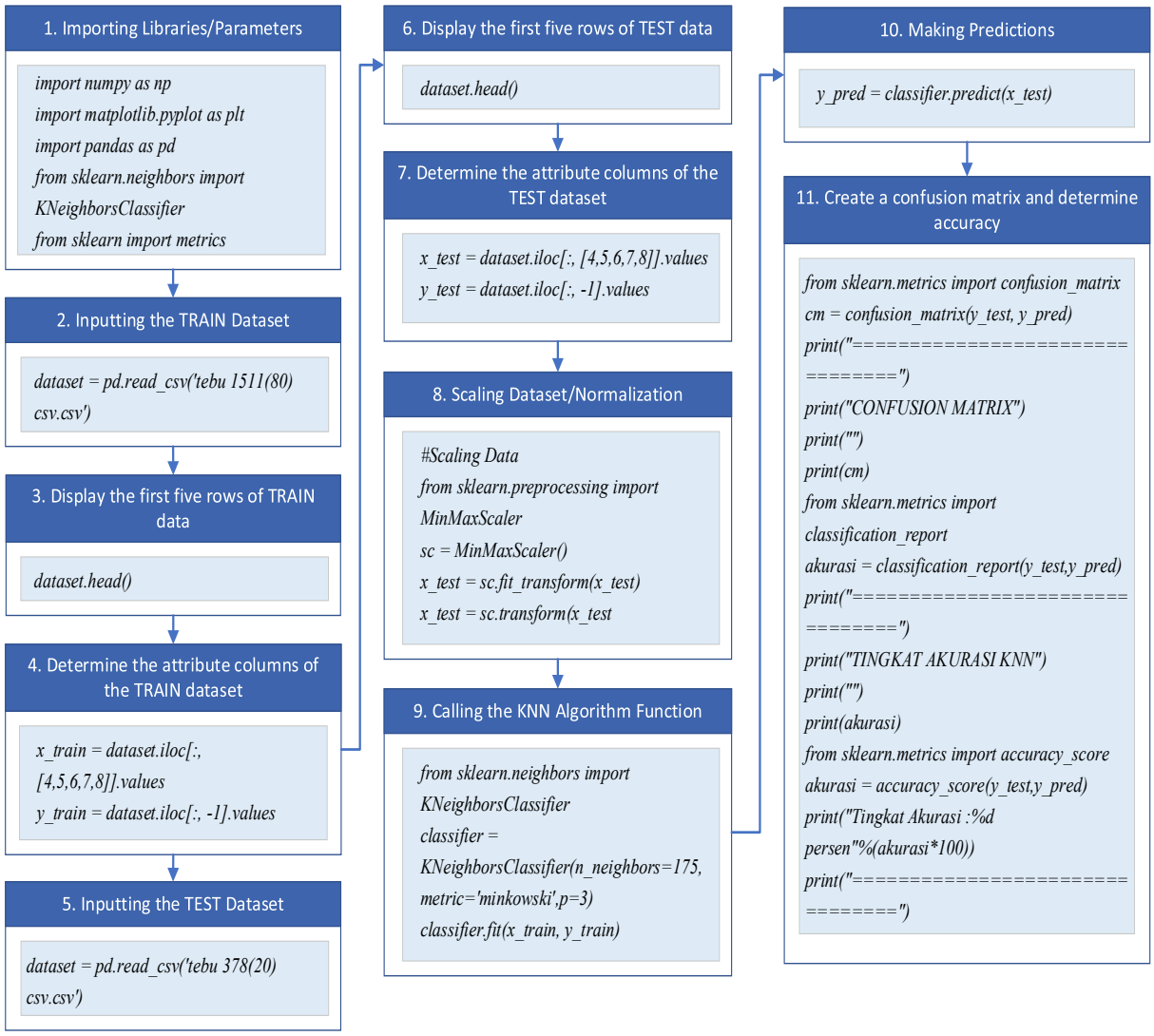

Figure 3. KNN modeling

Table 3. The first five rows in the TRAIN dataset

No	Cane No	Internode No	Code	Relative Distance	Internode Length (cm)	Mean Diameter (cm)	Circumference (cm)	Weight per cm (g/cm)	Brix (°Bx)	Class
1	1	1	100101	0.031062	17.0	3.90	11.0	9.529412	16.2	1
2	1	2	100102	0.084414	12.2	3.80	11.4	8.852459	15.8	0
3	1	3	100103	0.141787	19.2	3.80	11.1	8.437500	16.4	1
4	1	4	100104	0.201535	13.5	3.65	11.0	8.222222	16.1	1
5	1	5	100105	0.259090	18.0	3.55	10.8	8.222222	16.4	1

