Classification of Rice Variety using images of different rice grains

Ayush Kumar Mourya 2022126
Bhaskar Kashyap 2022137
Daksh Yadav 2022143
Ayush Kumar 2022124

Abstract

Rice, a vital global grain, varies in appearance, taste, and nutrition across varieties, necessitating accurate classification. Traditional manual methods are labor-intensive, costly, and inconsistent. Advances in machine vision and image processing provide efficient, non-destructive solutions by analyzing grain attributes like color, texture, and size. This project employs machine learning to classify rice varieties from grain images, ensuring faster, cost-effective, and reliable results.

Github Link for Rice image Classification using ML and CNN

1. Introduction

This project classifies rice varieties using machine learning on rice grain images. Various algorithms, including Linear Regression, Logistic Regression, Naive Bayes, Decision Tree, and Random Forest, were tested with features like shape, texture, and color. Naive Bayes achieved the highest accuracy, proving effective for reliable rice variety classification.

2. Literature Survey

In this study, a total of 75,000 rice grain images were obtained, with approximately 15,000 images for each of the five rice varieties. Each image was converted into a binary format for further feature extraction. Twelve morphological features were extracted, alongside four shape features derived from these morphological attributes. Additionally, color images were converted from the RGB (red, green, blue) color space into various other color spaces, including HSV (hue, saturation, value), Lab* (lightness, red/green value, blue/yellow value), YCbCr (luminance, chroma blue, chroma red), and XYZ. From these five color spaces, a total of 90 color features were extracted. The specific names of the morphological and shape features are provided in the feature extraction section, while details about the color features are discussed in the subsequent slide.

Table 2
List of features obtained from color spaces

Color Space	Mean	Standard Deviation	Skewness	Kurtosis	Entropy	Wavelet Decomposition
	Mean_RGB_R	StdDev_RGB_R	Skewness_RGB_R	Kurtosis_RGB_R	Entropy_RGB_R	Daub4_RGB_R
RGB	Mean_RGB_G	StdDev_RGB_G	Skewness_RGB_G	Kurtosis_RGB_G	Entropy_RGB_G	Daub4_RGB_G
	Mean_RGB_B	StdDev_RGB_B	Skewness_RGB_B	Kurtosis_RGB_B	Entropy_RGB_B	Daub4_RGB_B
	Mean_HSV_H	StdDev_HSV_H	Skewness_HSV_H	Kurtosis_HSV_H	Entropy_HSV_H	Daub4_HSV_H
HSV	Mean_HSV_S	StdDev_HSV_S	Skewness_HSV_S	Kurtosis_HSV_S	Entropy_HSV_S	Daub4_HSV_S
	Mean_HSV_V	StdDev_HSV_V	Skewness_HSV_V	Kurtosis_HSV_V	Entropy_HSV_V	Daub4_HSV_V
	Mean_LAB_L	StdDev_LAB_L	Skewness_LAB_L	Kurtosis_LAB_L	Entropy_LAB_L	Daub4_LAB_L
L*a*b*	Mean_LAB_A	StdDev_LAB_A	Skewness_LAB_A	Kurtosis_LAB_A	Entropy_LAB_A	Daub4_LAB_A
	Mean_LAB_B	StdDev_LAB_B	Skewness_LAB_B	Kurtosis_LAB_B	Entropy_LAB_B	Daub4_LAB_B
	Mean_ YCbCr_Y	StdDev_ YCbCr_Y	Skewness_ YCbCr_Y	Kurtosis_ YCbCr_Y	Entropy_ YCbCr_Y	Daub4_ YCbCr_Y
YCbCr	Mean_YCbCr_Cb	StdDev_ YCbCr _Cb	Skewness_ YCbCr_Cb	Kurtosis_ YCbCr_Cb	Entropy_ YCbCr_Cb	Daub4_ YCbCr_Cb
	Mean_ YCbCr_Cr	StdDev_ YCbCr_Cr	Skewness_ YCbCr_Cr	Kurtosis_ YCbCr_Cr	Entropy_ YCbCr_Cr	Daub4_ YCbCr_Cr
	Mean_XYZ_X	StdDev_XYZ_X	Skewness_XYZ_X	Kurtosis_XYZ_X	Entropy_XYZ_X	Daub4_XYZ_X
XYZ	Mean_XYZ_Y	StdDev_XYZ_Y	Skewness_XYZ_Y	Kurtosis_XYZ_Y	Entropy_XYZ_Y	Daub4_XYZ_Y
	Mean_XYZ_Z	StdDev_XYZ_Z	Skewness_XYZ_Z	Kurtosis_XYZ_Z	Entropy_XYZ_Z	Daub4_XYZ_Z

