

Fatigue Fracture Surfaces

Anthony Lino

2024-10-30

Abstract

Introduction

Fatigue

Fatigue is a common failure mechanism for commercial equipment. Fatigue is a phenomenon where cracks will grow as they solid experiences cycles of high and low stress, gradually weakening the solid and eventually causing failure well below the expected strength of the material. This is shown in figure 1, where the stress experienced near the crack tip is much greater than the stress experienced throughout the rest of the material, and the ratio between these two values is called the stress concentration factor (SCF).

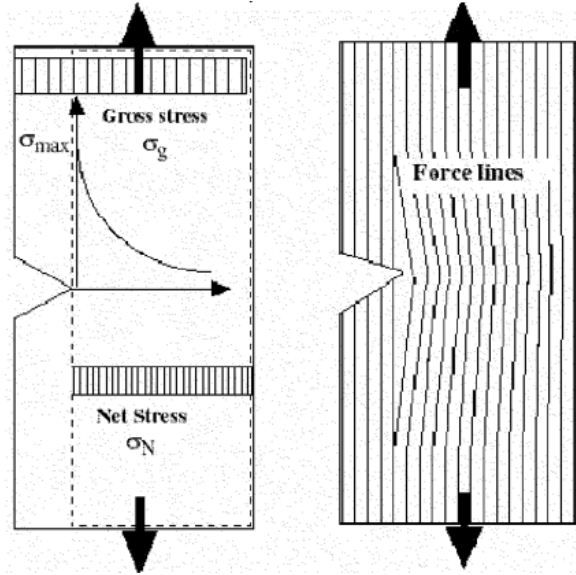


Figure 1: In the left figure the horizontal axis is the distance from the crack in the plane of the crack tip and the vertical axis is the stress as that position. We can see that as we get closer to the crack, the stress experienced at that point will go up to a maximum stress. On the right figure, this is visualized with force lines, which are meant to represent the force passing in a straight line from the bottom to the top of the material. Near the crack tip, several of the force lines are cut off, and are forced to pool near the crack tip, which causes the higher experienced stress.

As the material is stressed, the crack will grow, but eventually reach an equilibrium when the material stops stretching. This growth is also not linear, since the larger the crack the greater the concentration and the faster the growth. This feedback loop causes exponential crack growth, which shows up as linear on a log-log plot, such as figure 2. While this is easiest to explain with cracks, any irregularity in a solid, including sharp edges, surface roughness and internal defects can concentrate stress, which can cause a crack to form, and the crack can grow from there. As a result of this exponential relationship, the majority of the cycles are spent forming an initial crack from a defect, called crack nucleation, and when the crack is small. As a result, the shape, which determines the stress concentration, and the defect size, which determines the starting size of the crack, are key determiners of fatigue performance.

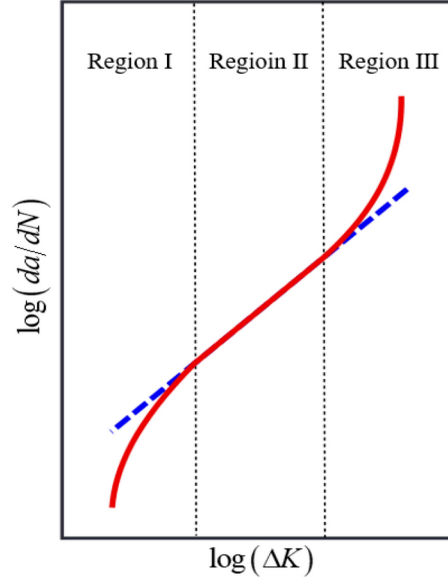


Figure 2: a is the crack length, N is the cycle and K is the stress intensity factor, which estimates the stress experienced near the tip of the crack based on its shape length and the applied stress. This means that $d(a)/dN$ is the growth in crack length per cycle and ΔK is the change in stress near the crack tip at the high and low point of the stress. This linear relationship near the center is called the Paris Regime and is a property of the material.

Laser Powder Bed Fusion

Laser Powder Bed Fusion is (LPBF) an emerging manufacturing method which allows for straightforward small batch manufacturing of complex designs. The ease of manufacturing small batches and the high design freedom allows for high performance designs. This makes it particularly well suited for adoption in the Aerospace industry. However, fatigue performance is a key metric in the Aerospace industry since cyclic loading is ubiquitous in Aerospace applications. This is problematic because LPBF inherently causes a rough surface and defects, which cause poor fatigue properties. A rough surface can be polished, but defects can be embedded below the surface, making them significantly harder to fix. The number of defects can be reduced by optimizing the settings used for printing, but that can be a very time consuming process. As a result, most studies focusing on process parameter optimization rely on these easier tests, leaving a gap in the literature for studies on fatigue optimization. A study in Professor Lewandowski's lab addressed this by tested the impact of different processing parameters on the fatigue properties. After fracture, the fractured surfaces were imaged under a microscope, revealing the impact of different defects.

Quantitative Fractography

Fractography is the study of fracture surfaces. This is a largely qualitative field, but there is an emerging field of quantitative fractography which relies on segmenting these images, and then extracting features from these masks. The Lewandowski lab explored the use of quantitative fractography for these images by performing 4 tasks:

1. Segmenting every defect from the fracture surface
2. Segmenting only the initiating defect from the fracture surface
3. Segmenting the region of fatigue crack growth
4. Segmenting the region of overload

While significant efforts were made, this is a time consuming process, it can take upwards of 4 hours for a single sample, making use of the entire dataset unfeasible. The existing manually annotated data can be used to train machine learning models to perform this segmentation task on the remainder of the data, allowing conclusions to be drawn on the remainder of the data. This is the primary task done in the remainder of this data.

Data Science Methods

Machine learning is used to segment defects and interesting characteristics from the images. Hypothesis testing is used to show the correlation between the extracted defects and the resulting mechanical properties.

Python Packages

- os - used for reading paths
- sys - used for adding folders to path
- cv2 - the most powerful library for image processing currently available in terms of both performance and capability
- numpy - allows images to be efficiently worked with as arrays. The foundation for cv2
- pandas - the most popular python library for working with dataframes
- math - used for mathematical functions
- random - used to create random variables
- matplotlib.pyplot - used for plotting
- re - python implementation of regex, a syntax for string processing
- datetime - used as a stopwatch to track performance
- ast - used to read in lists
- scipy - used for statistical tests
- seaborn - used to visualize standard statistical tests
- math - provides logarithmic and trigonometric functions

- joblib - package for parallelization of python scripts

organize_data is a script used to create the dataframe used in the rest of the report. It is imported here because some functions used to organize the data are also helpful in analysis.

```
import os
import sys
import cv2
import numpy as np
import pandas as pd
import math
import random
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import matplotlib.gridspec as gridspec
import re
import ast
import scipy
import seaborn
import math
import joblib
sys.path.append("/mnt/vstor/CSE_MSE_RXF131/cradle-members/mds3/aml334/mds3-advman-2/topics/ar")
import organize_data
```

organize_data

In addition to these package, the recent release of Meta's foundation models segment anything and segment anything 2 will be explored for use in characterizing the unsegmented initiating defects from high resolution images. [1] [2]

Exploratory Data Analysis

Explanation of your data set

The dataset mostly consists of unorganized images in 2 folders, and the numerical data is stored in a handful of spreadsheets. The dataframe was created in two parts: 1) The image dataframe, which contains the path to the images and 2) the numerical dataframe where are spreadsheets were merged. The id used to merge these dataframes is the sample_id, which was extracted from the image path using a regex and was cleaned from the existing Sample# column in the numerical spreadsheets. For the image dataframes, different categories were made, and each has a corresponding function which takes a file path as input and returns true

for whether the path leads to an image of that category. Data validation is done here to ensure that the categories are well defined.

