# 1-exploratory-data-analysis

May 12, 2024

```
[]: from google.colab import drive
     drive.mount('/content/drive')
    Mounted at /content/drive
[]: !pip install imagecodecs
    Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
    wheels/public/simple/
    Collecting imagecodecs
      Downloading
    {\tt imagecodecs-2021.11.20-cp37-cp37m-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl}
    (31.0 MB)
         I
                            | 31.0 MB 1.2 MB/s
    Requirement already satisfied: numpy>=1.16.5 in
    /usr/local/lib/python3.7/dist-packages (from imagecodecs) (1.21.6)
    Installing collected packages: imagecodecs
    Successfully installed imagecodecs-2021.11.20
[]: import cv2
     import datetime
     import gc
     import glob
     import math
     import matplotlib.pyplot as plt
     import numpy as np
     import os
     import pandas as pd
     import skimage.morphology
     import sys
     import imagecodecs
     import json
     import tifffile as tiff
     from matplotlib import colors
     from matplotlib import pyplot as plt
     from matplotlib.lines import Line2D
     from matplotlib_venn import venn2_unweighted
```

```
[]: plot_full_image = True
   # Number of glomeruli to display for each image
   num_glom_display = 5
   # Number of glomberuli to save as tiff files.
   num glom save = 5
   glob scale = 0.25
   base_path = '/content/drive/MyDrive/kidneysegmentation/'
   #Directory Contents
   print("Directory Contents")
   print('\n'.join(os.listdir(base_path)))
   train_images = sorted(glob.glob(os.path.join(base_path, 'train/*.tiff')))
   print(f'Number of training images: {len(train_images)}')
   print('\n'.join(train_images))
   #Test Images
   test_images = sorted(glob.glob(os.path.join(base_path, 'test/*.tiff')))
   print(f'Number of test images: {len(test_images)}')
   print('\n'.join(test_images))
   df train = pd.read csv(os.path.join(base path, 'train.csv'))
   df_info = pd.read_csv(os.path.join(base_path, 'HuBMAP-20-dataset_information.
    ⇔csv'))
   Directory Contents
   sample submission.csv
   train.csv
   .ipynb_checkpoints
```

```
/content/drive/MyDrive/kidneysegmentation/train/54f2eec69.tiff
    /content/drive/MyDrive/kidneysegmentation/train/8242609fa.tiff
    /content/drive/MyDrive/kidneysegmentation/train/aaa6a05cc.tiff
    /content/drive/MyDrive/kidneysegmentation/train/afa5e8098.tiff
    /content/drive/MyDrive/kidneysegmentation/train/b2dc8411c.tiff
    /content/drive/MyDrive/kidneysegmentation/train/b9a3865fc.tiff
    /content/drive/MyDrive/kidneysegmentation/train/c68fe75ea.tiff
    /content/drive/MyDrive/kidneysegmentation/train/cb2d976f4.tiff
    /content/drive/MyDrive/kidneysegmentation/train/e79de561c.tiff
    Number of test images: 5
    /content/drive/MyDrive/kidneysegmentation/test/2ec3f1bb9.tiff
    /content/drive/MyDrive/kidneysegmentation/test/3589adb90.tiff
    /content/drive/MyDrive/kidneysegmentation/test/57512b7f1.tiff
    /content/drive/MyDrive/kidneysegmentation/test/aa05346ff.tiff
    /content/drive/MyDrive/kidneysegmentation/test/d488c759a.tiff
[]: import cv2
    import glob
    import json
    import numpy as np
    import os
    import pandas as pd
    import tifffile as tiff
    from matplotlib import colors
    from matplotlib import pyplot as plt
    from matplotlib.lines import Line2D
    from matplotlib_venn import venn2_unweighted
[]: train images = list(map(lambda x: os.path.basename(x), train images))
    test_images = list(map(lambda x: os.path.basename(x), test_images))
[]: pd.set_option('display.max_colwidth', None)
    dataset_information = pd.read_csv(base_path + 'HuBMAP-20-dataset_information.
      ⇔csv')
    dataset_information['pixels_total'] = dataset_information.width_pixels *__
      →dataset_information.height_pixels
[]: train_dataset_information =
      odataset_information[dataset_information['image_file'].isin(train_images)].
      ⇔reset_index(drop=True)
    test dataset information = ___
      dataset_information[dataset_information['image_file'].isin(test_images)].
      ⇔reset_index(drop=True)
```

