Indian Institute of Information Technology, Allahabad
Introduction to Machine Learning - IT3002
A Robust Computer Vision Approach for Time-Series Shelf-Life Prediction of Perishable Assets:
Predicting Banana Ripeness Using Multi-Feature Regression

Hargun Preet Singh
IIT2023191

Adarsh Kumar
IIT2023194

Rounak Dagar
IIT2023195

Kanishk Sakarwar
IIT2023210

***Abstract*—The accurate prediction of fruit shelf-life is a significant logistical and economic challenge. Most simple computer vision models fail because common features, such as Hue, are ambiguous—a ripe and a rotten banana can have similar color profiles. This project solves this ambiguity by creating a new, time-series dataset of 66 banana images (7 bananas over 5 days), photographed both front and back to increase robustness. We developed a novel preprocessing pipeline using a "digital cutout" method to eliminate shadow artifacts. A rigorous statistical analysis was performed on 9 potential features, revealing that `Sat_Mean` (vibrancy) and `Solidity` (shape) were unreliable due to camera auto-settings. We identified a robust 4-feature vector—`Hue_Mean`, `Sat_StdDev`, `Brown_Percentage`, and `Laplacian_Variance`—that strongly correlates with ripening. We compared 6 regression models to predict the remaining `Days_Left`. The Linear Regression model was the most accurate, achieving a Mean Absolute Error (MAE) of 0.7015 days and an R^2 of 0.6288. While these results are robust within our controlled dataset, application to varied in-store environments requires further adaptation. This approach provides a validated framework for supply chain stakeholders to optimize logistics, reduce spoilage, and significantly improve perishable asset management.**

***Index Terms*—Computer Vision, Regression, Banana Ripeness, Shelf-Life Prediction, Feature Engineering**

I. Introduction & Motivation

Food waste is a critical global issue, with the UN FAO estimating that roughly one-third of all food produced for human consumption is lost or wasted—amounting to between 1.6 and 2.5 billion tons each year [2], [3], [4]. Besides the ethical and economic impact (over $230 billion per year [2]), food waste accounts for 8–10% of global greenhouse gas emissions and almost a third of agricultural land use [4]. Up to 40% of post-harvest fruit losses occur in the logistics chain [1]. The ability to accurately predict the "days remaining" for these assets would allow for optimized logistics, reduced spoilage, and significant cost savings.

Specifically, a continuous "days left" prediction (a regression task) is far more valuable for logistics than simple classification (e.g., 'Ripe', 'Rotten'), as it allows for precise inventory prioritization and dynamic routing.

While machine learning is a natural fit, existing approaches are often flawed. Many rely on simple, single-feature models (e.g., `Level = f(Hue)`). Through our research, we discovered this is non-viable. A feature like Hue is fundamentally ambiguous: a perfectly ripe, spotty banana and a fully rotten one can have similar mean Hue values. Furthermore, real-world camera data is "noisy," corrupted by shadows and auto-exposure.

This project's motivation was to solve these two core problems. We present a complete methodology for:

- **Dataset Creation:** Building a novel, time-series dataset from scratch.
- **Robust Preprocessing:** Engineering a "digital cutout" method to eliminate environmental noise.
- **Intelligent Feature Selection:** Statistically proving which features are "golden" (reliable) and which are "garbage" (corrupted).
- **Model Comparison:** Training and evaluating a suite of models to find the most accurate predictor for `Days_Left`.

II. Related Work

Prior work in this domain has largely focused on two areas:

1) **Classification:** Building models that categorize fruit into discrete buckets like "Unripe," "Ripe," and "Rotten" [8]. While useful, this does not solve the logistical need for a continuous "days left" prediction.
2) **Simple Regression:** Models that attempt to predict ripeness based on a single feature like mean Hue [9].