(a) All the colour features

List of morphological and shape features

Morphological Features						Shape Features	
1	Area	5	Eccentricity	9	Extent	1	Shape_Factor_1
2	Perimeter	6	Equivalent Diameter	10	Aspect Ratio	2	Shape_Factor_2
3	Major Axis Lenght	7	Solidity	11	Roundness	3	Shape_Factor_3
4	Minor Axis Lenght	8	Convex Area	12	Compactness	4	Shape_Factor_4

(b) Shape and morphological features

Features	ANOVA	X^2	Gain Ratio	Features	ANOVA	X^2	Gain Ratio	Features	ANOVA	X^2	Gain Ratio	Features	ANOVA	X^2	Gain Ratio
Roundness	1	1	1	Entropy_ YCbCr_Cb	28	54	25	Daub4_RGB_G	55	57	50	Extent	82	92	72
Compactness	2	2	2	StdDev_HSV_V	29	15	27	Daub4_ YCbCr_Y	56	61	56	Entropy_XYZ_Y	83	93	78
Shape_Factor_3	3	3	3	StdDev_LAB_L	30	19	30	Mean_RGB_G	57	58	51	Skewness_XYZ_X	84	96	96
Aspect Ratio	4	4	4	StdDev_ YCbCr_Y	31	21	32	Mean_ YCbCr_Y	58	62	57	Skewness_ YCbCr_Cb	85	75	70
Eccentricity	5	5	5	StdDev_RGB_G	32	22	33	Daub4_XYZ_X	59	65	60	Daub4_XYZ_Z	86	84	90
Minor Axis Lenght	6	7	9	StdDev_RGB_B	33	20	34	Mean_XYZ_X	60	66	61	Mean_XYZ_Z	87	86	91
Entropy_LAB_B	7	31	13	StdDev_RGB_R	34	16	31	Kurtosis_HSV_V	61	25	42	Daub4_RGB_B	88	87	92
Entropy_RGB_R	8	30	24	StdDev_LAB_B	35	29	40	StdDev_XYZ_X	62	47	63	Mean_RGB_B	89	88	93
Entropy_ YCbCr_Cr	9	24	23	Entropy_XYZ_X	36	68	38	Skewness_RGB_G	63	69	75	Entropy_XYZ_Z	90	95	94
Shape_Factor_2	10	9	7	Kurtosis_RGB_R	37	35	55	StdDev_XYZ_Y	64	49	64	Solidity	91	97	100
Daub4_HSV_H	11	71	66	StdDev_ YCbCr _Cb	38	28	37	Skewness_LAB_L	65	72	76	Shape_Factor_4	92	99	102
Mean_HSV_H	12	70	65	Daub4_RGB_R	39	51	45	Skewness_LAB_A	66	56	69	Mean_ YCbCr_Cr	93	81	84
Shape_Factor_1	13	6	8	Mean_RGB_R	40	52	47	Skewness_RGB_R	67	80	79	Daub4_ YCbCr_Cr	94	82	85
Entropy_LAB_A	14	38	26	Kurtosis_RGB_G	41	26	41	Skewness_ YCbCr_Y	68	78	80	Skewness_XYZ_Z	95	94	98
Entropy_ YCbCr_Y	15	34	21	Kurtosis_LAB_L	42	32	46	Skewness_LAB_B	69	74	71	StdDev_ YCbCr_Cr	96	90	87
Entropy_LAB_L	16	36	20	Kurtosis_ YCbCr_Y	43	33	48	Skewness_HSV_V	70	67	73	Entropy_HSV_H	97	101	77
Entropy_RGB_G	17	37	19	Mean_HSV_S	44	18	36	Kurtosis_XYZ_Z	71	50	62	Skewness_HSV_H	98	102	99
Major Axis Lenght	18	8	6	Daub4_HSV_S	45	17	35	Kurtosis_RGB_B	72	53	68	Kurtosis_HSV_S	99	89	89
Area	19	45	17	Kurtosis_XYZ_Y	46	23	39	Skewness_HSV_S	73	73	74	Kurtosis_LAB_B	100	91	97
Convex Area	20	42	16	Kurtosis_XYZ_X	47	27	43	Skewness_XYZ_Y	74	85	88	StdDev_HSV_H	101	100	95
Equivalent Diameter	21	44	18	StdDev_HSV_S	48	48	44	Daub4_HSV_V	75	76	81	Kurtosis_LAB_A	102	103	103
Perimeter	22	12	10	StdDev_LAB_A	49	43	49	Mean_HSV_V	76	77	82	Kurtosis_HSV_H	103	104	105
Daub4_LAB_B	23	14	15	Daub4_LAB_L	50	59	52	Entropy_HSV_S	77	55	67	Skewness_YCbCr_Cr	104	105	104
Mean_LAB_B	24	13	14	Mean_LAB_L	51	60	53	Entropy_HSV_V	78	83	86	Kurtosis_YCbCr_Cr	105	106	106
Daub4_ YCbCr_Cb	25	11	12	Daub4_XYZ_Y	52	63	58	Mean_LAB_A	79	39	28	Kurtosis_YCbCr_Cb	106	98	101
Mean_YCbCr_Cb	26	10	11	Mean_XYZ_Y	53	64	59	Daub4_LAB_A	80	40	29				
Entropy_RGB_B	27	41	22	StdDev_XYZ_Z	54	46	54	Skewness_RGB_B	81	79	83				