Table 4. The first five rows in the TEST dataset

No	Cane No	Internode No	Code	Relative Distance	Internode Length (cm)	Mean Diameter (cm)	Circumference (cm)	Weight per cm (g/cm)	Brix (°Bx)	Class
1	83	8	117108	0.522756	7.3	3.35	10.3	6.849315	16.2	1
2	83	9	117109	0.580736	11.3	3.35	10.3	7.168142	15.6	0
3	83	10	117110	0.641209	8.1	3.35	10.0	6.543210	16.2	1
4	83	11	117111	0.703865	12.0	3.20	9.8	6.083333	16.3	1
5	83	12	117112	0.766521	8.1	3.05	9.3	5.925926	17.0	1

The execution of the KNN workflow produced predicted quality classes for the testing dataset, which were evaluated using a confusion matrix. The classification results yielded TN = 105, TP = 187, FP = 75, and FN = 11, resulting in an overall accuracy of 77%. The relatively low number of false negatives indicates that most milling-feasible samples were correctly identified. In contrast, the higher number of false positives reflects a conservative classification bias that tends to favor identifying potentially feasible candidates for further validation. These results underpin the quantitative performance analysis presented in the Results section.

3. RESULTS

The results of the sequential experimental evaluation are shown in Table 5 to Table 8, which summarize the model's performance across varying train-test splits, normalization methods, distance formulas, and neighborhood sizes. Table 6 presents the results for Scenario Group 1, which aimed to determine the most stable data composition. Among the three configuration KNN-3 with an 80:20 split produced the highest accuracy of 77%, indicating that this proportion provides a balanced representation of both classes and sufficient sample variation for model generalization. Consequently, KNN-3 was selected as the baseline for Scenario Group 2.

Table 5. Scenarios and classification accuracy results with variations in train-test data split

No	Scenario	Train-test data split	Normalization Method	Distance Formula	k Value	Accuracy (%)
1	KNN-1	90:10	Min max	Minkowski	175	74
2	KNN-2	85:15	Min max	Minkowski	175	77
3	KNN-3	80:20	Min max	Minkowski	175	77

Table 6. Scenarios and classification accuracy results with variations in normalization formulas

No	Scenario	Train-test data split	Normalization Method	Distance Formula	k Value	Accuracy (%)
1	KNN-3	80:20	Min max	Minkowski	175	77
2	KNN-4	80:20	Standard	Minkowski	175	78
3	KNN-5	80:20	Robust	Minkowski	175	77

Table 7. Scenarios and classification accuracy results with variations in distance formulas

No	Scenario	Train-test data split	Normalization Method	Distance Formula	k Value	Accuracy (%)
1	KNN-4	80:20	Standard	Minkowski	175	78
2	KNN-6	80:20	Standard	Euclidean	175	77
3	KNN-7	80:20	Standard	Cityblock	175	78

Table 8. Scenarios and classification accuracy results with variations in k value

No	Scenario	Train-test data split	Normalization Method	Distance Formula	*k* Value	Accuracy (%)
1	KNN-8	80:20	Standard	Minkowski	1	69
2	KNN-9	80:20	Standard	Minkowski	2	62
3	KNN-10	80:20	Standard	Minkowski	3	71
4	KNN-11	80:20	Standard	Minkowski	5	73
5	KNN-12	80:20	Standard	Minkowski	10	76
6	KNN-13	80:20	Standard	Minkowski	15	77
7	KNN-14	80:20	Standard	Minkowski	25	78
8	KNN-15	80:20	Standard	Minkowski	50	78
9	KNN-16	80:20	Standard	Minkowski	75	78
10	KNN-17	80:20	Standard	Minkowski	100	78
11	KNN-4	80:20	Standard	Minkowski	175	78
12	KNN-18	80:20	Standard	Minkowski	300	77
13	KNN-19	80:20	Standard	Minkowski	500	77
14	KNN-20	80:20	Standard	Minkowski	750	76
15	KNN-21	80:20	Standard	Minkowski	900	74
16	KNN-22	80:20	Standard	Minkowski	1000	65

Table 6 shows the outcomes for Scenario Group 2, which evaluated different normalization methods. When the optimal split from Group 1 was fixed, Standard normalization (KNN-4) achieved the highest accuracy of 78%, outperforming Min max and Robust normalization. This suggests that centering and scaling the feature distributions improved the K-Nearest Neighbors distance calculations. Therefore, KNN-4 was used as the baseline configuration for Scenario Group 3.