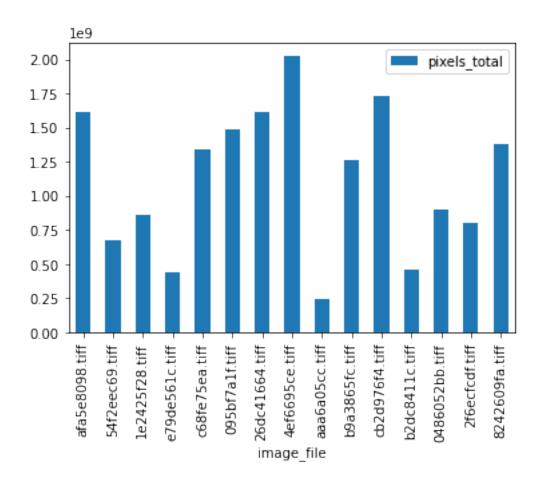
## 0.1 Training Dataset

Image resolution table and bar chart representation

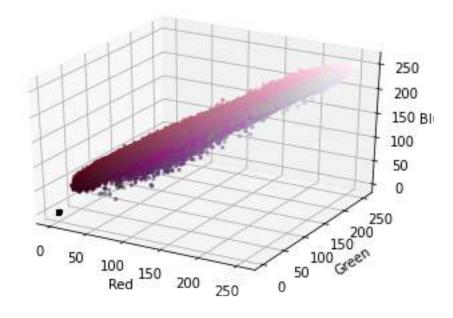
```
[]:
             image_file
                          width_pixels height_pixels
         4ef6695ce.tiff
                                 50680
                                                 39960
     0
     1
         cb2d976f4.tiff
                                  49548
                                                 34940
     2
         26dc41664.tiff
                                  42360
                                                 38160
     3
         afa5e8098.tiff
                                  43780
                                                 36800
     4
         095bf7a1f.tiff
                                  39000
                                                 38160
         8242609fa.tiff
                                                 31299
     5
                                 44066
     6
         c68fe75ea.tiff
                                 49780
                                                 26840
     7
         b9a3865fc.tiff
                                 40429
                                                 31295
         0486052bb.tiff
     8
                                  34937
                                                 25784
         1e2425f28.tiff
                                                 26780
                                  32220
         2f6ecfcdf.tiff
                                  25794
                                                 31278
     11
         54f2eec69.tiff
                                  22240
                                                 30440
     12 b2dc8411c.tiff
                                  31262
                                                 14844
         e79de561c.tiff
     13
                                  27020
                                                 16180
     14 aaa6a05cc.tiff
                                  13013
                                                 18484
```

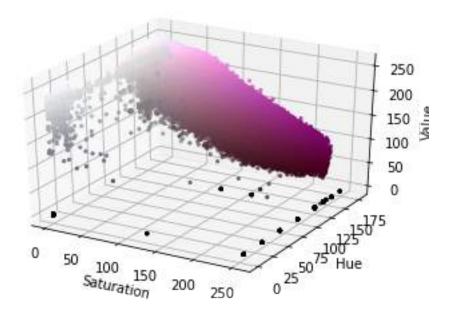
<sup>[]:</sup> train\_dataset\_information.plot.bar(x='image\_file', y='pixels\_total', rot=90)

<sup>[]: &</sup>lt;matplotlib.axes.\_subplots.AxesSubplot at 0x7efe299fc510>



## RGB and HSV color spaces of aaa6a05cc.tiff





#### Glomeruli Count

```
[]: train_glom_seg_files = __
      otrain_dataset_information['glomerulus_segmentation_file'].to_list()
     train_glomeruli_dict = {}
     for file_name in train_glom_seg_files:
         file_id = file_name[:9]
         with open(f'/content/drive/MyDrive/kidneysegmentation/train/{file_name}')_
      →as json_file:
             data = json.load(json_file)
             train_glomeruli_dict[file_id] = 0
             for entry in data:
                 if entry['type'] == 'Feature' and_
      →entry['properties']['classification']['name'] == 'glomerulus':
                         train_glomeruli_dict[file_id] += 1
                 else:
                     raise Exception(f"Unexpected json format: {entry['type']}, __

¬{entry['properties']['classification']['name']}")
     train_nr_glom = pd.DataFrame(list(train_glomeruli_dict.items()),__

→columns=['file_id', 'nr_glomeruli'])
     train_nr_glom.sort_values('nr_glomeruli', ascending=False).
      ⇔reset_index(drop=True)
```

```
[]: file_id nr_glomeruli
0 8242609fa 586
```

```
1
   b9a3865fc
                        469
2
    4ef6695ce
                        439
3
    095bf7a1f
                        350
4
    cb2d976f4
                        319
    26dc41664
                        245
6
   afa5e8098
                        235
7
   e79de561c
                        180
8
    1e2425f28
                        178
    2f6ecfcdf
                        160
10 54f2eec69
                        139
11 b2dc8411c
                        138
12 0486052bb
                        130
13 c68fe75ea
                        118
14 aaa6a05cc
                         99
```

```
[]: #total count train_nr_glom['nr_glomeruli'].sum()
```

