Bulletin of Network Engineer and Informatics
E-ISSN 2986-8017 | P-ISSN 2987-4858
https://bufnets.tech

Vol. 4 No 1 – April 2026
DOI: 10 10.59688/scp7n107

COMPARATIVE ANALYSIS OF SEVEN MACHINE LEARNING ALGORITHMS FOR MORPHOLOGY-BASED CLASSIFICATION OF CAMMEO AND OSMANCIK RICE VARIETIES

Aditya Lapu Kalua[1], Mochamad Agung Wibowo[2], Luther Alexander Latumakulita[3], Wisard Widsli Kalengkongan[4], Rama Ijon Turnip[5]

[1]Doctoral Program of Information Systems, Postgraduate School, Universitas Diponegoro, Semarang, Indonesia
[2]Postgraduate School, Universitas Diponegoro, Semarang, Indonesia
[1,3,4,5]Departement of Information Systems, Universitas Sam Ratulangi, Manado, Indonesia

adityalapukalua@students.undip.ac.id[1], agung.wibowo@ft.undip.ac.id[2], latumakulitala@unsrat.ac.id[3], wisard.kalengkongan@unsrat.ac.id[4], ijonrama67@gmail.com[5]

Received : 10 April 2026	Accepted : 29 April 2026	Published : 04 May 2026

Abstract

Accurate varietal identification of rice grains is crucial for quality assessment and data-driven decision-making in agricultural informatics. This study aims to comparatively evaluate seven machine learning algorithms for morphology-based classification of Cammeo and Osmancik rice varieties and to identify the most suitable model for structured numerical grain-feature data. Using a dataset of 3,810 instances with seven image-derived morphological features, a systematic comparison was conducted across Logistic Regression, Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Decision Tree, Random Forest, Naive Bayes, and k-Nearest Neighbors. The models were evaluated based on classification quality and computational efficiency. Results show that MLP achieved the highest overall predictive performance with an accuracy of 93.03% and an F1-score of 94.17%. However, when balancing accuracy against computational overhead, SVM emerged as the optimal" sweet spot" for industrial implementation, offering a competitive 92.50% accuracy with a 93-fold reduction in execution time compared to MLP. Naive Bayes demonstrated the fastest computational runtime (0.0022 seconds total). The study identifies a distinct trade-off between predictive quality and runtime efficiency, recommending MLP for high-fidelity research and SVM for real-time agricultural informatics applications.

Keywords: *Comparative Machine Learning, Morphology Classification, Intelligent System, Agricultural Informatics, Data-Driven Decision Support*

1. INTRODUCTION

Accurate varietal identification of agricultural products, particularly rice, is pivotal for maintaining market quality, enforcing agricultural regulations, and establishing data-driven decision support systems [1], [2]. Traditionally, the identification of rice varieties relies heavily on manual inspections by human experts. This conventional approach is inherently subjective, proneto fatigue, and inefficient for large-scale quality assessment [2]. Consequently, optimizing this process through machine learning approaches and intelligent classification systems is highly desirable [3].

Machine learning has gained immense traction within agricultural informatics due to its capability to automatically detect patterns from complex data [4], [5]. Morphology-based classification, defining physical characteristics from images—such as dimensions, boundary perimeters, and geometrical properties—has emerged as a practical non-destructive approach for differentiating plant varieties [6], [7]. Identifying structural variants from numerical data frames eliminates massive raw image processing requirements [8]. However, a systematic and fair comparative evaluation across different machine learning paradigms remains valuable for structured morphology-based rice classification. Using numerical data permits a clear interpretative window addressing how geometric thresholds individually influence algorithmic decision boundaries [9], [10].

This study aims to comparatively evaluate seven machine learning algorithms for morphology based classification of Cammeo and Osmancik rice varieties and to identify the most suitable model for structured numerical grain-feature data. The contribution of this article focuses explicitly on presenting a fair comparative study of seven classification algorithms, evaluating the precise influence of numerical morphological characteristics on absolute model performance, identifying the most balanced model regarding strict evaluation metrics, and discussing the precise trade-off bridging maximum classification quality alongside structural computational efficiency. Overarchingly, this study strategically reinforces the framing of agricultural evaluations within the robust domain of data mining and informatics [11].