Table 7 presents the evaluation of distance formulations. With the split and normalization fixed, the Minkowski metric (KNN-4) and the Cityblock metric (KNN-7) both achieved the highest accuracy of 78%, while the Euclidean distance achieved 77%. Minkowski was selected as the baseline for Scenario Group 4 because it offers greater flexibility by allowing distance computation to emphasize variations across multiple dimensions, consistent with the stepwise refinement strategy adopted in this study.

Table 8 provides the results for Scenario Group 4, which examined the influence of varying neighborhood sizes. Minimal values of k, such as 1, 2, and 3, produced unstable accuracy due to high sensitivity to noise, whereas tremendous values between 500 and 1000 led to oversmoothing and declining performance. A stability region was observed between $k = 75$ and $k = 175$, within which accuracy consistently reached 78%. The neighborhood size $k = 75$ represents the most balanced and computationally efficient setting within this region.

Together, the four scenario groups progressively refined the model configuration through systematic evaluation of data composition, feature scaling, distance formulation, and neighborhood size. The optimal settings identified from this sequential exploration were an 80:20 train-test split, Standard normalization, the Minkowski distance metric, and a neighborhood size of $k = 75$. These parameter choices served as the basis for the final K-Nearest Neighbors model evaluated in the subsequent Discussion section.

4. DISCUSSION

4.1. Interpretation of results

The experimental results demonstrate that the KNN algorithm can effectively classify sugarcane quality based solely on field-collected physical attributes. The four scenario groups, which varied in train-test proportions, normalization methods, distance metrics, and neighborhood sizes, achieved accuracies ranging from 74% to 78%. The configuration using an 80:20 data split, min max normalization, the Minkowski distance with $p = 3$, and $k = 175$ achieved an accuracy of 78% and an error rate of 22%. These values confirm that the model performs reliably despite the biological variability of field-collected samples, which typically lack the consistency found in laboratory-controlled datasets.

The confusion matrix generated under the optimal configuration shows a predominance of accurate classifications, with True Positives (TP) = 105 and True Negatives (TN) = 187, indicating that the model correctly identified most milling-feasible (Brix ≥ 16) and non-feasible (Brix < 16) samples. In contrast, False Positives (FP) = 75 and False Negatives (FN) = 11 represent relatively minor misclassifications, which can be attributed to natural variability among sugarcane stalks and overlapping °Brix values near the classification threshold. Such variability is common in agricultural datasets, where biological and environmental heterogeneity introduce noise that can influence decision boundaries [11], [13].

The stability of model performance across the different scenario groups confirms the predictive relevance of the five physical attributes. Internode length and internode weight per centimeter showed the strongest discriminative capability, consistent with previous findings linking these attributes to maturity and sucrose accumulation [1], [3], [5]. These results reinforce the idea that geometric and mass-based features can serve as practical proxies for biochemical indicators, enabling rapid non-destructive quality assessment.

The effect of neighborhood size k exhibited a clear, interpretable trend. Minimal k values (1, 2, 3, and 5) yielded unstable accuracy levels ranging from 62% to 73%, indicating high sensitivity to noise and outliers. Enormous k values between 500 and 1000 led to oversmoothing, reducing accuracy to 65%-77%, illustrating the classic bias-variance trade-off in KNN classification. A stability region was observed in which accuracy consistently reached 78 percent for k values between 25 and 175. The value $k = 75$ is a balanced, computationally efficient choice that avoids the risks of both underfitting and overfitting.

Compared with recent studies, the achieved accuracy is slightly lower than that of models trained on highly curated laboratory data, as reported by Sudipa et al. [11] and Gupta & Nahar [12], which obtained accuracies above 90%. However, the present study emphasizes field-based measurements, which are subject to variations in cane size, moisture, and environmental exposure. Therefore, achieving 78% accuracy under real-world field conditions represents a robust and meaningful performance level.