Data Cleaning

Data cleaning was an iterative process, with the organize data script being run to create a dataframe, then the results analyzed and verified with a jupyter notebook. The organize data script is in it's own section, and the smaller chunks used to analyze the results are below it. Several sections are not fully cleaned, with the majority of the effort going to the fatigue and overload region, since they are the easiest to verify and should be the simplest task for the model since they are rather large. Some forms of data validation include:

1. Use of a size threshold to validate the separation between stitched low resolution images and individual low resolution images.
2. A lower red pixel threshold to remove pixels in the greyscale images which are identified as colored because of noise and a higher red pixel threshold to identify images where every defect is segmented versus only the initiating defect segmented.

The full data cleaning and formation of the combined_df can be seen in organize_data.py

Dataset Counts

The most important part of this output is the amount of images in the path files. Among the raw images there are 475 high resolution images of the initiating defect and 496 low resolution images of the entire fracture surface. There are far fewer labeled images than unlabeled images.

```
combined_df = pd.read_csv('/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/AdvManu/fractography/c
organize_data.print_column_counts(combined_df)
```

Column name	Nulls	Values	dtype	Example
Unnamed: 0	0	7066	int64	0
build_id_x	3632	3434	object	CMU1
build_plate_position_x	3632	3434	object	1
testing_position_x	3632	3434	float64	1.0
sample_id	0	7066	object	CMU1-1-1
cycles	3632	3434	float64	356846.0
test_stress_Mpa	3632	3434	float64	1067.0
scan_power_W	3632	3434	float64	370.0
scan_velocity_mm_s	3632	3434	float64	950.0
energy_density_J_mm3	3632	3434	float64	92.731829573934

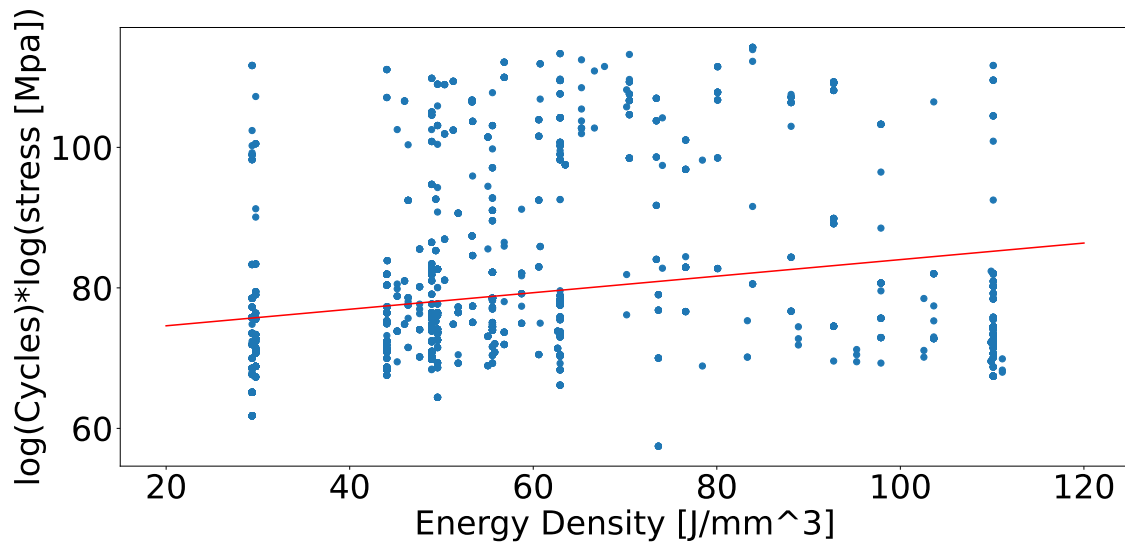
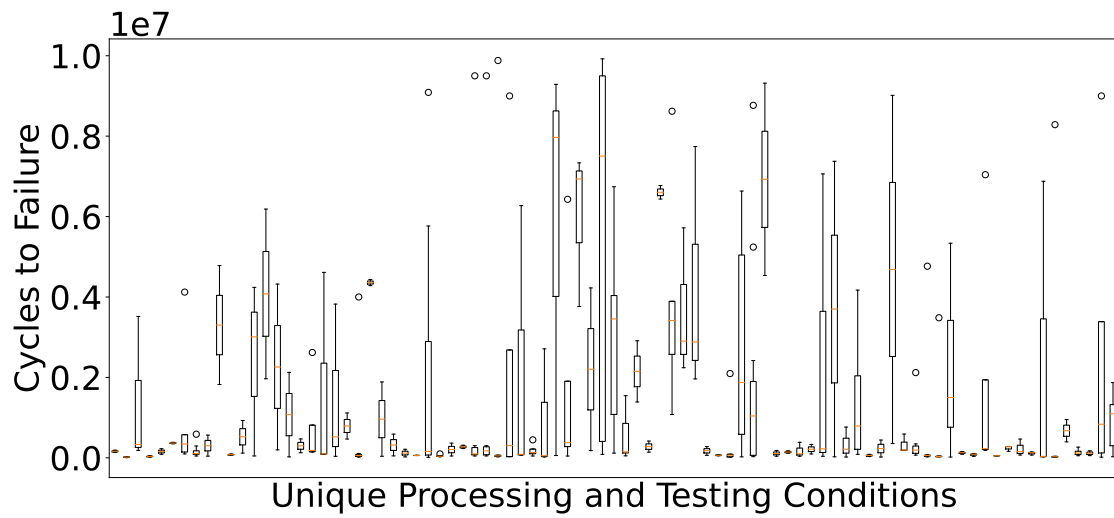
	image_path		110		6956		object		/mnt/vstor/CSE_	
	build_id_y		110		6956		object		CMU1	
	build_plate_position_y		110		6956		object		1	
	testing_position_y		1667		5399		float64		1.0	
	image_basename		110		6956		object		CMU01-1-1 INITI	
	image_class		110		6956		object		valid_image	
	points		5675		1391		object		[[742.0, 181.0]	

```
print(combined_df['image_class'].value_counts())
```