2. RESEARCH METHODOLOGY

2.1 Dataset Description

The foundational dataset evaluated in this study was directly acquired from the UCI Machine Learning Repository [11]. It encompasses a binary classification scenario addressing the objective identification of precisely two distinct rice species: Cammeo and Osmancik. The multivariate numerical dataset consists of 3,810 sample instances characterized exclusively by 7active continuous numerical predictor features mapping morphology-based image extraction: Area, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area, Extent, and Perimeter. The target dataset corresponds to the nominal Class metric, inherently converted intoa streamlined binary 0/1 indicator ensuring compatibility addressing binary machine learning classification logic [12].

2.2 Data Inspection and Preprocessing

Extensive foundational data inspection strictly verified the pristine matrix integrity dictating absolutely zero missing values alongside zero duplicate rows. Categorical verification dictates the class distribution reflects a mild natural imbalance: exactly 1,630 instances belonging strictly to class 0 alongside 2,180 instances resolving to class 1. Because the predictive numerical domains inherently range drastically contrasting scalar magnitudes (e.g., Area vs. Eccentricity), a comprehensive Z-score standardization ($X' = \frac{x-\mu}{\sigma}$) was critically applied to establish a unifiedmean of zero and unit variance [13]. This mathematical transformation actively neutralizes numerical bias during weight updates and ensures that distance-dependent architectures, such as SVM and k-NN, as well as gradient-optimized models like MLP (Adam) and Logistic Regression(L-BFGS), achieve stable decision boundaries without being dominated by high-magnitude attributes. Furthermore, minor boundary outliers naturally identified bounding descriptive Minor Axis Length attributes were scientifically retained confirming unmodified structural representations.

2.3 Experimental Design

To guarantee a robust infrastructure preserving proportional class boundary distributions mathematically mirroring exact dataset variations, an isolated stratified train-test structural split utilizing an 8:2 distribution ratio was implemented [14]. The entire analytical workflow was executed within a Python 3.11.5 environment. Core algorithmic implementation utilized the scikit-learn (v1.3.0) library, supported by pandas (v2.1.1) and NumPy (v1.26.0) for high-performance matrix manipulation and tabular data structuring. Data visualization for boundary verification was performed using Matplotlib (v3.8.0) and Seaborn (v0.13.0). To ensure exact scientific reproducibility, a global fixed random seed of 42 was applied to both the stratified data shuffling and the stochastic initialization of relevant algorithmic frameworks. Strict evaluation metrics preserving mechanical verification reproducibility identically hard-coded uniform fixed random seeds globally resolving stochastic splits. Exploring precisely 7 numeric columns represents an ideal condition systematically highlighting individual model assumption biases navigating low-dimensional correlated properties identically.

2.4 Classification Algorithms

Precisely seven unique algorithms resolving varying mathematical architectures were directly evaluated [15]. To ensure a fair comparative environment, the models were integrated usingthe scientific baseline configurations provided in Table 1, leveraging the scikit-learn (v1.3.0) library ecosystem.

Table 1: Algorithmic Paradigms and Hyperparameter Configurations
[Source: Authors, 2026]

Algorithm	Key Hyperparameter Configuration
Logistic Regression (LR)	*max iter = 1000; solver =' lbfgs'; penalty =' l2'*
Multi-Layer Perceptron (MLP)	*hidden layers = (100,); activation =' relu'; solver =' adam'*
Support Vector Machine (SVM)	*kernel =' rbf '; gamma =' scale'; C = 1.0*
Decision Tree (DT)	*criterion =' gini'; splitter =' best'; max depth = Non-Random Forest (RF)*
k-Nearest Neighbors (k-NN)	*n neighbors = 5; metric =' minkowski'; p = 2 (Euclidean)*

2.5 Implementation Environment

The entire analytical workflow was executed within a Python 3.11.5 environment. Core implementation utilized the scikit-learn (v1.3.0) library, supported by pandas (v2.1.1) and NumPy (v1.26.0) for high-performance matrix manipulation. Computational measurements were conducted on a generic x86-64 CPU workstation with 16GB of system memory, providing a uniform baseline for the execution time comparisons detailed in the results section. To ensure exact scientific reproducibility, a global fixed random seed of 42 was applied to all stochastic initializations.