The performance stability observed across the experimental scenarios suggests that the five physical features, namely relative distance ratio, internode length, mean diameter, circumference, and weight per centimeter, are effective predictors of sucrose content and overall sugarcane quality. Among these attributes, internode length and weight per centimeter exhibited stronger discriminatory capability, consistent with previous studies highlighting their correlation with maturity and sugar accumulation [1], [3], [5]. This reinforces the idea that geometric and mass-based features can serve as reliable proxies for biochemical quality indicators, thereby enabling non-destructive classification.

Compared with recent literature, the accuracy achieved in this study is consistent with, or slightly lower than, results reported from controlled environments. For instance, Sudipa et al. [11] achieved over 90% accuracy in fruit-quality prediction using KNN, while Gupta & Nahar [12] obtained similar performance in soil-quality classification using adaptive KNN models. However, those studies relied on highly curated or laboratory-calibrated datasets. In contrast, the present study emphasizes field-collected data, characterized by variability in cane size, moisture, and environmental exposure, which reflect actual industrial scenarios. Therefore, achieving 78% accuracy under such uncontrolled circumstances is a significant outcome that demonstrates the model's robustness and adaptability.

Furthermore, the use of the Minkowski distance metric (p = 3) was found to improve classification sensitivity compared with the Euclidean and Cityblock (Manhattan) metrics. This finding aligns with the observations of Suyal and Goyal [8] and Li & Ercisli [20], who reported that adjusting distance-computation parameters, particularly through nonlinear distance weighting, can significantly improve the discriminative capability of KNN models when feature variation is subtle. The stability of model performance across k values from 25 to 175 also suggests that the dataset's structure is sufficiently dense for neighborhood-based classification, supporting the appropriateness of KNN for small- to medium-sized agricultural datasets.

Overall, these findings confirm that the proposed KNN model, despite its computational simplicity, can provide meaningful classification outcomes comparable to more complex image-based or deep learning approaches [21]–[23]. The results validate the underlying assumption that physically measurable attributes can effectively capture quality-related variations in sugarcane, thereby bridging the gap between laboratory analytics and practical, in-field applications.

4.2. Implications and contributions

The findings of this study have several important implications for both theory and practice in industrial engineering, particularly for agro-industrial systems and quality control. The successful application of the KNN algorithm for classifying sugarcane milling feasibility based solely on physical attributes demonstrates that a low-complexity, interpretable model can provide reliable predictions in a real-field setting. This result bridges the gap between laboratory-based analytical models and practical decision-making tools in agricultural production environments.

From a practical standpoint, the developed model offers a viable framework for rapid, non-destructive assessment of sugarcane quality. The use of simple morphometric measurements, including internode length, diameter, circumference, and weight per centimeter, enables farmers, field inspectors, and mill operators to identify milling-feasible cane without the need for laboratory-based °Brix testing. Although the achieved classification accuracy of 78% is considered satisfactory for a field-based predictive model, it also implies that approximately 22% of samples may be incorrectly classified. In real-world industrial applications, this level of uncertainty must be managed carefully, depending on the decision context. For instance, in a sugar mill receiving 1,000 tons of cane per day, a 22% misclassification rate could result in approximately 220 tons being misclassified into the wrong quality category. If a portion of high-quality cane is misclassified as low-quality (false negatives), farmers may experience reduced incentives despite supplying premium raw material. Conversely, if low-quality cane is classified as milling-feasible (false positives), the mill could experience lower extraction efficiency and processing losses due to immature or low-sucrose cane.

From an industrial decision-making perspective, these two types of misclassification do not carry the same operational consequences. False Negative errors primarily affect farmers, as milling-feasible cane may be rejected or undervalued, potentially leading to dissatisfaction and reduced incentives. In contrast, False-Positive errors are generally more costly for sugar mills because processing low-quality cane can reduce extraction efficiency, increase energy consumption, and adversely affect overall mill performance. Consequently, sugar mills tend to prioritize minimizing False Positives to protect operational efficiency while managing False Negatives through secondary inspection or laboratory validation. Explicitly recognizing this asymmetry underscores the proposed KNN model's role as a pre-screening tool rather than a final decision mechanism. Nevertheless, when balanced against the cost and time savings of field-based, non-destructive inspection, the trade-off remains favorable. Conventional laboratory °Brix testing for 1,000 tons of cane may require several hours and multiple trained technicians, while the proposed KNN-based method can generate predictions within minutes using simple measurements. In practical terms, a moderate reduction in precision is offset by substantial gains in speed, scalability, and labor efficiency. Therefore, while an 78% accuracy rate is not perfect, it represents an acceptable level of operational performance for pre-screening or triage purposes, primarily when used as part of a multi-stage quality control process that combines physical and laboratory validation.