#### []: 3785

Glomeruli Size Distribution

```
[]: train_glom_seg_files =
     otrain_dataset_information['glomerulus_segmentation_file'].to_list()
    train_glomeruli_polys_dict = {}
    for file_name in train_glom_seg_files:
        with open(f'/content/drive/MyDrive/kidneysegmentation/train/{file name}'),
      →as json_file:
            data = json.load(json_file)
            train_glomeruli_polys_dict[file_name] = []
            for entry in data:
                if entry['type'] == 'Feature' and_
      →entry['properties']['classification']['name'] == 'glomerulus':
                    geom = np.array(entry['geometry']['coordinates']).astype(np.
      ⊶float32)
                    x,y,w,h = cv2.boundingRect(geom.squeeze(axis=0))
                    train_glomeruli_polys_dict[file_name].append((h,w)) # height,_
      \rightarrow width!
                else:
                    raise Exception(f"Unexpected json format: {entry['type']}, __
      train_res_glom = pd.DataFrame(list(train_glomeruli_polys_dict.items()),__

¬columns=['glomerulus_segmentation_file', 'glomeruli_height_width'])

    train_res_glom = train_res_glom.explode('glomeruli_height_width')
```

```
train_res_glom['height'], train_res_glom['width'] = ___
      \rightarrow zip(*train_res_glom['glomeruli_height_width'])
    train_res_glom = train_res_glom.drop(columns=['glomeruli_height_width'])
    train res glom.describe()
[]:
                height
                              width
    count 3785.000000 3785.000000
            322.601057
                         325.266579
    mean
    std
             84.198179
                          86.372246
    min
            94.000000
                         103.000000
    25%
            262.000000
                         263.000000
    50%
            318.000000
                         319.000000
    75%
            383.000000
                         385.000000
            626.000000
                         659.000000
    max
    Anatomical Structure
[]: def list_structures(seg_files, folder):
        train_anatomical_dict = {}
        folder_path = os.path.join(base_path, folder)
        for file_name in seg_files:
            file_id = file_name[:9]
            with open(os.path.join(folder_path, file_name)) as json_file:
                data = json.load(json_file)
                train_anatomical_dict[file_id] = []
                for entry in data:
                    if entry['type'] == 'Feature':
                        train_anatomical_dict[file_id].
      →append(entry['properties']['classification']['name'])
                    else:
                        raise Exception(f"Unexpected json format: {entry['type']}, __
      return pd.DataFrame(list(train_anatomical_dict.items()),__
      ⇔columns=['file_id', 'anatomical_structure'])
[]: train_anatomical_seg_files =_
      otrain_dataset_information['anatomical_structures_segmention_file'].to_list()
     train_anatomical = list_structures(train_anatomical_seg_files, 'train')
    train_anatomical
[]:
                                                   anatomical_structure
          file_id
                   [Outer Medulla, Outer Stripe, Cortex, Inner medulla]
        afa5e8098
    1
        54f2eec69
                                                      [Medulla, Cortex]
```

[Cortex, Medulla]

[Cortex]

1e2425f28

e79de561c

```
4
        c68fe75ea
                                                     [Cortex, Medulla]
        095bf7a1f
    5
                                                              [Cortex]
        26dc41664
    6
                   [Cortex, Outer Stripe, Inner medulla, Outer Medulla]
    7
        4ef6695ce
                                                     [Medulla, Cortex]
    8
        aaa6a05cc
                                                     [Medulla, Cortex]
        b9a3865fc
                                                     [Cortex, Medulla]
    9
    10 cb2d976f4
                                             [Cortex, Cortex, Medulla]
                                             [Cortex, Cortex, Medulla]
    11 b2dc8411c
                                                     [Medulla, Cortex]
    12 0486052bb
    13 2f6ecfcdf
                                                     [Medulla, Cortex]
                                                     [Medulla, Cortex]
    14 8242609fa
[]: #listing all unique anatomical structures
    pd.DataFrame(train_anatomical.
      -explode('anatomical structure')['anatomical structure'].unique(),
      []:
      unique_structures
          Outer Medulla
    0
    1
           Outer Stripe
```

Mask Visualization

Cortex

Medulla

Inner medulla

2

```
[]: def make_grid(shape, window=256, min_overlap=32):
         function to generate a grid layout for sliding window
         :param shape: height and width of the image
         :param window: size of the window
         :param min_overlap: minimal window overlap
         :return: array of window coordinates (x1, x2, y1, y2)
         x, y = shape
         nx = x // (window - min_overlap) + 1
         x1 = np.linspace(0, x, num=nx, endpoint=False, dtype=np.int64)
         x1[-1] = x - window
         x2 = (x1 + window).clip(0, x)
         ny = y // (window - min_overlap) + 1
         y1 = np.linspace(0, y, num=ny, endpoint=False, dtype=np.int64)
         y1[-1] = y - window
         y2 = (y1 + window).clip(0, y)
         slices = np.zeros((nx,ny, 4), dtype=np.int64)
         for i in range(nx):
             for j in range(ny):
```