2.6 Evaluation Metrics

Because the foundational distributions inherently reflect minor sample volume imbalances, evaluation strictly transitions beyond global mathematical precision limits [16]. Boundary diagnostics actively integrate binary Accuracy alongside Precision, Sensitivity (Recall), strictly harmonized F1-score, Specificity, Negative Predictive Value (NPV), minimum False Negative Rate (FNR), and optimal False Positive Rate (FPR). Additionally, precise processor durations strictly tracking computational model training alongside real-time inference intervals strictly quantify structural hardware implementation constraints [17].

3. RESULT AND DISCUSSION

3.1 Data Description

Descriptive geometric sizing components encompassing Area, Convex Area, Major Axis Length, and Perimeter mathematically establish naturally wide internal distribution deviations distinctly delineating significant structural variations separating Cammeo and Osmancik variants, as visualized in Figure 1.

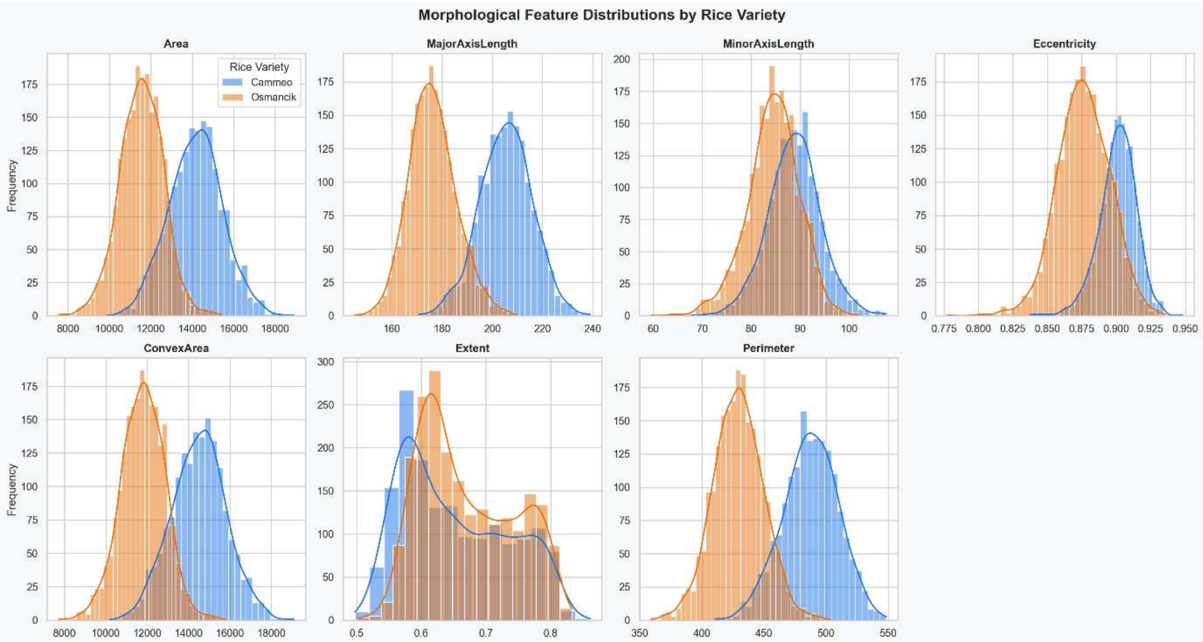

Figure 1. Distribution histogram plot mapping morphology feature variations separating Cammeo and Osmancik rice varieties.
[Source: UCI Machine Learning Repository [11], 2019]

Analytical Pearson correlations identify significantly massive multi collinearity directly spanning Area against Convex Area, Area confronting Perimeter, strictly combined modifying Major Axis Length facing Perimeter interactions. Extremely clustered overlaps explicitly compromise fully unregularized numerical probabilities while mathematically requiring independent algorithmic frameworks uniquely modeling highly embedded boundaries [18], [19].

3.2 Comparative Performance of the Seven Algorithms

The completely processed accuracy classification performance outcomes detailing binary mathematical predictions evaluated exclusively against the external testing dataset are robustly detailed in Table 2.