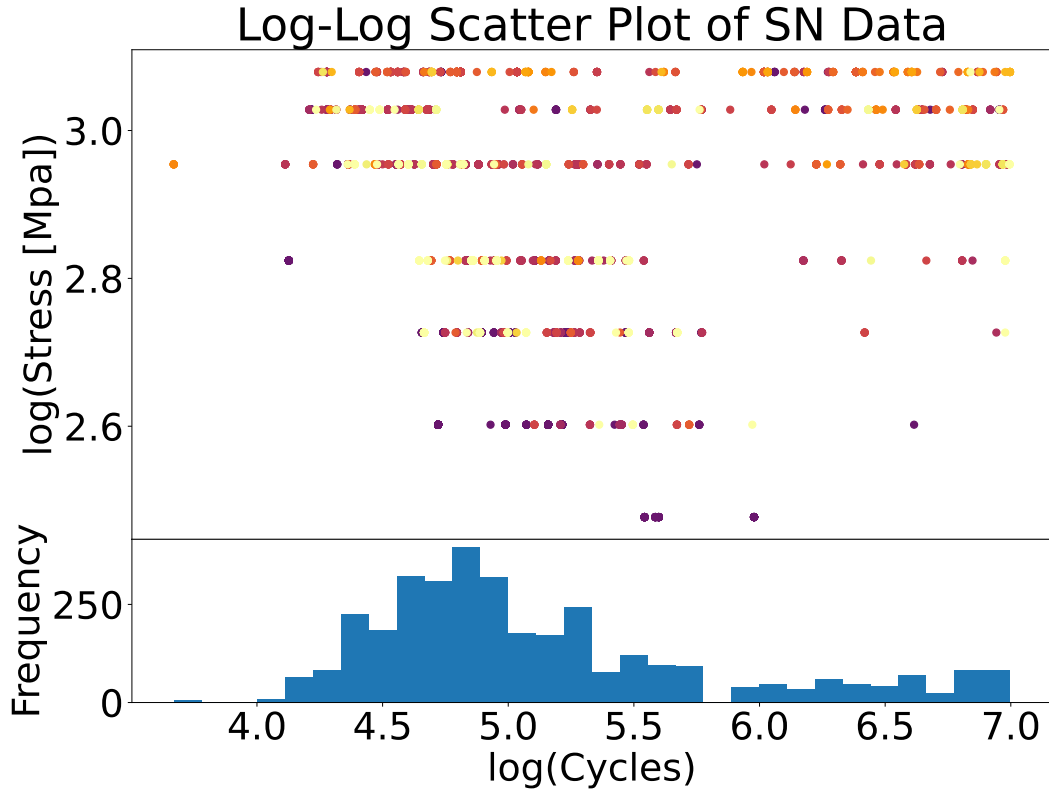
```
image_class
valid_image          4279
stitched             967
full_surface_unmarked 697
initiation           682
overload             112
fatigue              112
full_surface_marked   66
initiation_marked_stitched 41
Name: count, dtype: int64
```

Difference in Cycles to Failure

Below is a histogram and and scatter plot of the difference of two samples which had identical manufacturing and testing conditions. While this is a



The above figures show that the spreadsheets do not contain all of the necessary information, and from theory, we know that defects may be a leading cause. This is commonly visualized using $\log(\text{stress})$ vs $\log(\text{cycle})$ (SN) plots. We can also color the points according to their energy density.



Doing this, we can see that that the purple points, which is the samples with the most energy that have keyhole defects, were generally tested at a lower stress and failed with few cycles, while those with the least energy that have lack of fusion defects failed at higher stress with few cycles. The strongest samples, those tested at the highest stress and lasted the longest tend to be shades of orange. However, we can again see a lot of noise with some purple and yellow points being almost as strong as the other samples.

Statistical Learning

Fatigue Region

The fatigue region was selected as the first segmentation task because it is correlated with the fracture toughness, which is determined by the micro structure, which is otherwise unaccounted for. Since a tough micro structure is known to inhibit fatigue crack growth, this is a good candidate for a predictor of Cycles to failure difference. Additionally, the fatigue region is the largest part of the image and is the cleanest of the columns, so is the easiest to work with. The below script was used to train a unet model on segmenting the fatigue region.

Data augmentation was used to make the most of the small dataset. Since all stitched images are of different sizes, the first augmentation must standardize the size. The result was very successful. Initially, a randomized crop of the desired size was used, since that allows for a significantly larger and more diverse dataset than a more traditional resize. However, results in figure 3 show that the model’s training loss did not decrease with epochs, suggesting that the model is incapable of completing this task. All augmentations to the data were investigated, and the result was isolated to the random crop. A training loop with a resize transformation was used instead, shown in figure 4, and the training loss did decrease, though over training was a significant issue. This implies that the relationship of far away pixels is important for performing this task. Self-attention layers are known for their ability to learn global information, so AttentionUnet, a unet architecture which incorporates self-attention layers in skip connections was used, and it significantly outperformed the unet model with the same number of epochs and ultimately converged to a significantly lower training loss. Additionally, the wide interquartile range in both figure 3 and 4 was caused by a single mislabeled image. A well optimized training loop with the mislabeled data point corrected is shown in figure 5 and 6 for u-net and attention unet respectively.

