# Train-Test Split Report

## 1. Python Code for Train-Test Split

import os  
import shutil  
from sklearn.model\_selection import train\_test\_split  
  
# ----- CONFIG -----  
# Paths  
image\_folder = './04\_Preprocessed\_Dataset'   
label\_folder = './02\_ Annotations'   
  
# Output directories  
train\_image\_out = './Train\_Set/images'  
train\_label\_out = './Train\_Set/labels'  
test\_image\_out = './Test\_Set/images'  
test\_label\_out = './Test\_Set/labels'  
  
# ----- SETUP OUTPUT FOLDERS -----  
for d in [train\_image\_out, train\_label\_out, test\_image\_out, test\_label\_out]:  
 os.makedirs(d, exist\_ok=True)  
  
image\_files = [  
 f for f in os.listdir(image\_folder)  
 if f.endswith(('.jpg', '.jpeg', '.png')) and '\_aug' not in f  
]  
  
# ----- VERIFY EXAMPLE MATCH -----  
print("✅ Total original images found:", len(image\_files))  
if image\_files:  
 base = os.path.splitext(image\_files[0])[0]  
 label\_name = base + '.txt'  
 label\_path = os.path.join(label\_folder, label\_name)  
 print(f"🔍 Example check:\n - Image: {image\_files[0]}\n - Label: {label\_name}\n - Exists: {os.path.exists(label\_path)}")  
  
# ----- SPLIT -----  
train\_imgs, test\_imgs = train\_test\_split(image\_files, test\_size=0.2, random\_state=42)  
  
def copy\_pairs(img\_list, img\_out, label\_out):  
 skipped = 0  
 for img\_file in img\_list:  
 base = os.path.splitext(img\_file)[0]  
 label\_file = base + '.txt'  
   
 src\_img = os.path.join(image\_folder, img\_file)  
 src\_label = os.path.join(label\_folder, label\_file)  
  
 if not os.path.exists(src\_label):  
 print(f"⚠️ Missing label for {img\_file}, skipping.")  
 skipped += 1  
 continue  
  
 shutil.copy(src\_img, os.path.join(img\_out, img\_file))  
 shutil.copy(src\_label, os.path.join(label\_out, label\_file))  
   
 return skipped  
  
# ----- COPY FILES -----  
print("\n📂 Copying training data...")  
train\_skipped = copy\_pairs(train\_imgs, train\_image\_out, train\_label\_out)  
  
print("\n📂 Copying testing data...")  
test\_skipped = copy\_pairs(test\_imgs, test\_image\_out, test\_label\_out)  
  
# ----- SUMMARY -----  
print("\n✅ Dataset splitting complete!")  
print(f"🧠 Training images: {len(train\_imgs) - train\_skipped}")  
print(f"🧪 Testing images: {len(test\_imgs) - test\_skipped}")  
print(f"❌ Skipped due to missing labels: {train\_skipped + test\_skipped}")

## 2. Explanation of Data Split

The dataset initially contained 436 original (non-augmented) preprocessed images in YOLO format. Using the scikit-learn `train\_test\_split` function, these images were randomly split into a training set (80%) and a testing set (20%) while ensuring no overlap. Each image is paired with a corresponding `.txt` annotation file. If a label file was missing, the image was skipped to maintain data integrity.

The final directory structure consists of:

* • Train\_Set/images/
* • Train\_Set/labels/
* • Test\_Set/images/
* • Test\_Set/labels/

## 3. Confirmation

No image was used in both the training and testing sets. The `train\_test\_split` function guarantees a clean and exclusive separation of data between the two sets.