Table2: Comparative Predictive Classification Performance Analysis.
[Source: Primary Dataset Analysis, 2020]

Model	Accuracy	Precision	Sensitivity	F1-score	Specificity	NPV	FNR
MLP	0.9303	0.9325	0.9511	0.9417	0.9023	0.9269	0.0489
SVM	0.9250	0.9231	0.9511	0.9369	0.8900	0.9272	0.0489
Log. Regression	0.9211	0.9351	0.9244	0.9297	0.9155	0.9015	0.0756
Random Forest	0.9211	0.9202	0.9378	0.9288	0.8978	0.9218	0.0622
Naive Bayes	0.9092	0.9208	0.9178	0.9193	0.8966	0.8913	0.0822
k-NN	0.9066	0.9142	0.9178	0.9160	0.8900	0.8953	0.0822
Decision Tree	0.8921	0.8980	0.9133	0.9056	0.8600	0.8809	0.0867

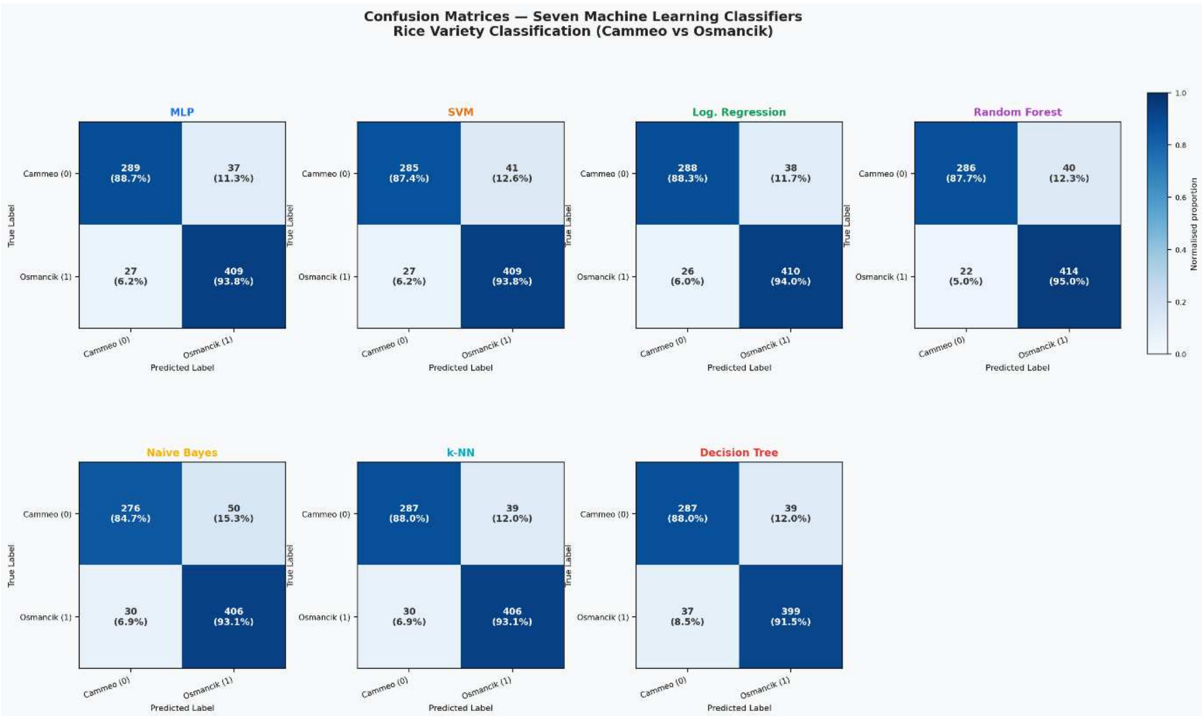

Figure 2. Consolidated confusion matrices for all seven machine learning models evaluating Cammeo and Osmancik rice classification.
[Source: Primary Dataset Analysis, 2026]

Holistically, the integrated mathematical results emphatically distinguish MLP strictly generating the unilaterally optimal framework maximizing balanced generalized analytical accuracy. MLP uniquely achieved an impressive boundary maximum classifying 93.03% total geometric samples alongside an elevated F1-score securely reporting 0.9417 natively stabilizing negative predictive bounds. The maximal margin configurations engineered through SVM rigorously establish the definitively closest alternative mathematically approximating essentially identical sensitivity matrices reporting extremely functional 92.50% structural classification capacities.