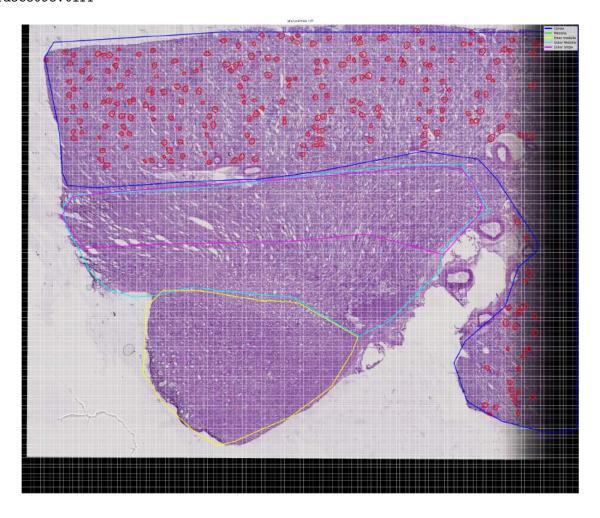
```
slices[i,j] = x1[i], x2[i], y1[j], y2[j]
return slices.reshape(nx*ny,4)
```

```
[]: def plot_masks(dataset_information, folder, frame_size, frame_overlap,__
      →plot_glom):
         folder_path = os.path.join(base_path, folder)
         for i in range(len(dataset_information)):
             # create new figure
             plt.figure(figsize=(32, 30))
             obj_line_thickness = 60
             # find metadata row for json file
             image_metadata = dataset_information.iloc[i]
             # open image file
             image = tiff.imread(os.path.join(folder_path,__
      →image_metadata['image_file']))
             print(image_metadata['image_file'])
             # reshape image if necessary
             if len(image.shape) == 5:
                 image = image.squeeze()
             if image.shape[0] == 3:
                 image = image.transpose(1, 2, 0)
             # create a copy of the image
             image = image.copy()
             # draw sliding window boxes
             frame_grid = make_grid((image.shape[1], image.shape[0]), frame_size,__
      →frame_overlap)
             for frame in frame_grid:
                 x1, y1 = frame[0], frame[2]
                 x2, y2 = frame[1], frame[3]
                 image = cv2.rectangle(image, (x1, y1), (x2, y2),__
      \hookrightarrowcolor=(255,255,255), thickness=16)
             # draw glomeruli polygons
             if plot glom:
                 # open glomeruli json file
                 read_glom_seg_file = open(os.path.join(folder_path,__
      →image_metadata['glomerulus_segmentation_file']), 'r')
                 glom_seg_data = json.load(read_glom_seg_file)
```