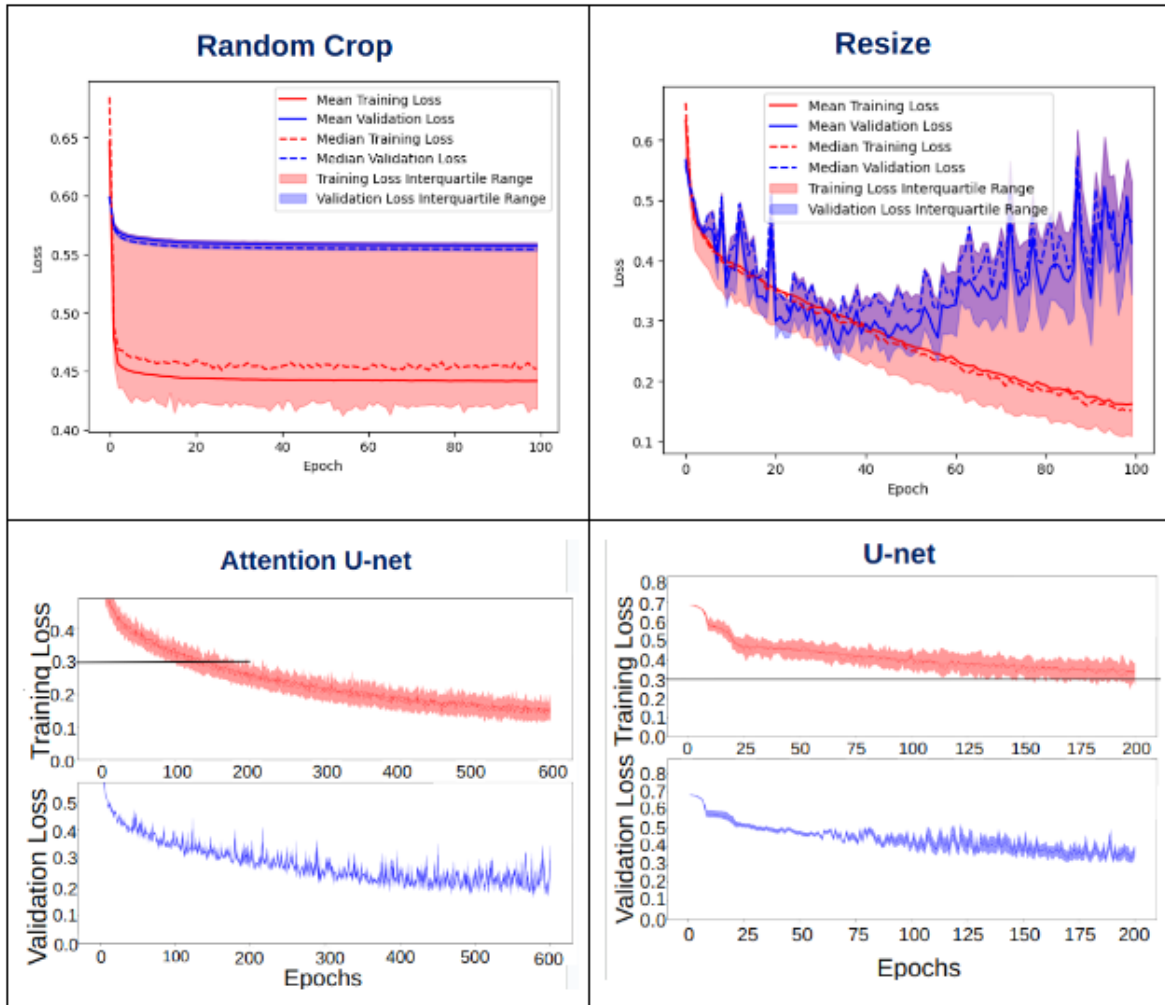


Figure 3: Model Training

Segment Anything Model on Initiating Defect

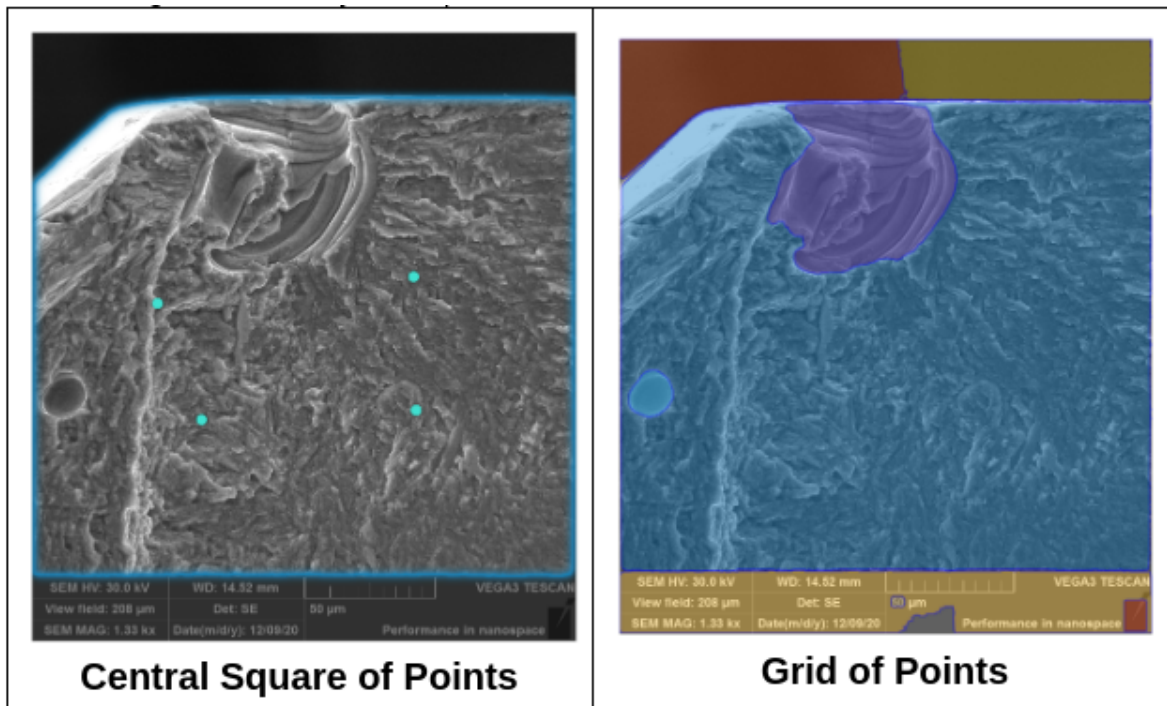


Figure 4: Example Shown for SAM Model

Figure 7 shows the segment anything model being used through it's webpage to segment an example initiating defect image. Based on this result, the surface can be segmented with a square of centralized points and the initiating defect is the largest defect inside of this surface. However, this is a high resolution image with good contrast between the surface, the foreground and the defect. An attempt was made to do this automatically, which was ultimately unsuccessful, so all images were labeled manually. The segmented results are saved as polygons whose points are in the points column of the dataframe.

The columns which were segmented.

```
points_df = combined_df[~combined_df['points'].isna()]
points_df['image_class'].value_counts()
```

```
image_class
valid_image    932
initiation     459
Name: count, dtype: int64
```

```
points_df['sample_id'].value_counts().head(10)
```

```
sample_id
NASA1-83-3    42
EP7-82-2      42
EP7-81-1      22
EP5-71-1      20
EP04-81-1     16
EP4-24-1      15
EP5-84-1      12
NASA1-83-1     12
CMU1-13-1     10
EP5-72-2      10
Name: count, dtype: int64
```

```
for group_string, group in points_df.groupby("sample_id"):
    imgs = group['image_path'].apply(cv2.imread)
    # point_lists = group['image_path'].apply(ast.literal_eval)
```