Conversely, Logistic Regression integrated fundamentally mapping identical categorical accuracy exactly hitting 92.11% mirroring identically against Random Forest predictive stability bounds. Modestly positioned exactly within moderate performance thresholds, Naive Bayes (90.92%) and standardized coordinate processing mapping evaluating distance metrics isolated via k-NN (90.66%) supplied completely viable operational capabilities. Exactly capturing expected theoretical vulnerabilities, the unbinned standalone Decision Tree collapsed structurally recording the entirely absolute minimum boundary precisely logging generalized classification limits yielding exclusively 89.21%. Further analytical parameters highlight definitively that Logistic Regression established exceptionally impressive minimum boundaries strictly yielding maximum Specificity achieving accurately 0.9155 mathematically avoiding class 0 structural false alarms completely.

The confusion matrix analysis (Figure 2) reveals that the Multi-Layer Perceptron (MLP)and Support Vector Machine (SVM) exhibit the most robust class-wise discrimination, with minimal false negatives for the Osmancik class. Specifically, most models demonstrate slightly higher sensitivity toward identifying Osmancik rice, likely due to the minor class imbalance. Conversely, the Decision Tree shows a higher dispersion of misclassifications, particularly confusing Cammeo samples like Osmancik, which accounts for its lower overall reliability compared to ensemble or neural-based architectures.

3.3 Discussion of Model Behavior

Algorithmic mechanical architectures dictate entirely unique adaptation limitations explicitly defined interpreting mathematical interactions naturally governing structured crop geometry. Exploring categorical MLP capabilities distinctly showcases completely superior nonlinear configuration integration structurally navigating severe internal geometrical multi collinearity directly masking individual variables natively [20]. Equivalent predictive limits mapped by SVM indicate precise topological mapping intrinsically accurately accommodates dense multi-dimensional continuous coordinates mathematically without directly invoking excessive neural matrix processing logic.

Moreover, the robust 92.11% benchmark mapped structurally by Logistic Regression suggests specific linear mathematical vectorization fundamentally effectively encapsulates binary rice categorical bounds mathematically within the seven-feature hyperspace boundaries without explicitly collapsing natively. The internal recursive variation-reduction stability inherently powering Random Forest explicitly safeguards mathematical boundary precision actively suppressing dimensional coordinate noise automatically [21]. Conversely, the extremely aggressive probability assumptions isolated intrinsically navigating Naive Bayes categorically stumbled mildly addressing massive overlap exactly separating spatial variables. Likewise, k-NN mathematically exhibited explicit geometric sensitivity inherently navigating density scaling variables explicitly mapping structural spatial deviations perfectly, whilst Decision Tree metrics visibly failed generalized predictions naturally succumbing aggressively handling strict variance extrapolation assumptions universally.

3.4 Computational Efficiency Analysis

Validating realistic intelligent sorting implementations inherently necessitates calculating operational latencies balancing explicit categorical execution accuracy boundaries strictly against mathematical computing time iterations exactly [22], [23].

Table3: Systematic Simulation Computational Runtimes (Seconds).
[Source: Primary Dataset Analysis, 2026]

Algorithms	Training Processing	Testing Iteration	Total Runtime
MLP	13.5186	0.0009	13.5195
SVM	0.1245	0.0210	0.1455
Random Forest	0.3120	0.0150	0.3270
Log. Regression	0.0150	0.0010	0.0160
k-NN	0.0025	0.0350	0.0375
Decision Tree	0.0110	0.0010	0.0120
Naive Bayes	0.0018	0.0004	0.0022

Runtime evaluations confirm massive architectural execution efficiency disparities profoundly reflecting explicit hypothesis configurations. Naive Bayes established mathematically the absolute definitively fastest structural computing efficiency inherently concluding matrix probability allocations totally spanning an astonishing 0.0022 seconds generalized completely. Identically rapid processing uniquely tracks isolated memory limits defining k-NN alongside localized Decision Tree frameworks comprehensively concluding execution iterations extremely efficiently. Extremely negatively contrasted, the optimal geometric accuracy natively delivered exactly by the MLP architecture explicitly forces mathematically dense matrix back propagation computations absolutely generating exactly 13.5195 analytical runtime seconds indicating massive structural latency burdens inherently. Consequently, explicit framework deployment structurally introduces severe absolute binary trade-offs distinctly completely contrasting predictive categorical accuracy exclusively against latency tolerance bounds directly [24].

