Machine Learning-Based Quality Grading of Cashew Kernels Using Linear Regression and XGBoost

Mr. Metta Satish Kumar[1], Dr. Kunjam Nageswara Rao[2]
[1] *PG Scholar, Department of Computer Science and Engineering (A) Visakhapatnam, Andhra Pradesh, India*
[2] *Professor, Department of Computer Science and Engineering (A) Visakhapatnam, Andhra Pradesh, India*

***Abstract*—Commercial cashew grading is commonly based on kernel size, shape, color, and the number of kernels per pound. Manual inspection is useful but can be slow and inconsistent when large batches are handled. This paper presents a data-driven approach for classifying whole cashew kernels into six commercial grades: W180, W210, W240, W320, W450, and W500. The study uses a labelled dataset of 6,000 kernel records with geometric, mass, and color attributes. Linear regression is used as a simple baseline, while an Extreme Gradient Boosting (XGBoost) classifier is trained using eleven raw and derived features. The workflow includes feature engineering, stratified train-test splitting, cross-validation, and a web interface for entering measured kernel values. In the project evaluation, the XGBoost classifier achieved 99.58% accuracy on the held-out test set and 99.40% mean cross-validation accuracy. The result suggests that supervised learning can support repeatable preliminary grading when measurements are collected consistently. The proposed interface is intended as a decision-support tool; it does not replace formal quality inspection or certification.**

***Index Terms*—cashew kernel, quality grading, XGBoost, linear regression, machine learning, classification**

I. INTRODUCTION

Cashew kernels are sold in recognized whole-kernel grades. Labels such as W180, W210, W240, W320, W450, and W500 indicate white whole kernels and are associated with size and kernel-count ranges. Larger kernels usually correspond to lower numerical counts. In routine handling, a grader must consider dimensional properties together with visible quality characteristics. This creates an opportunity for a measurement-based decision-support system that produces repeatable preliminary results. The objective of this work is to classify a single whole kernel from physical measurements. The work compares a simple linear-regression baseline with a nonlinear XGBoost classifier. The baseline helps show the value of a compact, interpretable relationship between dimensions and estimated thickness. XGBoost is then used because tree ensembles can model nonlinear interactions among dimensions, mass, ratios, volume, and color. The system is accompanied by a browser-based interface that collects values and sends them to a local prediction service. The contributions of this paper are: (i) a labelled six-class cashew-kernel dataset workflow, (ii) engineered geometric features relevant to size and shape, (iii) an XGBoost classification model evaluated with a hold-out set and stratified cross-validation, and (iv) a practical web interface for model-assisted grading.

II. METHODOLOGY

This study was carried out to develop a practical system for classifying cashew kernels into six common commercial grades: W180, W210, W240, W320, W450, and W500. The main idea was to use measurable kernel properties, such as size, shape, weight, and colour, to predict the most suitable grade. A dataset containing 6,000 cashew-kernel samples was used for this work. Each grade had 1,000 samples, which helped maintain a balanced dataset. For every kernel, the dataset included length, breadth, thickness, geometric mean diameter, sphericity, mass, colour, and the actual grade label.

S. No	Category	No of kernel for per 454g
1	W180	170 – 180

2	W210	200 – 210
3	W240	220 – 240
4	W320	300 – 320
5	W450	400 – 450
6	W500	450 – 500

Before training the model, the dataset was cleaned and prepared. Column names were standardized, and the colour values were converted into numerical form so that they could be used by the machine-learning model. In addition to the original measurements, four extra features were created: length-to-breadth ratio, length-to-thickness ratio, breadth-to-thickness ratio, and approximate kernel volume. These features were included because cashew grades depend not only on length but also on overall kernel shape and size.

The approximate volume of each kernel was calculated using the ellipsoid-volume formula:

$V = 4/3\pi(L/2)\ (B/2)\ (T/2)$

where (L), (B) and (T) represent length, breadth, and thickness, respectively.

The data was divided into training and testing sets. Eighty percent of the samples were used to train the models, while the remaining twenty percent were used to test their performance. Stratified sampling was applied so that all six grades were represented fairly in both sets.

Two models were considered in this study. Linear Regression was used as a basic comparison model because it is simple and easy to interpret. The main model was XGBoost, which was selected because it can handle complex relationships between several kernel properties. The XGBoost model used eleven features, including the original measurements and the derived shape-related values.

The performance of the models was evaluated using test accuracy, a confusion matrix, classification reports, feature importance, and five-fold cross-validation. The XGBoost model produced better results than the Linear Regression model. It achieved 99.58% accuracy on the test data and an average cross-validation accuracy of 99.40%.

Finally, a web-based grading interface was developed to make the system easier to use. The user can enter the physical measurements of a cashew kernel, such as length, breadth, thickness, mass, sphericity, and colour. The system then calculates the remaining features and predicts the appropriate cashew grade using the trained XGBoost model.

III. RELATED WORK AND BACKGROUND

Machine-learning systems are increasingly used when agricultural and food products can be represented by measured physical or visual attributes. In this setting, a classifier should be assessed not only by accuracy but also by whether its input variables can be measured reliably in operation. The present study therefore uses readily understandable features: length, breadth, thickness, mass, sphericity, color, and derived geometric terms.