Sharpness

From knowledge of fracture mechanics, we know there is a link between the shape of the initiating defect and the resulting fatigue properties, so it should be possible to extract features from these images which predict fatigue performance. The most common feature is the aspect ratio. However, as shown in figure 8, shape 1 and 3 have similar aspect ratio, but shape 1 should create a significantly worse concentrating factor because of the sharp edge in the bottom right corner. Sharpness measures the change in radial distance from the centroid, which better captures these sharp edges, as can be seen by the high value for both shape 1 and 2.

```
def find_sharpness(numpy_array):
    #Convert array to dataframe
    y,x= np.nonzero(numpy_array)
    df = pd.DataFrame({'x':x,'y':y})
    x0 = df['x'].sum()/df['x'].count()
    y0 = df['y'].sum()/df['y'].count()
    #Calculate polar coordinates
    df['x_rel'] = df['x'] - x0
    df['y_rel'] = df['y'] - y0
    df['angle'] = df.apply(lambda row:math.atan(row['y_rel']/row['x_rel']),axis=1)
```

```

df['distance'] = df.apply(lambda row:math.sqrt(row['y_rel']**2 + row['x_rel']**2),axis=1)
global_max = df['distance'].max()
#Find max for each bin
num_bins = 180
bin_edges = np.linspace(-math.pi/2, math.pi/2, num_bins + 1)
bins = pd.IntervalIndex.from_breaks(bin_edges,name='Angle_bin')
df.index = pd.cut(df['angle'],bins)
max_df = df.groupby(level=0,observed=False)['distance'].max()
max_diff = []
for i in range(0,len(max_df)-2):
    max_diff.append(abs(max_df.iloc[i]-max_df.iloc[i+1])/global_max)

return max_diff
def calculate_aspect_ratio(mask):
    # Find the non-zero mask coordinates
    y_coords, x_coords = np.where(mask > 0)

    # Calculate the bounding box dimensions
    height = y_coords.max() - y_coords.min() + 1
    width = x_coords.max() - x_coords.min() + 1

    # Calculate the aspect ratio
    aspect_ratio = width / height
    return aspect_ratio

def invert_colors(image):
    #From claude 3 Haiku
    """
    Invert the colors of the image.
    For 8-bit images, use 255 - image.
    For floating point images (0 to 1), use 1 - image.
    """
    if image.dtype == np.uint8:
        return 255 - image
    else:
        return 1.0 - image

img_1 = invert_colors(cv2.imread('img_1.png',cv2.IMREAD_GRAYSCALE))
img_2 = invert_colors(cv2.imread('img_2.png',cv2.IMREAD_GRAYSCALE))
img_3 = invert_colors(cv2.imread('img_3.png',cv2.IMREAD_GRAYSCALE))

```




Shape	 1	 2	 3
Maximum Sharpness	0.050	0.057	0.011
Aspect Ratio	1.180	3.990	1.0

Figure 5: Well Optimized Attention Unet Architecture

```

save_path = '/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/AdvManu/fractography/SAM_whole_surfa

def load_SAM__segmentation():
    paths_list = []
    basename_lists = []
    for path in os.listdir(save_path):
        if 'seg' in path:
            paths_list.append(save_path+'/'+path)
            basename_lists.append(path.removeprefix("whole_surface_seg_"))
    return pd.concat([pd.Series(paths_list,name='image_path'),pd.Series(basename_lists,name=
segmented_df = load_SAM__segmentation()

# print(combined_df['image_basename'].drop_duplicates().value_counts())
# for group_string, basenames in combined_df.groupby('image_basename'):
#     if 'stitched' in sample['image_class'].value_counts().index:
#

x_not_SAM_test = []
x_SAM_test = []
for group_string, group in combined_df.groupby('sample_id'):
    no_points = group[(group['image_class']=='initiation') & (group['points'].isna())]
    points = group[(group['image_class']=='initiation') & (~group['points'].isna())]
    if not (len(no_points.index)>=1 and len(points.index)>=1):
        if len(no_points.index)>=1:

```

```

        cycles =no_points['cycles'].iloc[0]
        stress =no_points['test_stress_Mpa'].iloc[0]
        x_not_SAM_test.append(math.log(cycles)*math.log(stress))
    elif len(points.index)>=1:
        cycles =points['cycles'].iloc[0]
        stress =points['test_stress_Mpa'].iloc[0]
        x_SAM_test.append(math.log(cycles)*math.log(stress))

data = {
    "screen_portion": portion_of_screen,
    "max_sharpness": max_sharpness,
    "aspect_ratio": aspect_ratio,
    "perimeter": perimeter,
    "pixels": pixels,
    "pixel_perimeter_ratio": pixel_perimeter_ratio,
    "energy_density":energy,
    "stress Mpa":stress
}
df = pd.DataFrame(data).replace([np.inf, -np.inf], np.nan).dropna()

def annotate_r2(x, y, **kwargs):
    slope, intercept, r_value, p_value, std_err = scipy.stats.linregress(x, y)
    r_squared = r_value ** 2
    ax = plt.gca()
    ax.annotate(f"$R^2$ = {r_squared:.2f}", xy=(0.6, 0.9), xycoords=ax.transAxes, fontsize=20)

# Create pairplot with linear regression
cmap = plt.cm.viridis
palette = [cmap(i / 3) for i in range(len(df.columns))]
g = seaborn.pairplot(df,kind="reg", plot_kws={"line_kws": {"color": "red"}},hue="energy_dens.

# Adjust the size of axis labels and ticks using matplotlib
for ax in g.axes.flatten():
    ax.set_xlabel(ax.get_xlabel(), fontsize=14) # Set the font size of x-axis labels
    ax.set_ylabel(ax.get_ylabel(), fontsize=14) # Set the font size of y-axis labels
    ax.tick_params(axis='both', labelsize=12) # Set the font size of ticks
# Add R-squared annotations to each plot
# g.map(annotate_r2)
plt.show()

```



```

coefficients, residuals, rank, s = scipy.linalg.lstsq(X, Y)
coefficients, residuals, rank, s = scipy.linalg.lstsq(X, Y)
spline = scipy.interpolate.UnivariateSpline(X,Y,s=5)

```

Citations

- [1] A. Kirillov *et al.*, “Segment Anything.” arXiv, 2023. doi: [10.48550/ARXIV.2304.02643](https://arxiv.org/abs/2304.02643).
- [2] N. Ravi *et al.*, “SAM 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024, Available: <https://arxiv.org/abs/2408.00714>

Appendix

organize_data.py

```

# %% Import
import pandas as pd
import re
import datetime
import cv2
import numpy as np
import math
import os
import PIL
import numpy as np
import concurrent.futures as futures

STITCHED_THRESHOLD = 3000000
UPPER_RED_THRESHOLD = 1000
LOWER_RED_THRESHOLD = 20
HATCH_SPACING = 0.14 #mm
LAYER_THICKNESS = 0.03 #mm

LOG_FILE = '/home/aml334/CSE_MSE_RXF131/lab-staging/mds3/AdvManu/fractography/combined_df_log.csv'
MESSAGE = 'Specified joins because only images with points were being kept in the csv'

manuel_mask_path = '/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/keyence-fractography/manuel_mask.csv'
fractography_path = '/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography'
polygons_path = '/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/AdvManu/fractography/shape_infos.csv'