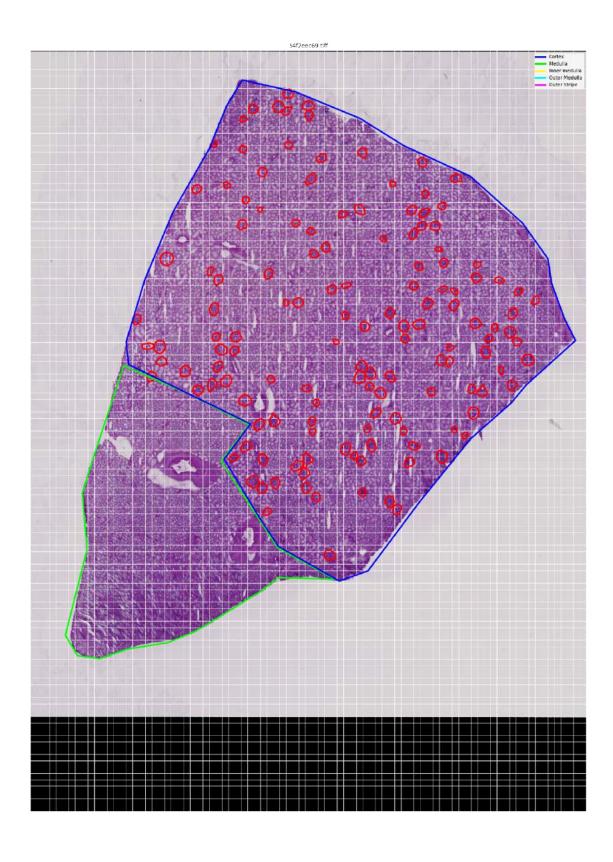
```
for k in range(len(glom_seg_data)):
              glom_poly = np.
⇒array(glom_seg_data[k]['geometry']['coordinates']).astype(np.int32) # qet_⊔
⇔coordinates of glomeruli
               cv2.polylines(image, glom_poly, True,(255,0,0),__
⇔thickness=obj line thickness)
       # open anatomical json file
      read_anatomical_seg_file = open(os.path.join(folder_path,__
→image_metadata['anatomical_structures_segmention_file']), 'r')
       anatomical_seg_data = json.load(read_anatomical_seg_file)
       # scan anatomical json file and draw lines
      for n in range(len(anatomical_seg_data)):
           obj_name =
→anatomical_seg_data[n]['properties']['classification']['name']
          obj_coords = anatomical_seg_data[n]['geometry']['coordinates']
          if (obj name == 'Cortex'): # draw line around cortex
               cv2.polylines(image, np.expand dims(np.array(obj coords[0]).
astype(np.int32), axis=0), True, (0,0,255), thickness=obj_line_thickness)
          elif (obj_name == 'Medulla'): # draw line around medulla
               cv2.polylines(image, np.array(obj_coords).astype(np.int32),_
→True, (0,255,0), thickness=obj_line_thickness)
          elif (obj_name == 'Inner medulla'): # draw line around inner medulla
              cv2.polylines(image, np.array(obj_coords).astype(np.int32),__
→True, (255,255,0), thickness=obj_line_thickness)
          elif (obj_name == 'Outer Medulla'): # draw line around outer medulla
               cv2.polylines(image, np.array(obj_coords).astype(np.int32),__
→True, (0,255,255), thickness=obj_line_thickness)
          elif (obj_name == 'Outer Stripe'): # draw line around outer stripe
              cv2.polylines(image, np.array(obj_coords).astype(np.int32),__
→True, (255,0,255), thickness=obj_line_thickness)
          else:
              raise Exception(f'Unknown anatomical object: {obj_name}')
       # down-scale the image
       image_resize = cv2.resize(image,(image.shape[1]//10,image.shape[0]//
→10), interpolation = cv2.INTER_CUBIC)
       # add legend and view the image
       custom_lines = [Line2D([0], [0], color=(0.,0.,1.), lw=4),
                   Line2D([0], [0], color=(0.,1.,0.), lw=4),
                   Line2D([0], [0], color=(1.,1.,0.), lw=4),
                   Line2D([0], [0], color=(0.,1.,1.), lw=4),
                   Line2D([0], [0], color=(1.,0.,1.), lw=4),
```

[]: plot\_masks(train\_dataset\_information, 'train', 1024, 256, True)

