Classification of Rice Variety using images of different rice grains

Ayush Kumar Mourya

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.

ist of feat	Mean	Standard	Skewnes	, K	urtosis	Fe	tropy	Wavel	
	Mean RGB R	Deviation StdDev RGB R	Showner	RGB R K	urtosis_RGB_R		tropy_RGB_R	Decon	RGB R
GB	Mean RGB G	SidDev RGB G	Skewner	RGB G Ki	urtosis RGB C	En		Daub4	RGB C
	Mean RGB B	SidDev RGB B	Skewnes	RGB B K	urtosis RGB B urtosis HSV H	En	tropy_RGB_B	Daubi	RGB B
ISV	Mean_HSV_H Mean_HSV_S	SidDev_HSV_H SidDev_HSV_S		HSV S K	urtosis_HSV_H urtosis_HSV_S	En En	tropy_HSV_H tropy_HSV_S	Daubi	HSV B
-	Mean HSV V	SidDev HSV V	Skownes	HSV V K	urtosis HSV V	- Ea	tropy_HSV_V	Daub4	HSV V
****	Mean LAB L Mean LAB A	SidDev LAB L SidDev LAB A	Skownes	LAB L K	urtosis LAB L urtosis LAB A		tropy_LAB_L tropy_LAB_A	Daub4	LAB L
-1-0-	Mean LAB B	SidDev LAB B			urtosis LAB B	En	tropy LAB B		LAB B
	Mean YCbCr Y	SidDev YORCE Y			urtosis YCbCr		tropy_YCbCr_Y	Daubi	
								YCEC Daub4	
CbCr	Mean_YCbCr_Cl	SMDev_YCbCr_6	b Skewnes	YONG CE K	urtosis_YCbCr	Cb En	tropy_YCbCr_Cb	YCEC	r Cb
	Mean_YCbCr_C	SidDev_YCbCr_C	r Skewnes	YChCr_Cr Ki	urtosis_YCbCr	Cr En	tropy_YCbCr_Cr	Daub4 YCbC	
	Mean XYZ X	StdDev_XYZ_X	Skewnes	XYZ_X K	urtosis XYZ X	- Be	tropy_XYZ_X		XYZ X
YZ	Mean XYZ Y Mean XYZ Z	StdDev_XYZ_Y StdDev_XYZ_Z	Skewnen		urtosis XYZ Y urtosis XYZ Z	En	tropy_XYZ_Y tropy_XYZ_Z		XYZ Y
	man Alk k	JAMES TO A LEGIS	Jacobson	CALL N	HUMIT ATT. I		mpy Alle E	LABOUR	
Perimeter Major Ax	is Lenght	6 Equivalent I 7 Solidity		11 Ros	sect Ratio undness		2 Shape_Fact 3 Shape_Fact	tor_3	
Minor Ax	b) Sh	ape and	l mo	rphol		al f	eatur	es	
(b) Sh	ape and	l mo	rphol			· ompound	es	X ¹ Gain Ratio
Features Soundances	b) Sh	ape and	l mo	Festives David ROB G	ogic	al f	eature	es	% 8 92 72
Features toundatess Compactness	b) Sha	Pentiers 1 Entropy_YCNC;Cls 2 Subber_HSV_V	d mo	Festares Dash4_RGB_G	ogic	al f	Feature	es	72 72 92 73 93 78
Features foundatess Compactness Bape_Factor_1	b) Sha	Pedage	d mc	Fedures Dubbl ROB, G Dubbl YCSC, Y Mem, RGB, G	ogic	al f	Feature cot cot cot cot cot cot cot cot cot cot	es	% 8 92 72 93 78 96 96
Features Soundness Compactness Hape_Factor_7 topect Ratio	b) Sha	Penhars Fenhars Emergy_YChC_Ch Safbe_JBSV_V Safbe_JBSV_LAB_L Safbe_JKChC_Y	d mc	Federes Dubbl_ROB_G Dubbl_ROB_G Dubbl_ROB_G Mean_ROB_G	ogic	al f	Feature Tentures	es	92 72 93 78 96 96 75 79
Features Compactness Daspe, Factor, J Papert Ratio Societimity	b) Sha	Penhara Fenhara Empay_YChC_Ch Midne_HAB_L Midne_YChC_Y Midne_YChC_Y Midne_YChC_Y Midne_YChC_Y	d mc	Fedures Dub4_ROB_G Dub4_YCSC_Y Mess_ROB_G Me	ogic	al f	Penture cer rop_XYZ_Y ryrcs_XYZ_X ryrcs_XYZ_X set_XYZ_X set_XYZ_X set_XYZ_X	es	% 8 92 72 93 78 96 96