```

```

path_list = [fractography_path,manuel_mask_path]
pd.set_option('display.max_rows', None) # Display all rows

check = re.compile(r'''
    ^
    (?:x)?
    (?:\d+)?
    (?:[a-b])?
    [-]?
    (?:Copy\ of\ |Overview|_STD_ETD_|Initiation)? # Optional prefixes
    [-]?
    [\d]?
    [-]?
    [x]?
    (?:\d+)?
    [-]?
    (EP|NASA|CMU) # Start with EP, NASA, or CMU (case insens
    [-_]? # Optional separator
    (\d+|0\d+) # Number or 0 followed by number
    [-_]? # Optional separator
    V? # Version number
    ([E\d]+|\d+) # Optional separator
    [-_]? # Optional additional number
    (\d+)? # Optional '_MARKED' suffix
    (?:_MARKED)? # Any characters in between (greedy by de
    (.*)? # File extension
    \.(png|tif|tiff|jpg)$
''', re.VERBOSE | re.IGNORECASE)

def log_message(file_path,message):
    log_entry = f'[{datetime.datetime}] {message}\n'

    # Append the log entry to the file
    with open(file_path, 'a') as file:
        file.write(log_entry)

def standardize_sample_num(x):
    try:
        m_f = re.match(r'^([A-Z]+)(\d+)-[V]?(\d+|E\d+)[-]?(\d+)?',x)
    except TypeError:
        return None
    if m_f:
        if(m_f.lastindex==4):

```

```

        return m_f.group(1)+m_f.group(2).lstrip('0').lstrip('0')+'-'+m_f.group(3)+'-'+str(
    else:
        return m_f.group(1)+m_f.group(2).lstrip('0').lstrip('0')+'-'+m_f.group(3)+'-'+str(
    else:
        return None
def name_to_power(name:str, position:int):
    for idx, option in enumerate(process_parameters['Test ID']):
        if(option[2] == name[position]):
            return process_parameters['P (W)'][idx]
def name_to_velocity(name:str,position:int):
    for idx, option in enumerate(process_parameters['Test ID']):
        if(option[2] == name[position]):
            return process_parameters['V (mm/s)'][idx]
def clean_name(input):
    return input.astype(str).str.replace('0','').str.rstrip('.')
def clean_BuildID(input):
    try:
        match = re.match(r'([A-Z]+)(\d+)',input,re.IGNORECASE)
        if match:
            prefix = match.group(1).upper().lstrip('0')
            numeric_part = match.group(2).lstrip('0')
            return prefix + numeric_part
    except TypeError:
        print(input)
#Adds File if Condition(path) return True
def recursive_search(condition,path:str, file_list:list):
    if os.path.isdir(path):
        for path_loop in os.listdir(path):
            recursive_search(condition,os.path.join(path,path_loop),file_list)
    else:
        if(condition(path)):
            file_list.append(path)
    return file_list
#Functions used for Validation
def size(image_path):
    img = cv2.imread(image_path)
    try:
        return img.shape[0] * img.shape[1]
    except AttributeError:
        print('File is corrupted: '+image_path)
        return -1
def size_red(image_path):

```

```

img = cv2.imread(image_path)
try:
    #Convert the image from BGR to HSV color space
    hsv_image = cv2.cvtColor(img, cv2.COLOR_BGR2HSV)

    # Define the lower and upper bounds for the red color in HSV space
    lower_red_1 = np.array([0, 50, 50])
    upper_red_1 = np.array([10, 255, 255])
    lower_red_2 = np.array([170, 50, 50])
    upper_red_2 = np.array([180, 255, 255])

    # Create masks for the red color ranges
    mask1 = cv2.inRange(hsv_image, lower_red_1, upper_red_1)
    mask2 = cv2.inRange(hsv_image, lower_red_2, upper_red_2)

    # Combine the masks
    red_mask = mask1 + mask2

    # Count the number of red pixels
    red_pixels = cv2.countNonZero(red_mask)

    return(red_pixels)
except AttributeError:
    print('File is corrupted: '+image_path)
    return -1
def is_greyscale(image_path):
    img = cv2.imread(image_path)
    if len(img.shape) < 3:
        return True
    if img.shape[2] == 1:
        return True
    # If the image is color, check if all channels are equal
    if np.allclose(img[:, :, 0], img[:, :, 1]) and np.allclose(img[:, :, 1], img[:, :, 2]):
        return True
    return False
def is_binary(path):
    try:
        img = PIL.Image.open(path)
        img = img.convert('L')
        img_data = np.array(img)
        unique_vals = np.unique(img_data)

```

```

        if len(unique_vals) == 2 and set(unique_vals) == {0, 255}:
            return True
        else:
            return False
    except PIL.UnidentifiedImageError:
        return False

def is_8bit(path):
    try:
        img = PIL.Image.open(path)
        img = img.convert('L')
        img_data = np.array(img)
        if img_data.min() < 0 or img_data.max() > 255:
            return False
        else:
            return True
    except PIL.UnidentifiedImageError:
        return False

#Different columns that would be valuable to have
def valid_image(path):
    for i in ['.csv', '.hdr', '.xlsx', 'Contaminated']:
        if i in path:
            return False
    if is_8bit(path):
        return True
    else:
        return False
def marked(path):
    if 'marked' in path.lower() and valid_image(path) and not is_greyscale(path):
        return True
    else:
        return False
def initiation(path):
    if ('_001' in path or 'initiation' in path.lower()) and valid_image(path) and size(path) > 0:
        return True
    else:
        return False
def stitched(path):
    if ('stitched' in path.lower() or 'composite' in path.lower()) and (not 'weka' in path) and size(path) > 0:
        return True
    else:

```

```

        return False
def full_surface_marked(path):
    if stitched(path) and (not is_greyscale(path)) and (size_red(path)>UPPER_RED_THRESHOLD):
        return True
    else:
        return False
def initation_marked_stitched(path):
    if stitched(path) and ('MARKED' in path) and (not is_greyscale(path)) and (LOWER_RED_THR
        return True
    else:
        return False
def fatigue(path):
    if 'fatigue' in path.lower() and valid_image(path) and is_binary(path):
        return True
    else:
        return False
def overload(path):
    if 'overload' in path.lower() and valid_image(path) and is_binary(path):
        return True
    else:
        return False
def full_surface_unmarked(path):
    if stitched(path) and (not 'marked' in path.lower()) and (not is_binary(path)):
        return True
    else:
        return False
def exclude(input):
    conditions = ['.hdr', '.csv', '.info', '.xlsx', '.info', '.ptx', '.s0001', '.zip', '.model']
    for condition in conditions:
        if condition in input: return True
    return False

def check_regex_basename(dict_to_search,search_function, exclude_conditions):
    i=0
    for key in dict_to_search:
        for field in dict_to_search[key]:
            basename = field.split('/')[ -1]
            if not search_function(basename) and not exclude(basename):
                print(basename)
                i+=1
    print('Unselected files: '+str(i))
def regex_basename(pattern):

