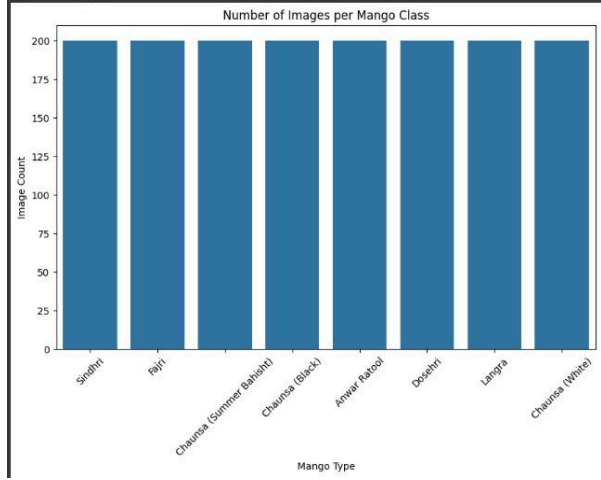


Data Collection and Preprocessing Phase

Date	4 July 2025
TeamID	SWTID1750180871
Project Title	Mangonet: A Vgg16-Based Neural Network For Mango Classification
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report:

Section	Description																		
Data Overview	Dimension: 1,600 images × 8 classes																		
Univariate Analysis	 <table border="1"> <caption>Number of Images per Mango Class</caption> <thead> <tr> <th>Mango Type</th> <th>Image Count</th> </tr> </thead> <tbody> <tr><td>Sahebi</td><td>200</td></tr> <tr><td>Fuji</td><td>200</td></tr> <tr><td>Chausa (Summer Bahiari)</td><td>200</td></tr> <tr><td>Chausa (Black)</td><td>200</td></tr> <tr><td>Anwar Ratool</td><td>200</td></tr> <tr><td>Dushehi</td><td>200</td></tr> <tr><td>Langra</td><td>200</td></tr> <tr><td>Chausa (White)</td><td>200</td></tr> </tbody> </table>	Mango Type	Image Count	Sahebi	200	Fuji	200	Chausa (Summer Bahiari)	200	Chausa (Black)	200	Anwar Ratool	200	Dushehi	200	Langra	200	Chausa (White)	200
Mango Type	Image Count																		
Sahebi	200																		
Fuji	200																		
Chausa (Summer Bahiari)	200																		
Chausa (Black)	200																		
Anwar Ratool	200																		
Dushehi	200																		
Langra	200																		
Chausa (White)	200																		
Bivariate Analysis	<i>Not applicable</i> —This is an image classification problem, and such analysis isn't directly applicable.																		
Multivariate Analysis	-																		
Outliers and Anomalies	No explicit detection for such outliers was performed.																		

Data Preprocessing Code Screenshots

Loading Data

```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("saurabhshahane/mango-varieties-classification")

print("Path to dataset files:", path)
```

Path to dataset files: /kaggle/input/mango-varieties-classification

Classification of data

```
path = "/kaggle/input/mango-varieties-classification/Dataset/Classification_dataset"
train_datagen = ImageDataGenerator(validation_split=0.2, rescale=1./255)
train_generator = train_datagen.flow_from_directory(
    directory=path,
    target_size=(224, 224),
    batch_size=32,
    class_mode='categorical',
    subset='training'
)

val_generator = train_datagen.flow_from_directory(
    directory=path,
    target_size=(224, 224),
    batch_size=32,
    class_mode='categorical',
    subset='validation'
)

test_datagen = ImageDataGenerator(rescale=1./255)
test_generator = test_datagen.flow_from_directory(
    directory=path,
    target_size=(224, 224),
    batch_size=32,
    class_mode='categorical',
    shuffle=False
)

... Found 1280 images belonging to 8 classes.
Found 320 images belonging to 8 classes.
Found 1600 images belonging to 8 classes.
```

Feature Engineering

```
from tensorflow.keras.preprocessing import image
import numpy as np

# Get the validation classes
img_path = "/kaggle/input/mango-varieties-classification/Dataset/Classification_dataset"

# Load and preprocess the image
img = image.load_img(img_path, target_size=(224, 224))
img = image.img_to_array(img)
img = img.astype('float32')
img = img / 255.0 # Normalize if your model expects normalized input

# Predict the class
predictions = model.predict(img)
predicted_class_index = np.argmax(predictions)
class_names = ['Amber Bala', 'Chausa Bala', 'Chausa Bala', 'Chausa Bala', 'Chausa Bala', 'Chausa Bala', 'Chausa Bala', 'Chausa Bala'] # Update as needed
predicted_class_name = class_names[predicted_class_index]
print("Predicted class: ", predicted_class_name)

# Save the image
img.save(img_path, target_size=(224, 224))
img = image.load_img(img_path, target_size=(224, 224))
img = image.img_to_array(img)
img = img.astype('float32')
img = img / 255.0 # Normalize if your model expects normalized input
```

Save Processed Data

-