afa5e8098.tiff

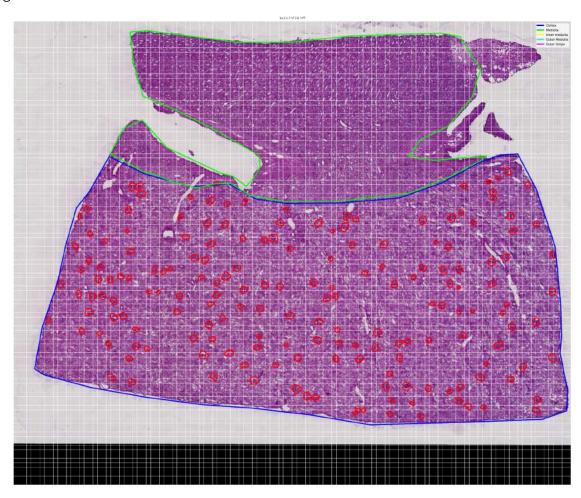


54f2eec69.tiff <Figure size 432x288 with 0 Axes>

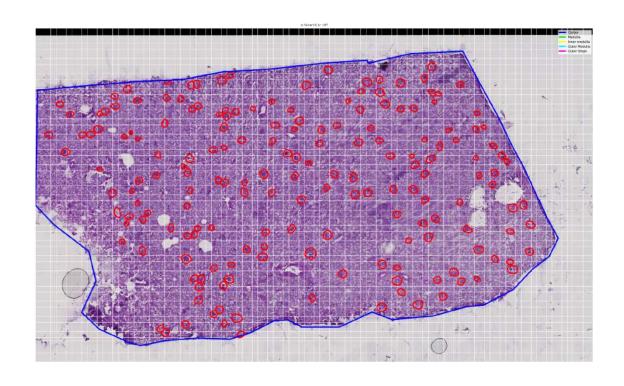


1e2425f28.tiff

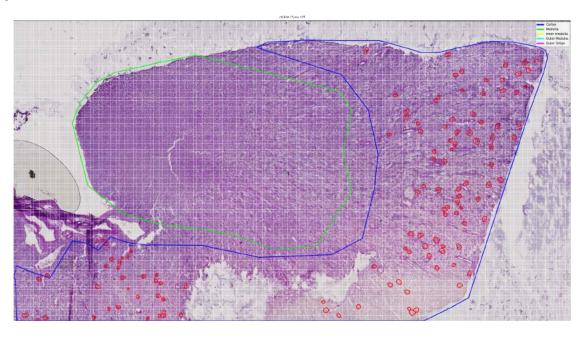
<Figure size 432x288 with 0 Axes>



e79de561c.tiff <Figure size 432x288 with 0 Axes>

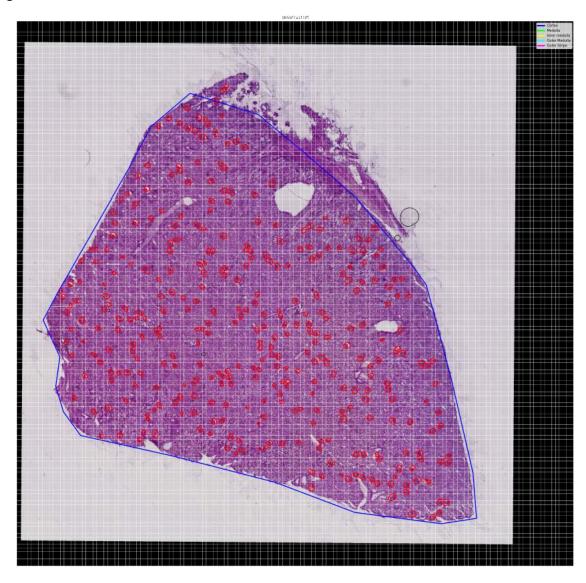


c68fe75ea.tiff <Figure size 432x288 with 0 Axes>

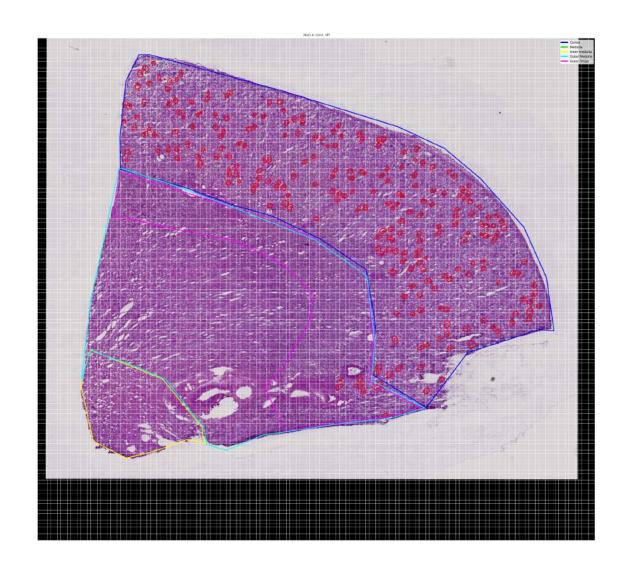


095bf7a1f.tiff

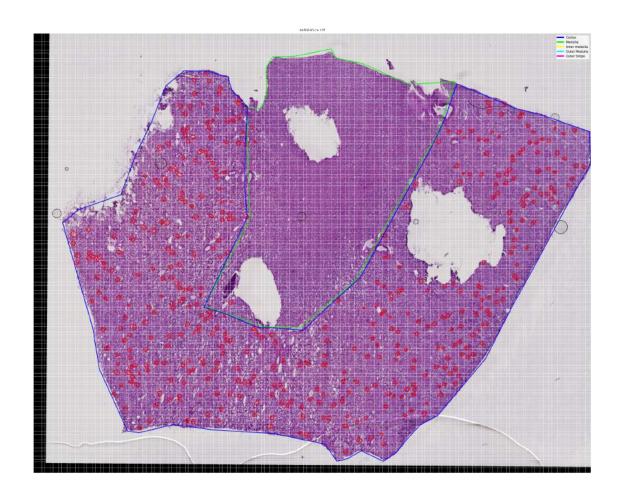
<Figure size 432x288 with 0 Axes>



26dc41664.tiff
<Figure size 432x288 with 0 Axes>



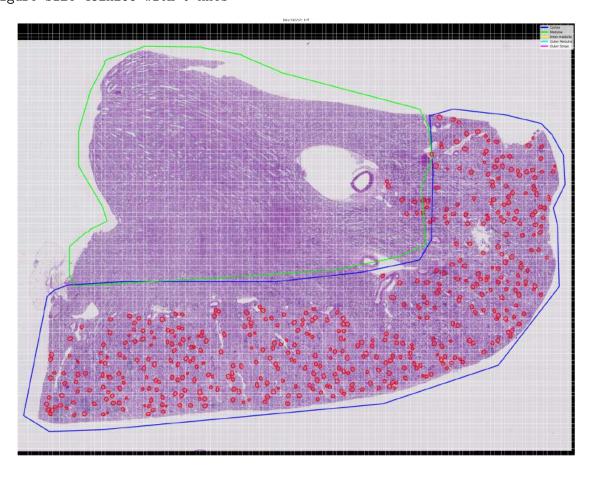
4ef6695ce.tiff
<Figure size 432x288 with 0 Axes>