XGBoost is a scalable implementation of gradient tree boosting that supports regularized tree ensembles and efficient training chen2016. It is suitable for tabular data with nonlinear decision boundaries. The preprocessing, split generation, metrics, and validation procedures were implemented with the Python machine-learning ecosystem pedregosa2011. Commercial cashew references describe W grades as white-whole grades linked to kernel-count limits and visible quality requirements aicc2024.

IV. MATERIALS AND METHODS

Subsection {Dataset The workbook used in this project contains 6,000 labelled records, with 1,000 samples for each target grade. The target label is one of W180, W210, W240, W320, W450, or W500. Each record contains length (mm), breadth (mm), thickness (mm), geometric mean diameter (GMD), sphericity, mass (g), color, and grade. Before modelling, column labels were cleaned and the grade distribution was inspected to confirm that all classes were represented.

Subsection {Feature Engineering Six raw measurements or attributes were used directly: length, breadth, thickness, GMD, sphericity, and mass. Color was transformed into an integer category. Four additional geometric features were calculated: length-to-breadth ratio, length-to-thickness ratio, breadth-to-thickness ratio, and approximate ellipsoidal volume. The volume was calculated as the approximate ellipsoidal volume was calculated from length, breadth, and thickness. where L, B, and T denote length, breadth, and thickness. These derived terms help distinguish kernels with similar length but different thickness or overall body size.

Subsection {Models and Evaluation An 80:20 stratified split was used so that every grade was represented in both training and test data. The XGBoost classifier used 400 estimators, maximum depth 7, learning rate 0.08, subsampling 0.80, column subsampling 0.80, minimum child weight 3, and a fixed random state of 42. Model performance was measured using classification accuracy, a confusion matrix, class-level reporting, and stratified five-fold cross-validation.

For comparison, the project also retained a linear-regression baseline that estimated a thickness-related score from length and breadth. This baseline is easy to explain but uses fewer inputs and cannot represent the full nonlinear relationship among kernel properties.

V. RESULTS AND DISCUSSION

The XGBoost model obtained 99.58% accuracy on the held-out test partition. Stratified cross-validation produced a mean accuracy of 99.40% with a standard deviation of 0.19%. These values indicate stable performance within the available dataset. The reported results should be interpreted as internal validation results; they do not prove performance on kernels from new farms, cameras, processing lines, or measurement instruments.

The improvement over the linear baseline is consistent with the fact that grade is not determined by one dimension alone. Mass, thickness, GMD, and ratios describe different aspects of kernel size and form. Color adds a quality-related attribute, although it should be collected under controlled illumination if it is used in a production environment.

The web application converts user-entered length, breadth, thickness, mass, sphericity, and color into the complete model feature vector. The local service computes the ratios, volume, and GMD before loading the saved classifier. This design prevents users from having to calculate derived values manually and makes the relationship between the interface and the training pipeline explicit.

VI. LIMITATIONS AND FUTURE WORK

The current evaluation uses one labelled dataset and fixed preprocessing assumptions. Future work should test the model on independently collected kernels, include samples with defects and broken pieces, and report per-class precision, recall, and calibration. A vision pipeline may also be developed to estimate dimensions and color from calibrated images; however, image-based measurements require reference objects, controlled lighting, and validation against physical instruments. Finally, the interface should record measurement uncertainty and warn users when inputs fall outside the training range.

VII. CONCLUSION

This work demonstrates a measurement-based cashew-kernel grading workflow using linear regression and XGBoost. The XGBoost model achieved higher internal validation accuracy than the simple baseline by combining raw measurements with engineered geometric features. The accompanying web interface provides a practical route for entering values and receiving a predicted W grade. The system is most appropriate as a consistent preliminary grading aid and should be validated on external data before commercial deployment.

REFERENCES

[1] S. N. Karnam, V. S. Vaddagallaiah, P. K. Rangnaik, A. Kumar, C. Kumar, and B. M. Vishwanath, “Precised Cashew Classification Using Machine Learning,” Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 17414–17421, Oct. 2024, doi: 10.48084/etasr.8052. This is highly relevant because it evaluates YOLOv5, YOLOv9, and CNN approaches for cashew classification. Article

[2] V. N. Pham, Q. H. Do Ba, D. A. Tran Le, Q. M. Nguyen, D. Do Van, and L. Nguyen, “A Low-Cost Deep-Learning-Based System for Grading Cashew Nuts,” Computers, vol. 13, no. 3, Art. no. 71, 2024, doi: 10.3390/computers13030071. Article

[3] S. N. Karnam and V. S. Vaddagallaiah, “Implementation and Assessment of New Hybrid Model for Cashew Kernel Classification,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 3, pp. 3492–3504, 2024. Article PDF

[4] Association of Food Industries, “AFI Specifications for Cashew Kernels,” online

standard/specification, accessed Sep. 1, 2026. This supports the W180–W500 grade definitions and kernel-count ranges. Specification

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.