CNN Cancer Detection Kaggle Mini-Project

December 10, 2024

[2]:
```
# Example Code Snippet for Data Overview
import pandas as pd

# Load the dataset (assuming paths are set correctly)
labels = pd.read_csv('train_labels.csv')

# Show the structure of the data
labels.head()
```

[2]:

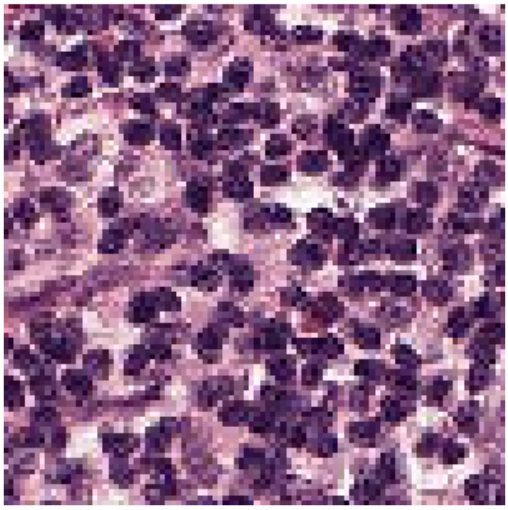
	id	label
0	f38a6374c348f90b587e046aac6079959adf3835	0
1	c18f2d887b7ae4f6742ee445113fa1aef383ed77	1
2	755db6279dae599ebb4d39a9123cce439965282d	0
3	bc3f0c64fb968ff4a8bd33af6971ecae77c75e08	0
4	068aba587a4950175d04c680d38943fd488d6a9d	0

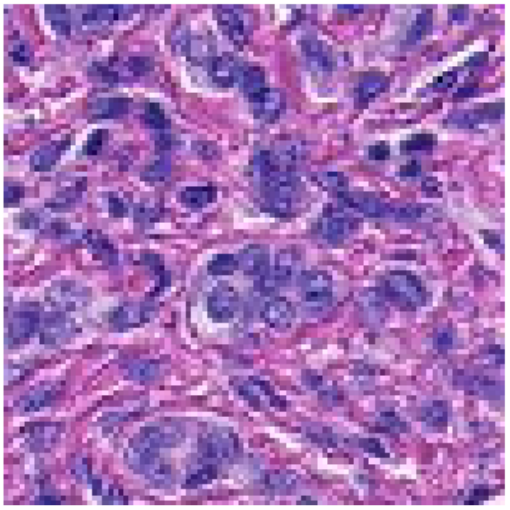
[7]:
```
# Example Code Snippet for EDA
import matplotlib.pyplot as plt
import seaborn as sns
from PIL import Image

# Load a few images to display
sample_images = ['test image 1.tif', 'test image 2.tif']  # Modify with real␣
 ↪paths

# Display sample images
fig, axes = plt.subplots(1, 2, figsize=(12, 6))
for ax, img_path in zip(axes, sample_images):
    img = Image.open(img_path)
    ax.imshow(img)
    ax.axis('off')

plt.show()
```



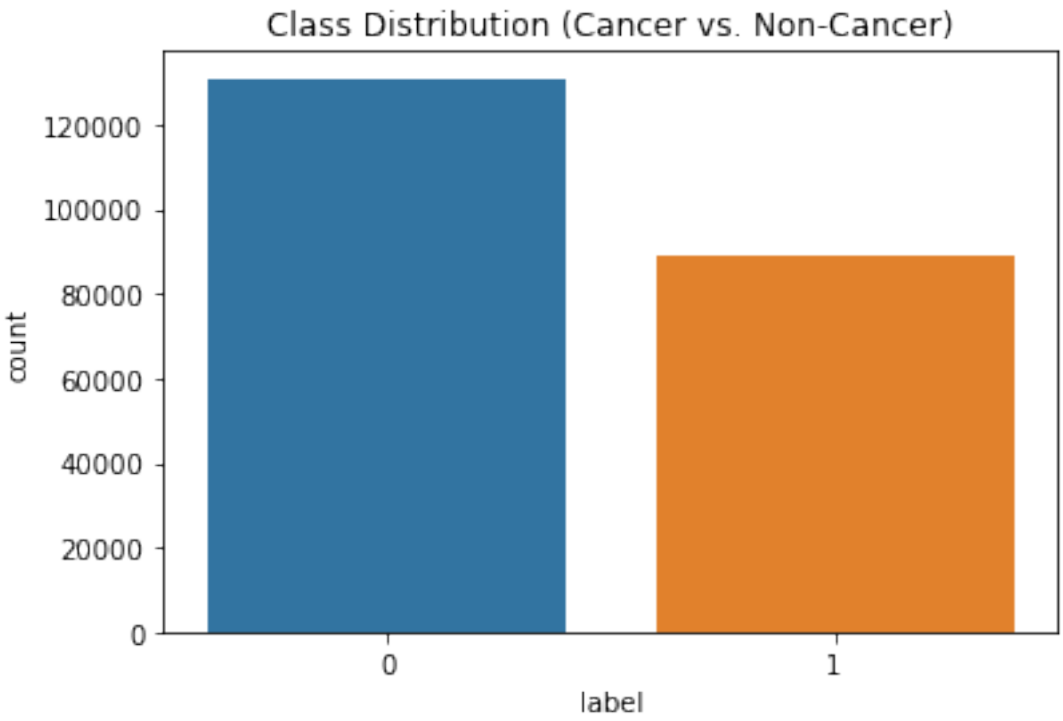


```
Index(['id', 'label'], dtype='object')
```

[8]:
```
# Plot the distribution of labels
# Check the column names
print(labels.columns)
```

```
Index(['id', 'label'], dtype='object')
```

[9]:
```
# Plot the distribution of labels (use the correct column name 'label')
sns.countplot(x='label', data=labels)
plt.title('Class Distribution (Cancer vs. Non-Cancer)')
plt.show()
```


Class Distribution (Cancer vs. Non-Cancer)
120000
100000
80000
60000
40000
20000
0
count
0
1
label