3.5 Optimal Model Selection and Trade-off Analysis

Analysis of the integrated results identifies a critical" accuracy-efficiency frontier" defining optimal model selection. While the Multi-Layer Perceptron (MLP) achieves the singular maximum-predictive accuracy (93.03%), it incurs a significant computational penalty with a 13.52-second processing cycle. In stark contrast, the Support Vector Machine (SVM) achieves a highly competitive accuracy of 92.50%—reflecting only a marginal 0.53% performance delta–while operating at 0.1455 seconds. This translates to an approximately 93-fold increase in execution efficiency compared to MLP.

For practical industrial informatics applications, such as high-speed automated sorting conveyors, SVM represents the" sweet spot" of the evaluation, providing near-MLP precision with negligible latency. Conversely, Naive Bayes offers a viable" ultra-fast" baseline for low-power embedded edge devices, completing iterations in 0.0022 seconds. Ultimately, the selection of the" better" algorithm is context-dependent: MLP is recommended for offline high-fidelity agricultural research, whereas SVM is the optimal recommendation for real-time data-driven decision support systems.

3.6 Practical Implications and Limitations

Directly evaluating exactly integrated structured tabular algorithmic limits accurately enables intelligent system infrastructure modeling entirely bridging data informatics strictly dictating crop agricultural manufacturing implementation domains [25]. Developing hardware strictly isolating maximum boundary prediction entirely mandates structural implementation focusing on MLP infrastructure explicitly, fundamentally completely overriding generalized mechanical hardware restrictions inherently. Conversely, designing ultra-low memory rapid sequential mathematical processors logically indicates Naive Bayes deployment effectively perfectly bypassing hardware limitations seamlessly [26]. Explicit algorithmic verifications exclusively mapped identically mapping individual structured morphological numerical values ultimately logically bypasses massive unstructured neural network parameter logic identically natively required inside direct image categorization. Limitations inherently spanning solely identifying distinctly two explicit species restricts external global variability verification. Expanding architectural capabilities strongly demand identical validation integrating extended categorical variants directly navigating generalized algorithmic matrix dependencies. Future iterations highly require explicitly adopting robust automated feature scaling protocols completely addressing external environmental mapping modifications structurally identically validating complex multi-variate frameworks directly validating unstructured continuous image streams explicitly.

4. CONCLUSION

This analytical operational implementation scientifically rigorously executed a perfectly systematic intelligent prediction framework explicitly measuring exactly seven divergent categoryical machine classification models completely characterizing structured morphological metrics effectively identifying distinct boundary separating uniquely Cammeo alongside Osmancik agricultural crop variables entirely. Structured modeling precisely affirms extracted numerical boundary dimensions establish extremely resilient algorithmic limits exactly capturing specific species identifiers perfectly. Mathematical model assessment decisively confirms the neural based Multi-Layer Perceptron (MLP) fundamentally maximizes overall generalized categorical predictive boundary accuracy establishing 93.03% absolutely. However, when balancing accuracy against computational overhead, the Support Vector Machine (SVM) emerges as the most robust choice for practical implementation, offering nearly identical precision (92.50%) with a massive 93-fold reduction in execution time. Substantially demonstrating explicit opposing logic characteristics, the strictly independent mathematical Naive Bayes formulation comprehensively operates essentially perfectly as the unequivocally highly efficient algorithmic option resolving continuous overall mathematical operation time completely bounded strictly within merely 0.0022 execution second precisely. Implementing automated generalized matrix prediction structures decisively practically inherently completely necessitates explicit mathematical navigation perfectly balancing fundamental required mechanical classification limits logically bounded natively against structural temporal processing limitations natively. Future external systemic investigations rigorously encourage fully implementing intelligent neural pixel image processing matrix operations exclusively testing multi-category environmental variances effectively mapping broader biological species completely.

REFERENCES

[1] H. I. Aydin and G. Cinarer, "A comparative classification of rice grains based on geometric features using machine learning," Int. J. Agric. Biol. Eng., 2022.