(c) Ranking of feature selection using ANOVA, Chi-square, and Gain Ratio tests

Figure 1. Overview of extracted features and feature selection methods

3. Dataset

The dataset used in this project includes images of five distinct rice varieties: Arborio, Basmati, Ipsala, Jasmine, and Karacadag. Each variety possesses unique characteristics in terms of grain size, shape, texture, and color, which aid in their differentiation. The dataset comprises a total of 75,000 images, with 15,000 images for each variety. These images capture various visual features of the rice grains, providing the necessary data for analysis and classification tasks. This large, diverse dataset enables the application of machine learning models to distinguish between the rice varieties effectively.

3.1. Data Preprocessing

To prepare the rice grain images for the training of the machine learning model, several preprocessing steps were performed:

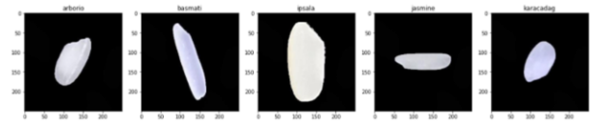
Uniform and Balanced Dataset: The dataset includes 15,000 images for each of the five rice varieties, eliminating class imbalance.

Image Resizing: All images were resized to 32x32 pixels for consistency and efficiency.

Data Augmentation (Horizontal Flipping): Random horizontal flipping introduced variability, which improved generalization.

Pixel Normalization: Pixel values were normalized to [-1, 1] for stable gradient updates.

Grayscale and Binary Conversion: Images were converted to grayscale, then binary, emphasizing structural features critical for classification.



(a) Rice grain varieties: Arborio, Basmati, Ipsala, Jasmine, and Karacadag

Label Encoding Categorical Data: The target labels (the rice varieties) are converted into numerical values using label encoding. The mapping is as follows:

Arborio: 0 Basmati: 1 Ipsala: 2 Jasmine: 3 Karacadag: 4 This step converts categorical labels into a format suitable for the model to process, creating a dataset with 75,000 entries and corresponding numerical labels.

Train and Test Dataloaders: The dataset is split into training and test sets, and the data is divided into batches of 32 images each. Random shuffling of the dataset occurs at each epoch to prevent the model from memorizing the data, thus helping it generalize better. Dataloaders not only improve memory efficiency but also allow for faster data retrieval and execution during training.

4. Methodology

Features Extracted: Morphological features like Area, Perimeter, Major and Minor Axis Lengths, Roundness, and Compactness, along with 90 color features from various color spaces (RGB, HSV, L*a*b*, YCbCr, XYZ) were extracted. Additional features like Color Histograms, HOG, LBP, and Edge Features were also considered.Then due to the poor performance of models we then extracted area , perimeter , major axis length and minor axis length .These new features had some outliers which were removed to enhance separability, reducing the dataset size from 60,000 to 50,000 entries. Machine learning models such as *Logistic Regression*, *Decision Trees*, *Random Forest*, *K-Nearest Neighbor (KNN)*, *Support Vector Machines (SVM)*, and *Multi-Layer Perceptron (MLP)* were implemented, with initial experiments revealing poor performance due to high-dimensional data and feature overlap. Re-evaluation post-outlier removal improved test accuracies (88–92%). Finally, a *Convolutional Neural Network (CNN)* achieved the best test accuracy of 96.12%.