Features foundates August States, Justes, Jusped Ratio Secretaristy	b) Sha	Festives I Entropy, VICE-C, Gs SulDes, 185°, V SulDes, 1AB, L SulDes, VICE, Gs SulDes, RGB, G SulDes, RGB, G	d mc	Federes Dubbl_ROB_G Dubbl_YCbC_Y Mean_ROB_G Mean_YCbC_X Mean_XYZ_X	ogic	al f	Feature Tentage Ten	es	32 0 92 72 93 78 96 96 75 70 84 90
Features Coundates Competens Hape, Factor, 3 Appert Ratio Covertinity Histor Axis Low Entropy_LAB_	b) Sha	Penters Entropy, VCC, CS Salbe, 1887, V Salbe, 1888, V	1 mc	Festives Dubbl_ROB_C Dubbl_YCG_Y Mens_HGB_C Mens_YGG_Y Dubbl_XYZ_X Kurnois_USV_V	ogic	al f	Penture cer rop_XYZ_Y ryrcs_XYZ_X ryrcs_XYZ_X set_XYZ_X set_XYZ_X set_XYZ_X	es	32 72 92 72 93 78 96 96 75 70 84 90 86 91
Features Soundances	b) Sha	Finishers I Brings, YCCC, Cs 2 SMDes, ISS, Cs 3 SMDes, LAS, L 4 SMDes, VCCC, SS 5 SMDes, RGB, G 9 SMDes, RGB, B 5 SMDes, RGB, R 4 SMDes, LAS, L	28 54 2 29 15 2 30 19 30 31 21 3 30 30 34 16 3	Pedares Dubbl.BGB_G Dubbl.BGB_G Dubbl.YGbC_Y Menn.HGB_G Dubbl.XYZ_X Menn.XYZ_X Mann.XYZ_X Subbl.XYZ_X Subbl.XYZ_X Subbl.XYZ_X	55 57 58 58 62 62 67 67 67 67 67 67 67 67 67 67 67 67 67	al f	Feature Feature OFF Property Victor	es	N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Features foundates Demperitates Bhape, Factor, J Hopper Ratio Societainly Minor Asis Lee Intropy, EAB, Intropy, State Bhape, Factor, J	b) Sha	Ape and Finites I Intropy, YCSC, Cs 2 Sulbe, ISS', V 3 Sulbe, IAB, L 4 Sulbe, IAB, B 5 Sulbe, IAB, B	28 54 2 29 15 2 30 19 3 31 21 3 32 22 3 33 20 3 34 16 3 3 25 6 68 3 37 35 5 5	Festives Dubb-J.ROB.O. Dubb-J.YOSO.Y. Mem.RU-GO.Y. Dubb-J.XYZ.X. Man.XYZ.X. Kurnois, JISY.Y. Salbox, XYZ.X. Salvenn. RBB.C. Salvenn. RBB.C.	Ogic	al f	Feature Fea	es	2x 0 92 72 93 78 96 96 75 79 84 99 86 91 87 92 88 99 95 94 97 100
Features foundates Compactness shape, Factor, J tapeet Ratio foundates fineropy, LAB , fineropy, RGB, fineropy, RGB, fineropy, RGB, Shape, Factor, South LHSV , H	b) Sha	ape and Finises I Bango, YCC, Cs 2 Silber, BSV, V 3 Silber, LAB, I 4 Silber, YCC, Y 5 Silber, BGB, G 9 Silber, BGB, G 4 Silber, LAB, B 5 Silber, BGB, B 6 Silber, YCC, Y 7 Kansia, BGB, B 6 Silber, YCC, Cs 6 Si	28 54 2 29 15 2 30 19 2 33 20 4 16 3 20 4 6 6 8 3 3 7 3 5 3 8 2 8 3 3	Federes Dash-J. Boll, G. Dash-J. V. Gor, Y. Men, J. Gor, Y. Man, XYZ, X. Karosi, JSY, Y. Salber, XYZ, X. Salver, XYZ, X. Salver, XYZ, X. Salver, XYZ, X. Salver, XYZ, X.	0gic	al f	Feeture cet syp_XYZ_Y synco_XYZ_X synco_XYZ_X synco_XYZ_Z a_XYZ_Z a_X	es vaccos 12 13 14 15 15 15 15 15 15 15 15 15 15 15 15 15	22 72 72 93 78 96 75 79 84 99 85 94 97 100 99 102