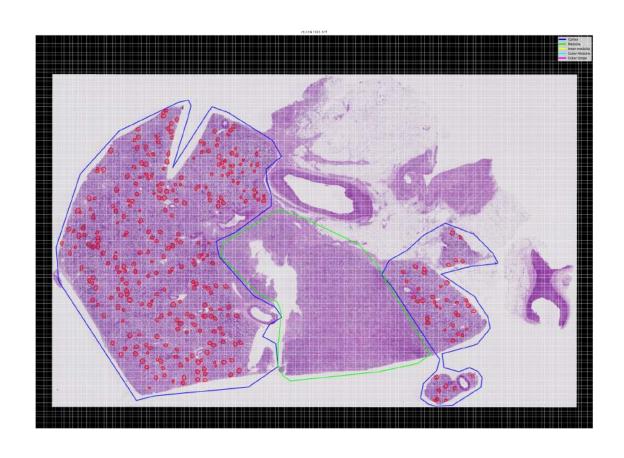
aaa6a05cc.tiff
<Figure size 432x288 with 0 Axes>



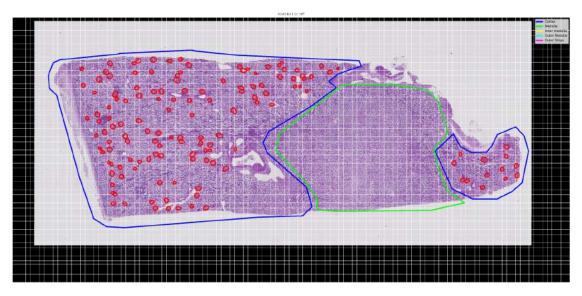
b9a3865fc.tiff
<Figure size 432x288 with 0 Axes>



cb2d976f4.tiff
<Figure size 432x288 with 0 Axes>



b2dc8411c.tiff
<Figure size 432x288 with 0 Axes>



```
KeyboardInterrupt
                                          Traceback (most recent call last)
<ipython-input-15-b5441024a68b> in <module>()
---> 1 plot_masks(train_dataset_information, 'train', 1024, 256, True)
<ipython-input-14-c87f6c46683c> in plot_masks(dataset_information, folder,__
 →frame_size, frame_overlap, plot_glom)
     12
     13
                # open image file
---> 14
                image = tiff.imread(os.path.join(folder_path,__
 →image metadata['image file']))
                print(image_metadata['image_file'])
     16
/usr/local/lib/python3.7/dist-packages/tifffile/tifffile.py in imread(files,,,
 ⇔aszarr, **kwargs)
    892
                        if aszarr:
    893
                            return tif.aszarr(**kwargs)
--> 894
                        return tif.asarray(**kwargs)
    895
    896
            with TiffSequence(files, **kwargs_seq) as imseq:
/usr/local/lib/python3.7/dist-packages/tifffile/tifffile.py in asarray(self, u
 ⇔key, series, level, squeeze, out, maxworkers)
   3372
  3373
                elif len(pages) == 1:
-> 3374
                    result = pages[0].asarray(out=out, maxworkers=maxworkers)
  3375
                else:
   3376
                    result = stack_pages(pages, out=out, maxworkers=maxworkers)
/usr/local/lib/python3.7/dist-packages/tifffile/tifffile.py in asarray(self, u
 →out, squeeze, lock, maxworkers)
   6815
                else:
   6816
                    # decode individual strips or tiles
-> 6817
                    result = create_output(out, keyframe.shaped, keyframe._dtyp
   6818
                    keyframe.decode # init TiffPage.decode function
   6819
/usr/local/lib/python3.7/dist-packages/tifffile/tifffile.py in_
 →create_output(out, shape, dtype, mode, suffix, fillvalue)
                    out[:] = fillvalue
  16911
  16912
                    return out
> 16913
               return numpy.zeros(shape, dtype)
  16914
            if isinstance(out, numpy.ndarray):
                if product(shape) != product(out.shape):
  16915
```