```

```

match = re.search(check,pattern)
if(match):
    type_func = clean_BuildID(match.group(1)+match.group(2))
    series_func = match.group(3).lstrip("0")
    if match.group(4):
        posit_idx_func = match.group(4).lstrip("0")
    else:
        posit_idx_func = None
    return type_func, series_func,posit_idx_func
else:
    return None
condition_list = [
    valid_image,
    initiation,
    stitched,
    full_surface_marked,
    initiation_marked_stitched,
    fatigue,
    overload,
    full_surface_unmarked
]
def print_column_counts(df,example=0):
    row_structure = '|{:~25}|{:~10}|{:~10}|{:~15}|{:~15}|'
    print(row_structure.format('Column name', 'Nulls', 'Values', 'dtype', 'Example'))
    for column in df.columns:
        nas = df[column].isna().sum()
        col_type = df[column].dtype
        print(row_structure.format(
            column,
            str(nas),
            str(len(df[column])-nas),
            str(col_type),
            str(df[column].iloc[example])[0:15],
        ))

# %%
'''Tidy EP and NADA data'''
if __name__=="__main__":
    # EP04,5,7 + NASA
    EP04 = pd.read_excel('/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography/EP04 (Complete)')
    EP05 = pd.read_excel('/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography/EP05/EP05 Fract')
    EP07 = pd.read_excel('/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography/EP07/EP07-Fract')
    NASA = pd.read_excel('/home/aml334/CSE_MSE_RXF131/staging/mds3/fractography/NASA03/NASA 1

```

```

EP_NASA_df = pd.concat([EP04,EP05,EP07,NASA])
EP_NASA_df['Sample#'] = EP_NASA_df['Sample#'].apply(standardize_sample_num)
del EP04
del EP05
del EP07
del NASA
EP_NASA_df = EP_NASA_df[['Sample#',' (Mpa)','Cycles']]
EP_NASA_df = EP_NASA_df.dropna()
EP_NASA_df = EP_NASA_df.rename(columns={
    'Sample#':'sample_id',
    ' (Mpa)':'test_stress_Mpa',
    'Cycles':'cycles'
})
EP_NASA_df = EP_NASA_df.astype({
    'sample_id':'string',
    'test_stress_Mpa':'float',
    'cycles':'int32'
})
# Needs to be done later
process_parameters = pd.read_csv('/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography/vari
EP_NASA_df['scan_power_W'] = EP_NASA_df['sample_id'].apply(lambda row: name_to_power(row,4))
EP_NASA_df['scan_velocity_mm_s'] = EP_NASA_df['sample_id'].apply(lambda row: name_to_velo
EP_NASA_df['energy_density_J_mm3'] = EP_NASA_df['scan_power_W']/(EP_NASA_df['scan_veloci
EP_NASA_df['build_id'] = EP_NASA_df['sample_id'].str.split("-").apply(lambda x: x[0])
EP_NASA_df['build_plate_position'] = EP_NASA_df['sample_id'].str.split("-").apply(lambda
EP_NASA_df['testing_position'] = EP_NASA_df['sample_id'].str.split("-").apply(lambda x:
del process_parameters

# %%
'''Tidying Brett's dataframe'''
#Brett Spreasheet
Brett_spreadsheet = pd.ExcelFile('/mnt/vstor/CSE_MSE_RXF131/staging/mds3/fractography/4-
Brett_df = pd.DataFrame()
for worksheet in Brett_spreadsheet.sheet_names:
    if worksheet not in ['Template','To Test','Retest']:
        Brett_df = pd.concat([Brett_df,pd.read_excel(Brett_spreadsheet,worksheet)])
del Brett_spreadsheet
Brett_df = Brett_df[['Build ID','Build #','Test #','Scan Power (W)','Scan velocity (mm/s)
Brett_df = Brett_df.rename(columns={
    'Build ID':'build_id',
    'Build #':'build_plate_position',
    'Test #':'testing_position',

```



```

    'Scan Power (W)': 'scan_power_W',
    'Scan velocity (mm/s)': 'scan_velocity_mm_s',
    ' max initiation (MPa)': 'test_stress_Mpa',
    'Cycles': 'cycles',
    })
#Filtering out NA values
Brett_df['testing_position'] = pd.to_numeric(Brett_df['testing_position'], errors='coerce')
Brett_df = Brett_df.dropna()
Brett_df = Brett_df.astype({
    'build_id': 'string',
    'build_plate_position': 'string',
    'testing_position': 'int32',
    'test_stress_Mpa': 'float',
    'cycles': 'int32'
})
Brett_df['build_id'] = Brett_df['build_id'].apply(clean_BuildID)
Brett_df['sample_id'] = Brett_df['build_id'] + '-' + Brett_df['build_plate_position'].apply(clean_BuildID)
#Filtering out mechanical data from times the sample didn't break (we get no metallography)
Brett_df = Brett_df.reset_index(drop=True)
Brett_df = Brett_df.iloc[Brett_df.groupby('sample_id')['Retest'].idxmax()]
Brett_df = Brett_df.drop(columns=['Retest'])
Brett_df['energy_density_J_mm3'] = Brett_df['scan_power_W'] / (Brett_df['scan_velocity_mm_s'] * Brett_df['cycles'])

# %%
'''Tidying Austin's Data'''
#Austin's spreadsheet
Austin_spreadsheet = pd.ExcelFile('/home/aml334/CSE_MSE_RXF131/staging/mds3/fractography/Austin.xlsx')
Austin_df = pd.DataFrame()
for worksheet in ['Fatigue Test Table', 'K calculation']:
    if 'Fatigue Test Table' in worksheet:
        x = pd.read_excel(Austin_spreadsheet, worksheet, skiprows=1)
        x['cycles'] = x['Cycles @ Failure']
        x['test_stress_Mpa'] = x['MPa']

    elif 'K calculation' in worksheet:
        x = pd.read_excel(Austin_spreadsheet, worksheet, skiprows=0)
        x['test_stress_Mpa'] = x['Mpa']
    x = x.loc[:, ~x.columns.str.startswith('Unnamed')]
    Austin_df = pd.concat([Austin_df, x])
del x
del Austin_spreadsheet
#Extracting information from spreadsheet