From a theoretical perspective, this study contributes to the understanding of the relationship between the physical and geometric features of sugarcane stalks and sucrose concentration. It reinforces the notion that morphometric data can serve as meaningful proxies for biochemical indicators when supported by suitable machine-learning frameworks. The demonstrated performance of KNN also supports recent discussions in agricultural informatics, suggesting that model interpretability and ease of deployment can be more valuable than marginal gains in predictive accuracy [11], [13]. By emphasizing a balance between simplicity and predictive reliability, this research contributes to the broader discourse on developing accessible, domain-relevant algorithms for agricultural decision support.

In the context of industrial engineering, the study contributes to the advancement of data-driven quality control systems that integrate computational intelligence with process optimization. The proposed framework exemplifies how artificial intelligence can enhance productivity, reliability, and standardization within agro-industrial supply chains. It aligns with the principles of Total Quality Management (TQM) and Continuous Improvement, as the proposed model transforms raw measurement data into actionable insights that support scheduling, capacity planning, and process efficiency. These insights not only improve decision-making but also encourage a culture of data utilization and evidence-based management within traditional agricultural industries.

The research also presents a novel methodological contribution by positioning KNN as an effective, lightweight tool for non-destructive classification of agricultural commodities. While earlier studies on KNN

have focused on fruit and soil datasets, this study is among the first to implement KNN for sugarcane quality classification under field conditions. The findings confirm that the method maintains robust performance despite biological variability, providing a foundation for scalable applications across other crop types. This contributes to both the body of knowledge in applied machine learning and to the development of intelligent quality management systems in industrial engineering.

Beyond its technical and theoretical contributions, this study holds broader relevance to sustainable industrial development. By reducing dependence on destructive laboratory analysis and enabling efficient use of resources, the model supports the objectives of the United Nations Sustainable Development Goals (SDGs), particularly SDG 2 (Zero Hunger) through agricultural productivity improvement, SDG 9 (Industry, Innovation, and Infrastructure) through technological innovation in production systems, and SDG 12 (Responsible Consumption and Production) through resource efficiency and waste minimization. Thus, the proposed approach not only enhances operational effectiveness but also embodies the principles of sustainability and inclusivity that underpin modern industrial systems.

Addressing this problem also contributes to Indonesia's strategic goals for sugar self-sufficiency and bioethanol production under the Presidential Regulation No. 40 of 2023, consistent with broader international agendas for sustainable agriculture and renewable energy [24]–[26].

4.3. Validation and reliability of results

The validity and reliability of the results obtained in this study were examined from methodological, empirical, and industrial perspectives. From a methodological standpoint, the KNN model was tested across four groups of experimental scenarios that varied the proportions of training and test data, normalization formulas, distance metrics, and the number of neighbors (k). The model consistently achieved accuracy between 74% and 78%, indicating stable performance across configurations. This consistency indicates strong internal validity, suggesting that the applied preprocessing and normalization procedures effectively minimized potential bias arising from differences in measurement scales. The stepwise evaluation design, in which each optimal configuration was carried forward to the next scenario group, also strengthened the methodological robustness by preventing overfitting and ensuring the reproducibility of results.

To further strengthen the validation of the proposed model and address potential interaction effects not fully captured by the sequential optimization strategy, an additional five-fold cross-validation was conducted. The validation employed a stratified k-fold scheme with the same preprocessing pipeline, including min max normalization and a KNN classifier with Minkowski distance ($p = 3$) and $k = 175$. The cross-validation results yielded accuracies of 77.23%, 76.16%, 77.15%, 76.82%, and 78.15% across the five folds, with an average accuracy of 77.10% and a standard deviation of 0.64%. The low variability across folds indicates consistent performance and confirms that the selected configuration generalizes well beyond a single train–test split. Notably, the average cross-validation accuracy closely aligns with the accuracy obtained from the 80:20 split and the confusion matrix analysis, thereby reinforcing the robustness and reliability of the proposed KNN-based classification framework.