Features foundates foundates hape, Factor, J tapeet flatio focerativity flater Avic Lee factory, LAB, flatory, RGB, flatory, WCM, flator, Avic Lee flatory, MCM, flatory, MCM, flator, Avic Lee flatory, MCM, flator,	b) Sha	Ape and Festives I Entropy, YCSC_CS 2 Sellow_18N*_V 3 Sellow_18N*_V 5 Sellow_18GB_B 9 Sellow_18GB_B 3 Sellow_18GB_B 3 Sellow_18GB_B 3 Sellow_18GB_B 5 Sellow_18GB_B 5 Sellow_18GB_B 5 Sellow_18GB_B 5 Debb_18GB_B 5 Debb_18GB_B	28 54 22 39 15 22 33 30 30 34 16 6 33 37 35 5 5 5 9 61 64 39 51 64 66 39 51 64 66 39 51 65 66 39 51 66 66 51 66 66 66 51 66 66 66 51 66 66 66 51 66 66 66 51 66 66 66 66 66 66 66 66 66 66 66 66 66	Federer Dubb JIGB, G Dubb JIGB, G Dubb JIGB, G Dubb JICB, G Men, NGC, G Me	0gic	al f	Feature on on on on on on on on on o	es v.coov 12 13 14 15 15 15 15 15 15 15 15 15 15 15 15 15	72 07 17 17 17 17 17 17 17 17 17 17 17 17 17
Festiness foundation Compactation Support Ratio Socientinity Hanne Asia Lee Entropy, SCIB, Entropy, SCIB, Entropy, SCIB, Entropy, SCIB, Happ, Factor, 3 Suith J. 185V. H. Holipp, Factor, 3	b) Sha	ape and Finites I Barago, YCAC, CS 2 Silber, 1851, V 3 Silber, 1AB, L 4 Silber, YCAC, Y 5 Silber, 1661, G 5 Dabbl, 1661, R 6 Silber, 1760, CS 5 Dabbl, 1661, R 6 Meng, 1861, R	28 54 2 29 15 2 30 19 30 31 21 8 33 20 3 34 16 3 35 26 66 8 3 37 35 5 8 28 3 3 40 52 4 40 52 4	Feditors Doubl-Holl, M. G. G. G. S. Doubl-Holl, M. G. G. G. Men, R. G. R. G. M. G. G. M. G. G. G. M. G.	Ogic	al f	Feature on on on on on on on on on o	es	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Features Completes Completes Completes Completes Completes Controlly Mane Axis Lee Entropy_CAB_intery, WGB Charpy, Yack Chapp, Factor, J Sub-4_HSV_H chen_HSV_H chen_Factor, J intery_CAB_interva_Int	b) Shi	Penters I Entropy, YCSC_CS 2 Silbo_18S'_V 3 Silbo_18S'_V 4 Silbo_YCSC_S 5 Silbo_18G_S 5 Silbo_18G_S 6 Silbo_18G_S 6 Silbo_1G_S 6 Silbo	28 54 22 39 15 22 33 30 30 34 16 6 33 37 35 5 5 5 9 61 64 39 51 64 66 39 51 64 66 39 51 65 66 39 51 66 66 51 66 66 66 51 66 66 66 51 66 66 66 51 66 66 66 51 66 66 66 66 66 66 66 66 66 66 66 66 66	Profess. Date J. Bode J. Vico. y. Men. 1801. 0 Date J. Bode J. Vico. y. Men. 1801. 0 Date J. Vico. y. Men. 1801. 0 Servena J. Bol. o Servena J. O Se	Ogic	al f	Feature one one one one one one one o	**************************************	72 07 17 17 17 17 17 17 17 17 17 17 17 17 17