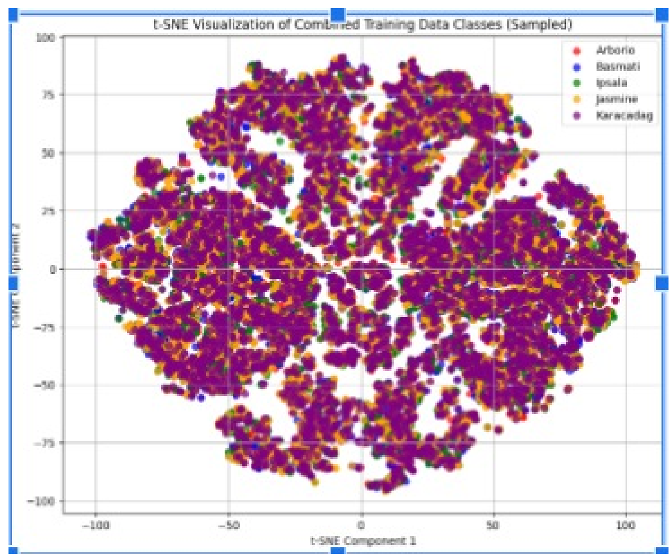


Figure 3. Features analysis based on paper

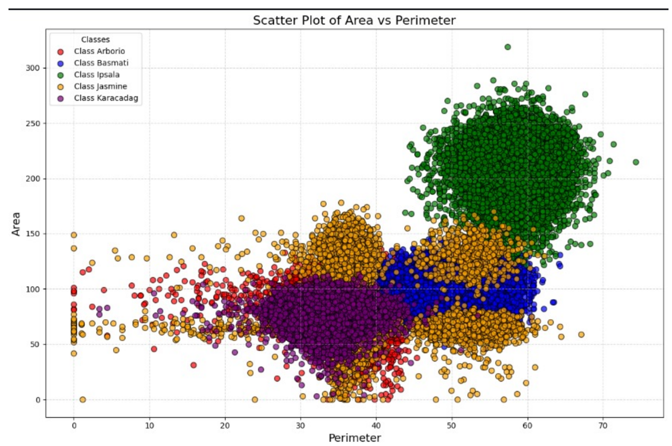


Figure 4. Feature analysis after first procedure

Model Details: Linear models like *Linear Regression* and *Logistic Regression* struggled due to linearity and overlapping clusters. Ensemble models like *Decision Trees* and *Random Forests* suffered from overfitting. *KNN* failed to handle high-dimensional clustering, and *SVM* showed marginal improvements after outlier removal. *MLP* saw limited success in addressing feature overlap. The *CNN* excelled by leveraging image-based feature extraction, achieving a superior test accuracy of 96.12%.

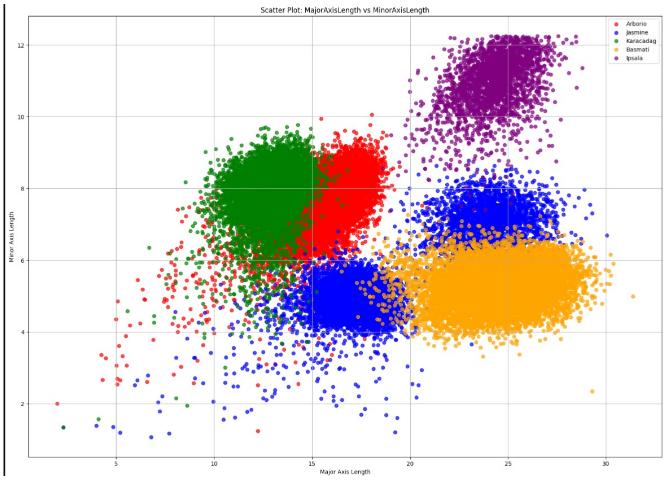


Figure 5. feature analysis after Removal of outliers

5. Results

The project *Classification of Rice Variety Using Images of Different Rice Grains* highlights the impact of feature extraction, preprocessing, and advanced modeling techniques.

Initial Results: Linear models like Logistic and Linear Regression struggled due to high dimensionality and overlapping clusters. Ensemble models like Decision Trees and Random Forest overfitted, while models like KNN, MLP, Naive Bayes, and SVM also faced challenges due to high-dimensional data, resulting in low accuracy.