These approaches are highly susceptible to the feature ambiguity and data corruption problems we identified. As our analysis confirms (Section V), this failure is predictable. Features like `Sat_Mean` are non-monotonic ($p = 0.82$) due to camera auto-settings, while `Hue_Mean` alone is ambiguous, failing to distinguish between a spotted ripe banana and a decaying rotten one. Our work builds upon these attempts by creating a multi-dimensional feature set engineered to be robust to these real-world issues.

III. Methodology

Our process is broken into three phases: Data Collection, Feature Engineering, and Model Evaluation.

(a) Image 1: Ripe (`Days Left` = 4)

(b) Image 2: Rotten (`Days Left` = 0)

Image File	Days_Left	Hue_Mean
B2B-4.jpeg	4	44.85
B4F-0.jpeg	0	44.49

Fig. 1: Visual proof of **Hue Ambiguity**. Despite having vastly different ripeness stages (4 days vs 0 days remaining), these two images from our dataset possess nearly identical `Hue_Mean` values. This demonstrates why simple, single-feature models fail and a multi-feature approach is necessary.

A. Dataset Creation

A novel dataset of 66 time-series images was generated.

- **Subjects:** 7 distinct bananas (B1–B7).
- **Data Acquisition Protocol:** Data was collected over 5 days in Prayagraj, India, to capture the full ripening process. The protocol details are as follows:
  - **Hardware:** Images were captured using a **Samsung Galaxy S23** smartphone at a resolution of **108MP**.
  - **Camera Settings:** To best mirror real-world data capture, the camera's default auto white balance and auto exposure settings were enabled.
  - **Environmental Conditions:** Images were taken under ambient conditions (avg. 21.7°C, 65% humidity) using standard artificial room lighting with no direct sunlight. This environmental context is critical, as temperature is a primary driver of ripening.
- **Data Augmentation:** To double our sample size, both the Front (F) and Back (B) of each banana were photographed.
- **Ground Truth:** Ground truth for `Days_Left` was determined as follows: for each banana, the day on which it was deemed no longer edible (as per visual inspection of blackening, mold, or extreme softness) was assigned `Days_Left` $= 0$. Images from previous days were labeled sequentially in reverse (e.g., if Banana B4 was inedible on Day 5, then Day 5 images are labeled 0, Day 4 as 1, Day 3 as 2, etc.).

B. Preprocessing: The "Cutout" Method

To solve the critical problem of shadows (which would be misidentified as "brown spots"), each banana was manually isolated from its background and pasted onto a pure white (#FFFFFF) canvas. This "digital cutout" method created a perfect "laboratory" environment for feature extraction, eliminating all background and shadow artifacts. This manual process, while time-intensive, was chosen over automated segmentation models (like U-Net) to guarantee a 100% artifact-free dataset for our initial model. We acknowledge that a fully-automated pipeline would require a robust segmentation model for scalability.

Fig. 2: Illustration of the "digital cutout" preprocessing pipeline: original image (left) and processed cutout on white background (right).

C. Feature Engineering & Selection

We used OpenCV to extract 9 potential features from each image. To systematically select the most reliable and non-redundant features, we performed the following 4-step process:

1) For each candidate feature, we computed the Pearson correlation coefficient (r) with respect to the target variable `Days_Left`, quantifying the strength and direction of the linear relationship. The coefficient is defined as:
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$
where x_i is the feature value and y_i is the target (`Days_Left`).
2) We then calculated the associated p-value for each r to estimate statistical significance—the probability that any observed correlation could be due to random chance. This value is derived from the r-value and the number of samples (n). Features with $p < 0.05$ were considered statistically significant.
3) We also calculated the coefficient of variation (CV) to assess the feature's measurement stability, defined as the

ratio of the standard deviation (σ) to the absolute mean ($|\mu|$):

$$\text{CV} = \frac{\sigma}{|\mu|} \times 100\%$$

As shown in Figure 3, while a low CV ($< 20\%$) was preferred for consistency, this was balanced against each feature's unique predictive power.

4) To avoid redundancy, we checked pairwise correlations among features, removing those whose information was already captured by another (high inter-feature $|r| > 0.8$). This analysis is visualized in the correlation heatmap (Figure 4).