Features Compactures Compactur	b) Shi	Finishers 1 Battery, YCC-C, Ch 2 Salber, BNC, YCC-C, Ch 2 Salber, BNC, YCC-C, Ch 3 Salber, LAS, L 4 Salber, YCC-C, Y 5 Salber, BCB, G 6 Salber, BCB, G 8 Salber, YCC-C 5 Dabel, BCB, G 8 Salber, BCB, G 8 Karrisis, BC	No. 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Federace Foods (Note Company) Foods (Note	Ogic	al f	Feature on on on on on on on on on o	es v.coev 12 80 84 85 86 87 99 99 99 99 99 99 99 99 99 99 99 99 99	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Features Roundness Compactness Shape, Factor, J Supert Ratio Socientisisty Maner Asis Law Interopy, EAB, Enteropy, EAB, Enteropy, EAB, Enteropy, EAB, Enteropy, EAB, Enteropy, EAB, Enteropy, LAB, Enteropy, Enteropy, Enteropy, Enteropy, Enteropy, Enteropy, Enteropy, Enteropy, Enteropy, Entero	b) Shi	Finishers 1 Battery, YCC-C, Ch 2 Salber, BNC, YCC-C, Ch 2 Salber, BNC, YCC-C, Ch 3 Salber, LAS, L 4 Salber, YCC-C, Y 5 Salber, BCB, G 6 Salber, BCB, G 8 Salber, YCC-C 5 Dabel, BCB, G 8 Salber, BCB, G 8 Karrisis, BC	1 mc	Profines Dated ROB, G Dated ROB, G Dated ROB, G Dated ROB, G Dated RVZ, X Man, RVSC, Y Dated RVZ, X Salven, RVSC, Y Salven, ROB, G Serven, ROB, G S	Ogic	al f	Feature Fea	es v.0000 12 13 18 18 18 18 18 18 18 18 18 18 18 18 18	2 0 92 72 93 78 96 96 75 79 84 99 86 91 87 92 88 99 97 100 99 102 81 84 82 83 94 98 90 87
Festimes Toundates Toundat	b) Sh: 1	I Interest Tenton Tento	1 mc	Fedders Dudd-JiGI, G Dudd-JiGI, G Dudd-JiCI, G Dudd-JiCI, G Men, JVGC, Y Men, MGI, G Men, JVGC, Y Men, MGI, G Men, JVGC, Y Men, MJC, X Men, MJC, X Men, MJC, X Menn, MGI, G Men, MGI, G Me	Ogic	al f	Feature one of the property o	es vacore 12 18 18 18 18 18 18 18 18 18 18 18 18 18	2
Features foundates foundates foundates lappe, Flactor, J lupped Batio focetainity flatency, EAB, flatency, EAB, flatency, EAS, flatency, EAS, flatency, LAB,	b) Sh: 1	Integration of Temporary CVIC, CS States of T	1 mc	Protection Dead-Motil Go Dead-	Ogic	al f	Feature on on on on on on on on on o	VACCOV 12 13 15 15 15 15 15 15 15 15 15 15 15 15 15	2
Features Connectors Connectors Compectation Construction Control of the Control Contro	b) Sh: 1	Finders I Entropy, VIAC, Co. Solley, 1879. VIAC, 1879	NAME	Festeres. Dash J.Gill, G. Dash J. Vick., Y. Dash J. J. Gill, G. Dash J. Vick., Y. Dash J. Vick., Y. Man, J. Y.Z., X. Man, J. Y. Z. Saller, J. Y. Z. Saller, J. Z. S	Ogic	25 Ext. 25 Sh Ext. 25	Ceature on Systyly on Systyl	es vaccov 122 183 184 185 187 187 187 187 187 187 187 187 187 187	2