Model	Metric	Train Value	Test Value
Linear Regression	Mean Squared Error	1.9536	3355377.4801
Logistic Regression	Accuracy	0.2752	0.2236
	Log Loss	1.5697	1.6434
Naive Bayes	Accuracy	0.2113	0.2595
	Log Loss	15.8304	16.9667
Decision Tree	Accuracy	1.0000	0.2033
	Log Loss	15.8304	28.7148
Random Forest	Accuracy	1.0000	0.1921
	Log Loss	0.3508	1.6433
KNN	Accuracy	0.4554	0.1868
	Log Loss	1.1191	13.4604
MLP	Accuracy	0.8774	0.1804
	Log Loss	0.3385	8.7799

Figure 6. Initial Results

Improvements After Preprocessing: Outlier removal reduced the dataset size from 60,000 to 50,000 entries, improving class separability. Feature reduction and normalization improved performance, with test accuracies ranging from **88% to 92%** and training accuracies between **94% and 100%**.

Final Model Performance: Convolutional Neural Networks (CNNs) outperformed all models with a test accuracy of **96.12%**, showcasing their ability to handle high-dimensional image data and extract meaningful features.

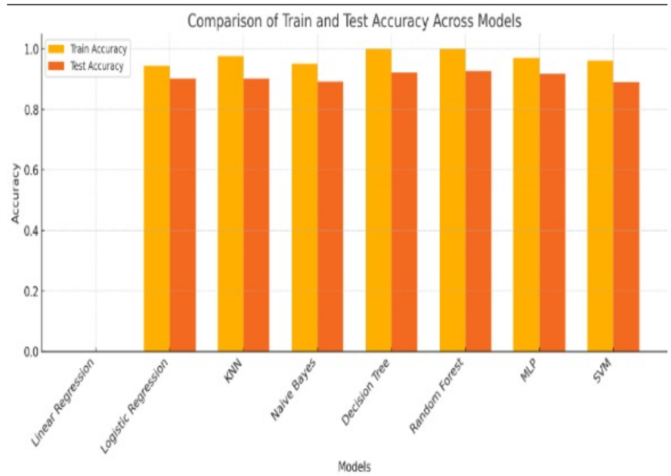


Figure 7. Performance comparison of different models after removing outliers

6. Analysis

Challenges in Classification: The dataset's high dimensionality (106 features) caused computational complexity and overlapping class groups, making linear models like Logistic Regression and Naive Bayes ineffective. Significant feature overlap among rice varieties hindered classifiers' ability to distinguish classes. Numerous outliers in features like area and axis lengths distorted class boundaries, degrading model performance. Ensemble models such as Random Forest and Decision Trees suffered from overfitting, achieving perfect training accuracy but poor generalization on testing data.

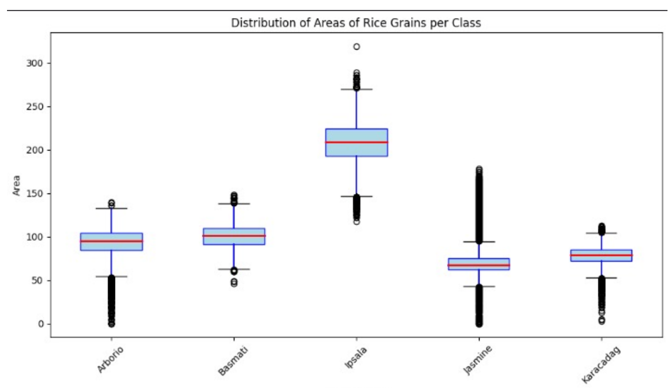


Figure 8. Removal of outliers for areas

Steps Taken to Address the Challenges: Outliers were removed, reducing the dataset from 60,000 to 50,000 en-