Empirically, the model was validated using 1,889 field-collected samples of the Bululawang sugarcane variety grown in Sitirejo Village, Malang, East Java. These samples represent genuine biological variation in stalk morphology, maturity, and environmental exposure. Achieving 78% accuracy under these uncontrolled conditions confirms the model's empirical reliability and its ability to generalize within the studied context. The results suggest that the five morphometric parameters, namely relative distance ratio, internode length, mean diameter, circumference, and weight per centimeter, capture sufficient variability to serve as non-destructive indicators of Brix-based milling feasibility. Nevertheless, further empirical validation using datasets from other sugarcane varieties, climatic zones, or harvest seasons is recommended to verify the model's external generalizability.

From an industrial reliability perspective, the stability of model performance and its reliance on easily measurable physical attributes demonstrate the feasibility of adopting this method in actual mill operations. The model's simplicity allows for integration into existing inspection workflows without additional instrumentation or specialized labor. Although the 22% error rate indicates that manual verification or secondary testing may still be required for borderline cases, the trade-off between speed, cost, and precision remains favorable compared with conventional laboratory °Brix testing. The framework thus provides a

reliable preliminary screening tool for process scheduling, raw material allocation, and quality-based procurement decisions.

Although the sequential experimental design employed in this study enables efficient identification of a stable baseline configuration, it does not exhaustively explore all possible hyperparameter combinations and therefore does not guarantee a global optimum. Potential interaction effects between parameters such as normalization methods, distance metrics, and neighborhood size may not be fully captured. Nevertheless, the additional five-fold cross-validation analysis demonstrates that the selected configuration yields consistent and reliable performance across multiple data partitions, supporting its suitability as a practical and computationally efficient baseline for field-based applications. Future studies may further enhance methodological robustness by adopting full grid search or nested cross-validation to explore parameter interactions more comprehensively.

In summary, the results of this study are reliable, empirically grounded, and methodologically sound. The combination of sequential experimental evaluation and additional k-fold cross-validation demonstrates both computational stability and practical robustness of the proposed model. While the current validation confirms suitability as a field-oriented baseline, future research may further enhance robustness through independent testing datasets, extended grid-search strategies, and pilot implementation within operational sugar mills to assess long-term stability and industrial scalability.

5. CONCLUSION

This study developed and evaluated a KNN–based model for non-destructive classification of sugarcane milling feasibility using measurable physical attributes of the stalk. The model successfully classified sugarcane samples into two quality categories (Brix < 16 and Brix ≥ 16) with an optimal accuracy of 78% under field conditions. The results demonstrated that a lightweight and interpretable algorithm, when combined with well-defined morphometric parameters, namely relative distance ratio, internode length, mean diameter, circumference, and weight per centimeter, can serve as a practical alternative to laboratory-based sugarcane quality assessment. The consistent performance across multiple experimental scenarios confirmed the model's methodological robustness and empirical reliability.

The research contributes to the growing body of knowledge on machine learning applications in agro-industrial systems by showing that physically grounded, feature-based models can balance predictive performance with interpretability and operational feasibility. From an industrial engineering perspective, the proposed framework demonstrates how simple, data-driven tools can improve inspection efficiency, facilitate quality-based scheduling, and support decision-making in sugar mills. Furthermore, by reducing dependence on destructive laboratory testing, the model aligns with sustainable production practices and supports the objectives of the United Nations Sustainable Development Goals (SDGs), particularly SDG 2 (Zero Hunger), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 12 (Responsible Consumption and Production).

Future research should extend model validation by conducting cross-validation and testing on independent datasets from different sugarcane varieties and geographical regions to ensure broader generalizability. Incorporating temporal and environmental variables, such as temperature, moisture, and soil characteristics, may further strengthen predictive performance. Pilot testing in operational sugar mills is also recommended to verify the relationship between predicted quality classes and actual sugar yield, thereby assessing the model's practical accuracy in real production environments. Integrating the K-Nearest Neighbors framework with IoT-based field measurement tools or mobile decision-support platforms could enable real-time implementation and support wider scalability in industrial applications. In addition, the ongoing digital transformation of agro-industrial supply chains underscores the importance of integrating technological innovation with collaborative, sustainable production practices. Finally, longitudinal testing across multiple harvest seasons is needed to evaluate the long-term stability of the model under varying agronomic and climatic conditions.