Features Soundrose Joseph Ratio Societies Shape, Placter, J Soperit Ratio Societies Shape, Placter, J Soperit Ratio Societies Shape, Placter, J Shape, Shape	b) Shi	Interport VIDE CAS STATE AND ADDRESS OF THE PROPERTY OF THE PR	Name of the control o	Federes Dobal, MCR, G. Dobal, MCR, G. Dobal, MCR, G. Dobal, MCR, G. Mon, YGGC, V. Mon, YGGC, V. Dobal, XYZ, X. Borens, LBY, V. Silverson, LBY, V. Silverson, LBY, V. Silverson, LB, G. Silverson, LB, G. Silverson, LB, G. Silverson, VAG.	Ogici 10 10 10 10 10 10 10 1	al f	Enteres est ors ors ory, NYZ, Y orson, NYZ, X orson, NYG, C	es vacces 122 183 184 185 185 185 185 185 185 185 185 185 185	2
Features Conneliscus Compactuse Compactuse Compactuse Compactuse Control of Control Co	b) Sh: 1 1 2 2 2 2 2 2 2 2	In factors: I Berry, YUSCA: I	VMON 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Festeres. Dockl. M. Gill, G. D. Dockl. M. Gill, G. Sarresi, G. S. Sarresi, G. S. Sarresi, G. S. Sarresi, J. Sarresi, J. S. Sarresi, J. Sarresi, J. S. Sarresi, J. Sarresi, J. S. Sarresi, J. Sarr	Ogic	al f	Cature ori NYZ,Y wyp,XYZ,Y wyp,XYZ,Y wyp,XYZ,X at,XYZ,Z at,XZ,Z at,X	es v.0000 12 13 13 13 13 13 13 13 13 13 13 13 13 13	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Festione Competition of the Comp	b) Shi	Interest of the state of the st	No. 00 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Federes Doubl. JUGU. G. Doubl. JUGU. G. Doubl. JUGU. G. Doubl. JUGU. G. Men. JUGU. G. Men. JUGU. G. Men. JUGU. G. Doubl. JUZU. J. Doubl. JUZU. J. Doubl. JUZU. J. Salvenes, J. G. Sal	Ogici State of the state of th	al f	Entherse ore syp_XYZ_Y syp_XYZ_Y syp_XYZ_Y syp_XYZ_X syp_XYZ_X syp_XYZ_X sh_Ror_B sh_XYZ_Z sh_Ror_B sh_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_X s	PS VV.0000 12 13 13 14 15 15 15 15 15 15 15 15 15 15 15 15 15	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Fediness Josephers Josephe	b) Sh: 1 1 2 2 2 2 2 2 2 2	In fasters: I Barry, YUCC, O. I Barry, San, O. I Barry, S	VACO 10 10 10 10 10 10 10 10 10 10 10 10 10	Festares Dard J. ROLL J. Dard J. J. ROLL J. S. Dard J. ROLL J. Dard J. NYL J. S. Barrens, J. BOLL J. S. Sarrens, J. B. J. S. Sarrens, J. S. J. Sarrens, J. Sarrens, J. S. J. Sarrens, J. Sarrens, J. S. J. Sarrens, J. Sarrens, J. Sarrens, J. Sarrens, J. Sarrens, J. Sarrens, J. Sarr	OSIC: Vivil 100 1	21 f	Pentures on on on on on on on on on o	PS VACON 12 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Festione Competition of the Comp	b) Shi	Interest States of States	No. 00 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Problem Company Compan	Ogici State of the state of th	21 f	Entherse ore syp_XYZ_Y syp_XYZ_Y syp_XYZ_Y syp_XYZ_X syp_XYZ_X syp_XYZ_X sh_Ror_B sh_XYZ_Z sh_Ror_B sh_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_Z shy syp_XYZ_X s	PS VACON 12 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