```

```

Austin_df = Austin_df.reset_index(drop=True)
ID_regex = re.compile(r"^(EP\d+|CMU\d+|NASA\d+)-(V?\d+)-(\d)",re.IGNORECASE)
result = Austin_df.loc[~Austin_df['ID'].isna(), 'ID'].apply(lambda x: re.search(ID_regex
result = result[~result.isna()]
Austin_df['build_id'] = result.apply(lambda x: clean_BuildID(x.group(1)))
Austin_df['build_plate_position'] = result.apply(lambda x: x.group(2))
Austin_df['testing_position'] = result.apply(lambda x: x.group(3))
Austin_df = Austin_df[~Austin_df['build_id'].isna()]
Austin_df['sample_id'] = Austin_df['build_id']+"-"+Austin_df['build_plate_position']+"-"
Austin_df = Austin_df[['build_id','build_plate_position','testing_position','sample_id'],
Austin_df = Austin_df.dropna()
Austin_df = Austin_df.astype({
    'build_id':'string',
    'build_plate_position':'string',
    'sample_id':'string',
    'testing_position':'int32',
    'test_stress_Mpa':'float',
    'cycles':'int32'
})
CMU_master_spreadsheet = pd.ExcelFile('/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/AdvMan
CMU_NASA_process_parameters_df = pd.DataFrame()
for worksheet in ['CMU01','CMU02','CMU03','CMU04','CMU05','CMU07','CMU08','CMU09','CMU10
    x = pd.read_excel(CMU_master_spreadsheet,worksheet,skiprows=23)
    x = x[['Unnamed: 0','power','velocity']]
    x = x.rename(columns={
        'Unnamed: 0':'build_plate_position',
        'power': 'scan_power_W',
        'velocity': 'scan_velocity_mm_s'
    })
    x['energy_density_J_mm3'] = x['scan_power_W']/(x['scan_velocity_mm_s'].apply(lambda
    x['build_id'] = clean_BuildID(worksheet)
    x = x.dropna().reset_index(drop=True)
    x['build_plate_position'] = x['build_plate_position'].astype(np.int64)
    CMU_NASA_process_parameters_df = pd.concat([CMU_NASA_process_parameters_df,x])
CMU_NASA_process_parameters_df=CMU_NASA_process_parameters_df.dropna().reset_index(drop=
CMU_NASA_process_parameters_df = CMU_NASA_process_parameters_df.astype({
    'build_id':'string',
    'build_plate_position':'string',
    'scan_power_W':'float',
    'scan_velocity_mm_s':'float',
    'energy_density_J_mm3':'float',
})

```

```

Austin_df = pd.merge(Austin_df,CMU_NASA_process_parameters_df,on=['build_id','build_plate_position'])
df = pd.concat([Austin_df,Brett_df,EP_NASA_df]).drop_duplicates()
df = df.astype({
    'sample_id':'string',
    'build_id':'string',
    'build_plate_position':'string',
    'testing_position':'int',
    'scan_power_W':'float',
    'energy_density_J_mm3':'float',
    'test_stress_Mpa':'float',
    'cycles':'int',
})
df['build_plate_position'] = df['build_plate_position'].str.removesuffix(".0")
df=df[df['cycles']!=10000000]
df = df.drop_duplicates().reset_index(drop=True)
x = df['sample_id'].value_counts()
x = x[x!=1].index
for sample in x:
    df = df[df['sample_id'].apply(lambda y: not sample in y)].reset_index(drop=True)
print('Mechanical Data')
print_column_counts(df)
# %%
'''Extracting Image file List'''
name = []
column_dict = {}
i = 0
column_list = []

def process_folder(column, top_folder):
    # Function to process each top_folder
    return recursive_search(column, top_folder, [])

for column in condition_list:
    temp_list = []
    with futures.ThreadPoolExecutor() as executor:
        # Submit tasks for processing each folder in parallel
        results = executor.map(process_folder, [column]*len(path_list), path_list)

        # Combine results
        for result in results:
            temp_list.extend(result)

```

```

# Store results
name.append((column.__name__, len(temp_list)))
column_dict[name[i][0]] = temp_list
print(str(name[i]) + f'\tPosition: {i}')
i += 1

# %%
'''Add Combine Dataframes'''
dataframe_list = []
for i, key in enumerate(column_dict):
    build_num_column = []
    build_plate_position_column = []
    test_position_column = []
    basename = []
    Sample_num = []
    path_column = []
    for j, field in enumerate(column_dict[key]):
        if regex_basename(field.split('/')[1]):
            path_column.append(field)
            type_inst, series_inst, posit_idx_inst = regex_basename(field.split('/')[1])
            build_num_column.append(type_inst)
            build_plate_position_column.append(series_inst)
            test_position_column.append(posit_idx_inst)
            basename.append(field.split('/')[1])
            if posit_idx_inst == None:
                Sample_num.append(type_inst.upper()+'-'+str(series_inst)+'-1')
            else:
                Sample_num.append(type_inst.upper()+'-'+str(series_inst)+'-'+str(posit_idx_inst))
    path_column = pd.Series(path_column, name='image_path')
    build_num_column = pd.Series(build_num_column, name='build_id')
    build_plate_position_column = pd.Series(build_plate_position_column, name='build_plate_position')
    test_position_column = pd.Series(test_position_column, name='testing_position')
    basename_column = pd.Series(basename, name='image_basename')
    Sample_num_column = pd.Series(Sample_num, name='sample_id')
    df_temp = pd.concat(
        [
            Sample_num_column,
            path_column,
            build_num_column,
            build_plate_position_column,
            test_position_column,
            basename_column,
        ],

```

```

no_retake = group[group['Retest']=='float(0)'\n",
"
    try:\n",
"
        max_difference_cycles.append(no_retake['Cycles'].max()-no_retake['Cycles'].min())\n",
"
    except ValueError:\n",
"
        print('In group: '+str(name)+' something weird is going on')\n",
plt.rcParams.update({'font.size': 16}) # Set font size for all elements\n",
fig, (ax1,ax2) = plt.subplots(1, 2, figsize=(12, 6))\n",
ax1.hist(max_difference_cycles,bins=50)\n",
ax1.set_xlabel('Max Difference in key Cycles to Failure')\n",
ax1.set_ylabel('Count')\n",
df_polygons = pd.read_csv(df_polygons_path+['File named if for use cycles']\n",
df_polygons = df_polygons.rename(columns={'columns': 'Failure'})\n",
ax2 = df_polygons.plot(kind='scatter',\n",
    'Points': 'points'\n",
} })\n",
df_polygons['points'] = df_polygons['points'].apply(lambda x: x.lstrip("\n"))\n",
df_imgs = pd.concat(dataframe_list, axis=0).merge(df_polygons,on='image_basename',how='left')\n",
df_final = pd.merge(df,df_imgs,on='sample_id',how='outer').reset_index(drop=True)\n",
df['display_name'] = "venv",\n",
df['language'] = "python",\n",
df['final_python'] = "/mnt/vstor/CSE_MSE_RXF131/lab-staging/mds3/AdvManu/fractography/combined"\n",
}, log_message(LOG_FILE, str(len(df_final))+ ' :'+MESSAGE)\n",
df['row_structure'] = '{:~50}|{:~10}|{:~10}|{:~15}|'\n",
df['row_structure'].format('Column name', 'Nulls', 'Values', 'Position'))\n",
df['name'] = "ipython",\n",
for column in df_final.columns:\n",
    nas = df_final[column].isna().sum()\n",
    df['row_structure'].format(column, str(nas), str(len(df_final[column])-nas), str(i))\n",
    log_message(LOG_FILE, str(row_structure.format(column, str(nas), str(len(df_final[column])-nas), str(i))))\n",
    df['name'] += "python",\n",
    nbconvert_exporter": "python",\n",
else:\n",
    pygments_lexer": "ipython3",\n",
    print("\nname_10.12")\n",
}\n",
},\n",
},\n",
}

def organize_data4.py
"nbformat_minor": 2
}

```