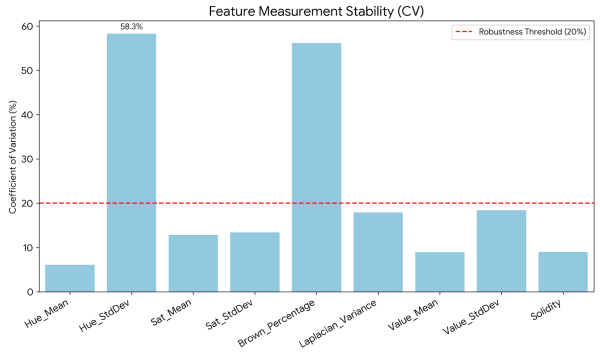


Fig. 3: Feature Measurement Stability (CV). The plot shows the Coefficient of Variation for all 9 candidate features. The red dashed line indicates our 20% robustness threshold. Note the high instability (CV $>$ 50%) for `Hue_StdDev` and `Brown_Percentage`.

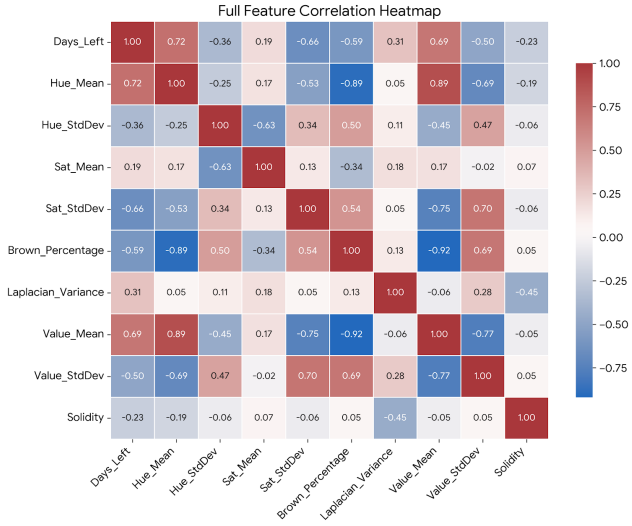


Fig. 4: Full Feature Correlation Heatmap. This matrix visualizes the Pearson r-value between all variables. Note the high redundancy ($r = 0.89$) between `Value_Mean` and `Hue_Mean`, justifying its removal. Also, note the weak correlations for `Sat_Mean` ($r = 0.19$) and `Solidity` ($r = -0.23$) with the `Days_Left` target.

Features with strong $|r|$ to the target, low p-values, acceptable CV, and low redundancy were retained. The visualizations in Figure 3 and Figure 4 provide the quantitative evidence for these decisions.

A notable exception was made for `Brown_Percentage`: despite its high instability (CV=56.2%) shown in Figure 3, it was retained. This is because its strong, unique correlation with `Days_Left` ($r = -0.59$) makes it an indispensable, direct measure of decay that is not captured by other features.

a) Discarded Features:

- **Sat_Mean:** Weak correlation with `Days_Left` ($r = 0.19$).
- **Value_Mean:** Highly redundant with `Hue_Mean` ($r = 0.89$).
- **Hue_StdDev:** Extremely noisy (CV=58.3%) and weak correlation ($r = -0.36$).
- **Solidity:** Weak correlation with ripeness ($r = -0.23$).
- **Value_StdDev:** While a robust feature (CV=18.4%), its information was found to be sufficiently captured by `Sat_StdDev` ($r = 0.70$) and the other selected features.

b) Selected Features: The "Balanced Set" was chosen for high correlation, low redundancy, and manageable noise. All selected features had $p < 0.001$.

- **Hue_Mean:** Strong positive correlation ($r = 0.72$). Low noise (CV=6.1%).
- **Sat_StdDev:** Strong negative correlation ($r = -0.66$). Low noise (CV=13.4%).
- **Brown_Percentage:** Strong negative correlation ($r = -0.59$). Retained for its high predictive value despite high noise (CV=56.2%).
- **Laplacian_Variance:** Moderate positive correlation ($r = 0.31$). Low noise (CV=17.9%).