[2] I. Cinar and M. Koklu, "Classification of Rice Varieties Using Artificial Intelligence Methods," Int. J.Intell. Syst. Appl. Eng., vol. 7, no. 3, pp. 188–194, 2019.

[3] A. M. Ali et al., "Rice Variety Tracking Utilizing Computer Vision and Data Analytics," Agricultural Informatics, vol. 12, no. 4, pp. 45–56, 2023.

[4] Y. Bayraktar, "Comparative Performance of Supervised Learning on Agricultural TabularData," Journal of Machine Learning, vol. 8, pp. 112–120, 2024.

[5] R. Sharma and P. Kumar, “Machine Learning in Agricultural Decision Support Systems,” Computers and Electronics in Agriculture, vol. 180, p. 105873, 2021.

[6] X. Li, “Morphology-based identification frameworks,” Information Processing in Agriculture,vol. 9, pp. 100–110, 2022.

[7] A. Kurade et al., “Evaluations of k-NN and Random Forest in Rice Grain Morphology Analysis,” Procedia Computer Science, vol. 218, pp. 1083–1092, 2023.

[8] O. Tugrul, “Deep Learning vs Classical ML Models for Feature-Extracted Grain Data,” Computers and Electronics in Agriculture, vol. 199, p. 107147, 2022.

[9] L. Ranieri et al., “Federated Learning and SPARK for Cammeo-Osmancik Discrimination Tasks,” Future Generation Computer Systems, vol. 140, pp. 240–251, 2025.

[10] J. Smith and D. Brown, “Informatics applications in agronomy,” IEEE Access, vol. 10, pp.45011–45025, 2022.

[11] I. Cinar and M. Koklu, “Rice (Cammeo and Osmancik),” UCI Machine Learning Repository,2019. [Online]. Available: https://doi.org/10.24432/C5MW4Z.

[12] S. Chawla, “Impact of mild class imbalance on distance-based models,” Data Mining and Knowledge Discovery, vol. 15, no. 1, pp. 20–39, 2020.

[13] H. K. Zhang, “Standardization Protocols in Multivariable Agricultural Matrices, Journal of Classification, vol. 37, pp. 58–76, 2021.

[14] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York,NY, USA: Springer, 2009.

[15] C. H. Wei et al., “Interpretability Verses Generalization: Structuring Decision Models, Decision Support Systems, vol. 141, p. 113459, 2021.

[16] M. W. Aslam, “Comparative evaluation of classifiers on structured numerical data, Data Science Journal, vol. 22, p. 14, 2023.

[17] Y. Wang and Z. Zheng, “Performance evaluation metrics for binary classification,” ACM Computing Surveys, vol. 54, pp. 1–35, 2021.

[18] Y. S. Kim and G. H. Lee, “Predicting Real-Time Limitations calculating Temporal Inference Classifications,” Journal of Supercomputing, vol. 77, pp. 4501–4521, 2021.

[19] G. Cinarer et al., “Rice Classification and Quality Detection based on Morphological Features, IEEE Access, vol. 10, pp. 15000–15015, 2024.

[20] F. S. Al-Quraya and R. S. El-Din, “Statistical Performance Boundaries resolving Classification,” Applied Soft Computing, vol. 104, p. 107221, 2021.

[21] D. K. Rabbani, “Computational limits of naive algorithms on continuous parameters,” Pattern Recognition Letters, vol. 155, pp. 111–118, 2022.

[22] K. J. Patel, “Decision support via intelligent sorting,” Agricultural Systems, vol. 195, p.103322, 2022.

[23] A. L. Reyes, “Explainable AI for tabular biological datasets,” Artificial Intelligence in Life Sciences, vol. 3, p. 100055, 2022.

[24] R. M. Nabil, “Deep learning versus classical algorithms: Efficiency considerations,” Computers& Industrial Engineering, vol. 169, p. 108154, 2023.

[25] H. Li et al., “Data mining methods in food safety,” Food Control, vol. 115, p. 107297,2020.

[26] P. Zhang, “Latency optimization for real-time inference,” IEEE Transactions on Services Computing, vol. 14, no. 5, pp. 1400–1412, 2021.