ACKNOWLEDGMENT

This study was supported by the Ministry of Education, Science, and Technology through the Fundamental Research Grant and the Inter-University Collaborative Research Program 2025, under Contract No. 503/UN62.21/PG.00.01/2025. The authors also express their appreciation to Universitas Pembangunan Nasional Veteran Yogyakarta for the institutional support and research facilities provided throughout this

work. In addition, the authors express their gratitude to the editor and the anonymous reviewers for their insightful comments and constructive suggestions, which have significantly strengthened the overall quality of this manuscript.

REFERENCES

[1] P. Chauhan, M. Kaushal, D. Vaidya, A. Gupta, F. Ansari, and S. Patidar, "Storage Stability of Sugarcane Stalks," *Asian J. Dairy Food Res.*, no. Of, 2024, doi: 10.18805/ajdfr.dr-2126.

[2] S. Bhatia, Jyoti, S. K. Uppal, K. S. Thind, and S. K. Batta, "Post harvest quality deterioration in sugarcane under different environmental conditions," *Sugar Tech*, vol. 11, no. 2, 2009, doi: 10.1007/s12355-009-0023-7.

[3] F. Ergasi, A. Q. Khan, and E. O. Keyata, "Effect of storage periods on quality characteristics and sugar yield of pre-harvest burnt and unburnt cane of sugarcane varieties (Saccharum spp. hybrid) at Finchaa Sugar Factory, Oromia, Ethiopia," *Cogent Food Agric.*, vol. 9, no. 1, 2023, doi: 10.1080/23311932.2023.2258776.

[4] G. D. Urgesa and E. O. Keyata, "Effect of storage periods and conditions on juice quality characteristics of sugarcane (Saccharum officinarum sp. Hybrid) at Finchaa sugar factory, Oromia, Ethiopia," *Vegetos*, vol. 37, no. 4, 2024, doi: 10.1007/s42535-023-00678-2.

[5] V. Misra, A. K. Mall, S. Solomon, and M. I. Ansari, "Post-harvest biology and recent advances of storage technologies in sugarcane," *Biotechnology Reports*, vol. 33. 2022, doi: 10.1016/j.btre.2022.e00705.

[6] South African Sugarcane Research Institute, "Determining crop maturity for purposes of cane quality management (Information Sheet 4.7)," Mount Edgecombe, South Africa, 2022. [Online]. Available: https://sasri.org.za/wp-content/uploads/2022/08/4.7-Determining-crop-maturity-for-purposes-of-cane-quality-management.pdf.

[7] L. de P. Corrêdo, J. P. Molin, and R. Canal Filho, "Is It Possible to Measure the Quality of Sugarcane in Real-Time during Harvesting Using Onboard NIR Spectroscopy?," *AgriEngineering*, vol. 6, no. 1, 2024, doi: 10.3390/agriengineering6010005.

[8] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7. 2022, doi: 10.14445/22315381/IJETT-V70I7P205.

[9] M. Waleed, T. W. Um, T. Kamal, and S. M. Usman, "Classification of agriculture farm machinery using machine learning and internet of things," *Symmetry (Basel).*, vol. 13, no. 3, 2021, doi: 10.3390/sym13030403.

[10] M. K. Senapaty, A. Ray, and N. Padhy, "A decision support system for crop recommendation using machine learning classification algorithms," *Agriculture*, vol. 14, no. 8, p. 1256, 2024.

[11] I. G. I. Sudipa, R. A. Azdy, I. Arfiani, N. M. Setiohardjo, and others, "Leveraging k-nearest neighbors for enhanced fruit classification and quality assessment," *Indones. J. Data Sci.*, vol. 5, no. 1, pp. 30–36, 2024.

[12] A. Gupta and P. Nahar, "Classification and yield prediction in smart agriculture system using IoT," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 8, 2023, doi: 10.1007/s12652-021-03685-w.

[13] S. O. Araujo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata, "Machine learning applications in agriculture: current trends, challenges, and future perspectives," *Agronomy*, vol. 13, no. 12, p. 2976, 2023.