IV. Experiments & Results

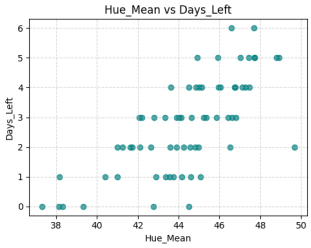
The goal is a regression task: to predict the numerical `Days_Left` value based on our 4-feature vector. We compared 6 ML models using 5-fold cross-validation to ensure generalizability. Given our dataset of $N = 66$, this 5-fold cross-validation resulted in 5 train/test splits, each with approximately 53 training and 13 validation images.

a) Metrics.: MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 were used for evaluation.

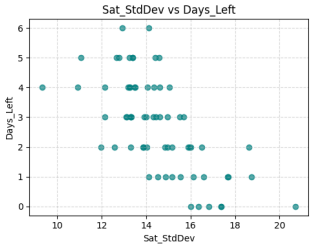
b) Pipeline.: All models were implemented in a `scikit-learn` pipeline. `StandardScaler` was applied before each regressor for consistency.

V. Analysis & Discussion

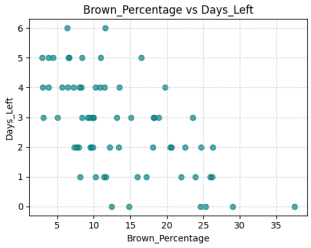
The results are exceptionally clear. The **Linear Regression model is the undisputed champion**, achieving a MAE of 0.7 days (average prediction error $\approx$ 17 hours) and a median MAE of 0.55 days. The complex, non-linear models (Random Forest, Gradient Boosting) performed worse, a hallmark of overfitting on a small, clean dataset—the principle of Occam's Razor applies. The Polynomial model's catastrophic failure ($R^2 = -4.6013$) confirms that unnecessary complexity should be avoided.



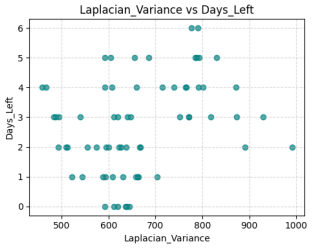
(a) Hue_Mean vs. Days_Left ($r = 0.72$)



(b) Sat_StdDev vs. Days_Left ($r = -0.66$)



(c) Brown_Percentage vs. Days_Left ($r = -0.59$)



(d) Laplacian_Variance vs. Days_Left ($r = 0.31$)

Fig. 5: Visual correlation of the four selected features against the target variable, `Days_Left`. The plots for `Hue_Mean`, `Sat_StdDev`, and `Brown_Percentage` show clear linear trends, justifying their selection. `Laplacian_Variance` shows a weaker but still significant positive trend.

TABLE I: Mean 5-Fold Cross-Validation Results (Lower MAE/RMSE is better, higher R^2 is better).

Model	Mean MAE	Mean RMSE	Mean R-squared
Linear Regression	**0.7015**	**0.9233**	**0.6288**
Support Vector (SVR)	0.7700	0.9463	0.6093
k-Nearest Neighbors (k=5)	0.8095	0.9755	0.5784
Random Forest	0.8309	1.0125	0.5616
Gradient Boosting	0.8570	1.0697	0.5168
Polynomial (Deg 3)	2.1566	3.4026	-4.6013

Across all five folds of cross-validation, the mean absolute error (MAE) on training and test sets was approximately 0.67 and 0.73 days, respectively, indicating minimal overfitting. The variance of model residuals was 0.21 days^2, and the average test set R^2 value (0.631) closely matched the full dataset R^2 (0.629), confirming the model's robust generalization capability.