(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:

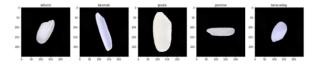
Uniform and Balanced Dataset: The dataset includes 15,000 images for each of the five rice varieties, eliminating class imbalance.

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 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%.

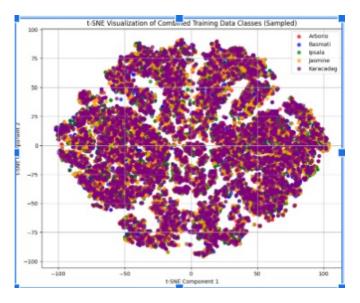


Figure 3. Features analysis based on paper

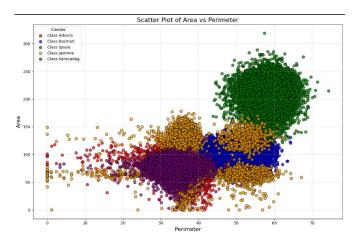


Figure 4. Feature analysis after first procedure

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

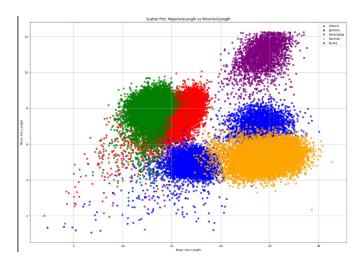


Figure 5. feature analysis after Removal of outliers

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.

		The state of the s	W// 100000	
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 Net-

works (CNNs) outperformed all models with a test accuracy of **96.12%**, showcasing their ability to handle high-dimensional image data and extract meaningful features.

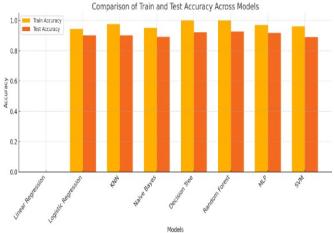


Figure 7. Performance comparison of different models after removing outliers

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.

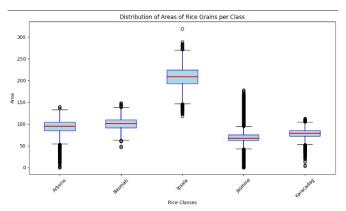


Figure 8. Removal of outliers for areas

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

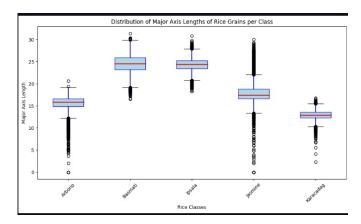


Figure 9. Removal of outliers major axis length

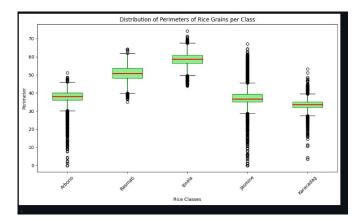


Figure 10. Removal of outliers for perimeters

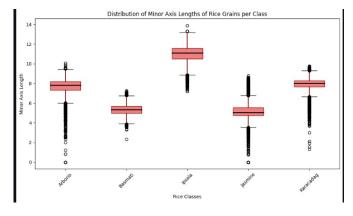


Figure 11. Removal of outliers for minor axis length

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 accu-

racy 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