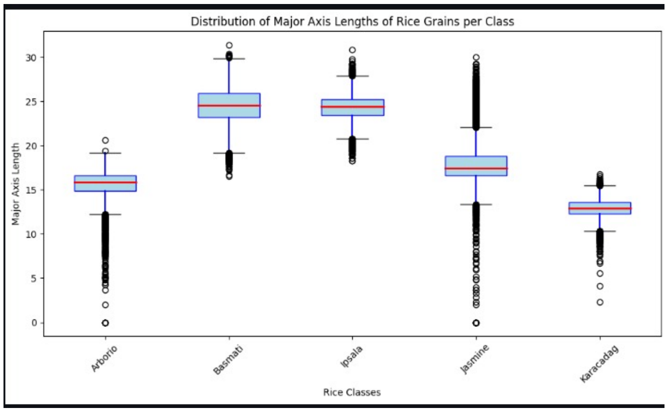


Figure 9. Removal of outliers major axis length

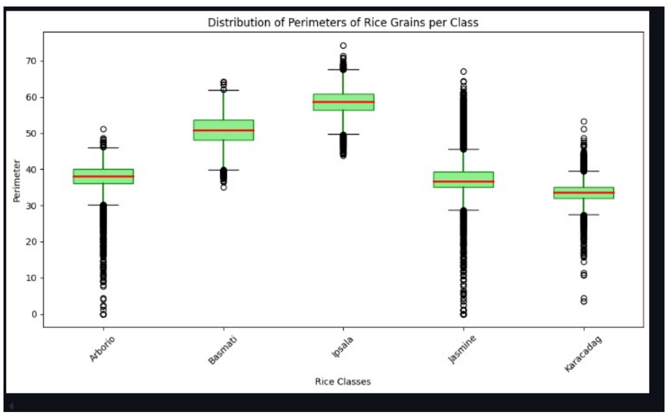


Figure 10. Removal of outliers for perimeters

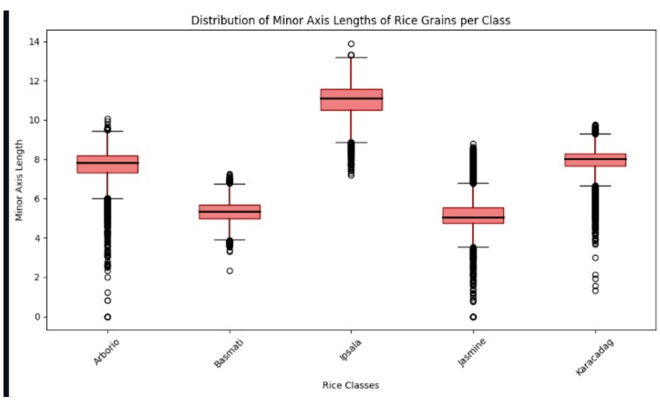


Figure 11. Removal of outliers for minor axis length

tries, which improved class separability.Additional features like Color Histograms, HOG, and LBP were introduced for better class discrimination. Preprocessing and feature refinement improved the models' robustness, resulting in better training and testing performance.

Model Performance Analysis: Linear models like Logistic Regression and Naive Bayes struggled with low accuracy due to the dataset's nonlinear nature. Ensemble models like Random Forest and Decision Trees achieved test accuracies of 88–92% post-outlier removal but required careful tuning.KNN ,MLP and SVM model showed great test accuracy.The CNN model demonstrated superior performance, achieving a test accuracy of 96.12%, effectively handling the high-dimensional image data.

7. Final Conclusion: Classification of Rice Variety Using Images of Different Rice Grains

This project focused on automating the classification of rice varieties using machine learning models and image data of different rice grains. A Convolutional Neural Network (CNN) achieved the highest accuracy of 96.12%, significantly outperforming traditional models, which achieved test accuracies in the range of 88%–92%.

Outlier removal played a crucial role in enhancing the performance of the models. Outliers were identified and removed based on the assumption that rice grains belonging to the same variety should exhibit similar shape and size features. This step was critical for improving accuracy, as it reduced noise and improved the separability of data points belonging to different classes.

The analysis revealed that simpler linear models struggled with the high-dimensional feature space of image data. Ensemble learning techniques and clustering-based models were recommended as more effective alternatives, given their ability to handle complex patterns and feature interactions.

We observed high test accuray on all the models.Training accuracies ranged between 94% and 100%, while test accuracies were lower, between 88% and 92%.

Another key observation was that increasing the number of features led to more overlapping between the clusters of different classes, making classification more challenging. By addressing these overlaps through feature selection and outlier removal, the classification performance improved significantly.

In conclusion, this project highlights the importance of data preprocessing, such as outlier removal, in achieving high accuracy for high-dimensional image data.

8. Contributions

Bhaskar Kashyap: Feature extraction of Research Paper, Model training ,CNN
Ayush Kumar Mourya: Literature survey, Clustering Analysis ,Outlier Removal , Model training and analysis after outlier removal
Daksh Yadav: Preprocessing,Linear and Logistic Regression, report making, ppt
Ayush Kumar: Report, ppt, Preprocessing , Linear and Logistic Regression

9. References

Study on Machine Learning Techniques for Rice Classification

A Comprehensive Review of Rice Image Processing Techniques

Github Link for Rice image Classification using ML and CNN