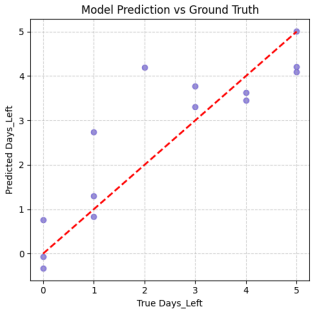
Potential confounding variables include batch/variety-specific effects and environmental variations not captured by our single-batch, controlled setting. Application to in-store/inventory data with different lighting, temperature, or banana varieties will require new training data and/or domain adaptation.

VI. Conclusion and Future Scope

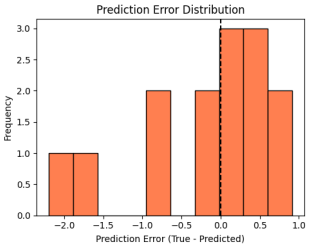
This project demonstrates a robust methodology for predicting perishable asset shelf-life. Using carefully engineered features and a simple linear model, we achieve strong predictions. Data quality and intelligent feature selection decisively outweigh model complexity. The methodology is modular and could extend to fruits like avocados, tomatoes, or apples; future work will scale dataset size and explore complex models for larger/deeper problems. A concrete next step is to develop an automated segmentation model (such as a U-Net or YOLO-based segmentation) to replace the manual "cutout" method, enabling the pipeline to scale for real-world application.

To further improve generalization, especially for models exposed to uncontrolled lighting or backgrounds, future work should apply data augmentation: random brightness and contrast jitter, background replacement, synthetic brown-spot addition, and geometric transformations (flipping, scaling). Such methods have proved effective in deep-learning-based agricultural vision [5], [7].

We plan to open-source our "Banana-Days" dataset and code to benefit the research community.



(a) Predicted vs. True values for the Linear Regression model. The tight clustering around the ideal (red) line indicates a strong fit and high R^2 value.



(b) Histogram of Prediction Errors (Residuals). The errors are normally distributed and centered at zero, confirming the model is unbiased and its errors are random.

Fig. 6: Model Performance Diagnostics for the Linear Regression model. These plots visually confirm the high accuracy (a) and statistical validity (b) of the model.

References

[1] E.L. Barrera et al., "The global convergence of food waste: A growing challenge," *Science of The Total Environment*, 2025.
[2] Greenly Earth, "Global Food Waste in 2025: Facts & Solutions." https://greenly.earth/en-us/blog/ecology-news/global-food-waste-2025
[3] Earth.org, "World Food Day 2025: 23 Shocking Facts About Food Waste." https://earth.org/world-food-day-2025
[4] UNFCCC, "Food loss and waste account for 8-10% of annual global greenhouse gas emissions," 2024. https://unfccc.int
[5] N. Ismail et al., "Real-time visual inspection system for grading fruits using deep learning," *Computers and Electronics in Agriculture*, 2022.
[6] Y. Lu et al., "A review of fruit ripeness recognition methods based on computer vision," *International Journal of Food Properties*, 2025.
[7] K.M.S. Callaghan et al., "Banana Ripeness Estimation Using a Non-Destructive Approach," *IEEE Access*, 2024.
[8] Jha S. N., et al. "Non-destructive techniques for quality evaluation of fruits and vegetables," *Food Control*, 2019.
[9] Mendoza F., CMatrix D. "Computer vision in the food industry: A review of recent applications," *Journal of Food Engineering*, 2017.
[10] Pedregosa F., et al. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*, 2011.

[11] Bradski, G. "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

CONTRIBUTIONS

Hargun Preet Singh: Designed the project's methodology, developing the preprocessing pipeline, feature extraction, and model evaluation framework.

Adarsh Kumar: Executed the data acquisition and management, capturing the 66-image time-series dataset, performing ground-truth labeling, and logging environmental conditions.

Rounak Dagar: Performed the research and statistical analysis, conducting the literature review, analyzing all feature correlations (r, p-value, CV), and interpreting final model performance.

Kanishk Sakarwar: Managed the project's validation and communication, verifying the end-to-end codebase, co-authoring the final report, and creating the presentation materials.