[14] R. Yunita, R. S. Hartati, S. Suhesti, and Syafaruddin, "Response of bululawang sugarcane variety to salt stress," in *IOP Conference Series: Earth and Environmental Science*, 2020, vol. 418, no. 1, doi: 10.1088/1755-1315/418/1/012060.

[15] S. Budi, A. E. Prihatiningrum, N. D. P. Budiono, N. E. W. Budianto, and R. Agustina, "Genetic diversity and yields of promising sugarcane clones under ratoon crop in a rainfed land," *Aust. J. Crop Sci.*, vol. 19, no. 5, pp. 539–547, 2025.

[16] N. M. Nawi, "Development of New Measurement Methods to Determine Sugarcane Quality from Stalk Samples," University of Southern Queensland, 2014.

[17] M. N. Alam, U. K. Nath, K. M. R. Karim, M. M. Ahmed, and R. Y. Mitul, "Genetic Variability of Exotic Sugarcane Genotypes," *Scientifica (Cairo).*, vol. 2017, 2017, doi: 10.1155/2017/5202913.

[18] S. M. I. Bachoosh, "Correlation and Path Coefficient Analyses in Sugarcane," *Egypt. J. Plant Breed.*, vol. 25, no. 1, pp. 15– 23, 2021, [Online]. Available: https://ejpb.journals.ekb.eg/article_220547_d748e5ea8671b993408a8ac48f4e8338.pdf.

[19] J. Sun, C. Sun, Z. Li, Y. Qian, and T. Li, "Prediction method of sugarcane important phenotype data based on multi-model and multi-task," *PLoS One*, vol. 19, no. 12, 2024, doi: 10.1371/journal.pone.0312444.

[20] Y. Li and S. Ercisli, "Data-efficient crop pest recognition based on KNN distance entropy," *Sustain. Comput. Informatics Syst.*, vol. 38, 2023, doi: 10.1016/j.suscom.2023.100860.

[21] J. Li *et al.*, "KRA: K-Nearest Neighbor Retrieval Augmented Model for Text Classification," *Electronics*, vol. 13, no. 16, p. 3237, 2024.

[22] V. Gurunathan, T. Sathiya Priya, J. Dhanasekar, M. Ishwarya Niranjana, and S. Suganya, "Plant Leaf Diseases Detection Using KNN Classifier," in *9th International Conference on Advanced Computing and Communication Systems, (ICACCS)*, 2023, pp. 2157–2162, doi: 10.1109/ICACCS57279.2023.10112901.

[23] A. Taner, M. T. Mengstu, K. Ç. Selvi, H. Duran, İ. Gür, and N. Ungureanu, "Apple Varieties Classification Using Deep Features and Machine Learning," *Agric.*, vol. 14, no. 2, 2024, doi: 10.3390/agriculture14020252.

[24] OECD/FAO, "OECD-FAO Agricultural Outlook 2020-2029," Rome/OECD Publishing, Paris, 2020. doi: https://doi.org/10.1787/1112c23b-en.

[25] International Energy Agency, "Renewables 2021: Analysis and forecast to 2026," 2021. [Online]. Available: https://www.iea.org/reports/renewables-2021.

[26] Climate Action Tracker, "Indonesia: Policies and action," 2024. [Online]. Available: https://climateactiontracker.org/countries/indonesia/policies-action/.

[27] M. V. Subbarao, J. T. S. Sindhu, Y. C. A. Padmanabha Reddy, V. Ravuri, K. P. Vasavi, and G. C. Ram, "Performance Analysis of Feature Selection Algorithms in the Classification of Dry Beans using KNN and Neural Networks," 2023, doi: 10.1109/ICSCDS56580.2023.10104809.

[28] T. K. Mishra, S. K. Mishra, K. J. Sai, B. S. Alekhya, and A. R. Nishith, "Crop Recommendation System using KNN and Random Forest considering Indian Data set," 2021, doi: 10.1109/OCIT53463.2021.00068.

[29] X. Chao and Y. Li, "Semisupervised Few-Shot Remote Sensing Image Classification Based on KNN Distance Entropy," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, 2022, doi: 10.1109/JSTARS.2022.3213749.

[30] S. Zhang, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, 2022, doi: 10.1109/TKDE.2021.3049250.