## KeyboardInterrupt:

```
<Figure size 432x288 with 0 Axes>
<Figure size 2304x2160 with 0 Axes>
```

## 0.2 Test Data

Image Resolution

1 d488c759a.tiff

3 2ec3f1bb9.tiff

4 3589adb90.tiff

```
[]: test_dataset_information.sort_values('pixels_total',__
      →ascending=False)[['image_file', 'width_pixels','height_pixels']]
[]:
           image_file width_pixels height_pixels
    0 aa05346ff.tiff
                              47340
                                             30720
    2 57512b7f1.tiff
```

33240

46660

23990

29433

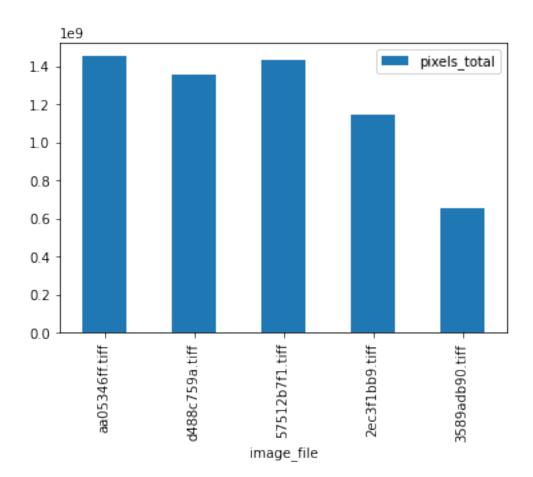
```
[]: test_dataset_information.plot.bar(x='image_file', y='pixels_total')
```

[]: <matplotlib.axes.\_subplots.AxesSubplot at 0x7efde9e65290>

43160

29020

47723



## Anatomical Structure

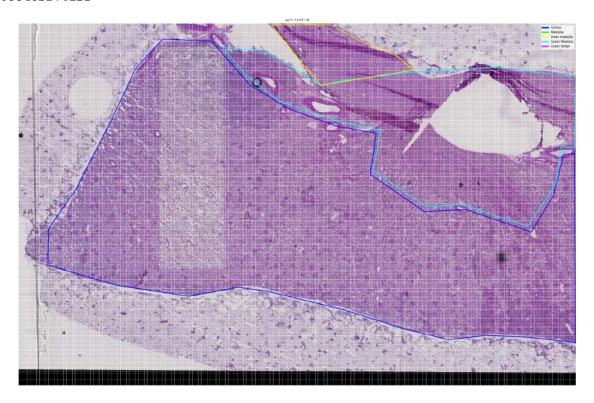
```
[]: test_anatomical_seg_files =__
      otest_dataset_information['anatomical_structures_segmention_file'].to_list()
     test_anatomical = list_structures(test_anatomical_seg_files, 'test')
[]: pd.DataFrame(test_anatomical.
      ⊖explode('anatomical_structure')['anatomical_structure'].unique(),⊔

→columns=['unique_structures'])
[]:
      unique_structures
                  Cortex
           Outer Medulla
     1
     2
           Inner medulla
                Medulla
```

## Mask Visualization

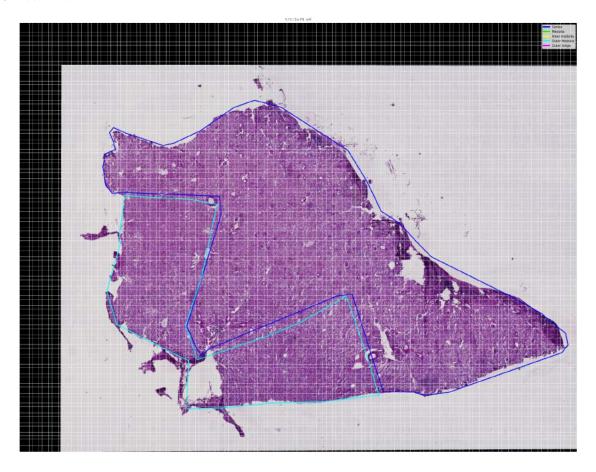
```
[]: plot_masks(test_dataset_information, 'test', 1024, 256, False)
```

aa05346ff.tiff

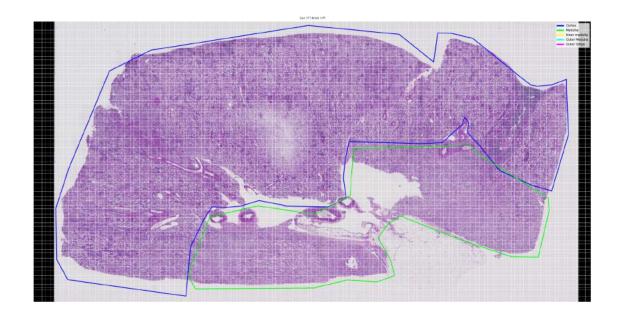


d488c759a.tiff

57512b7f1.tiff



2ec3f1bb9.tiff



3589adb90.tiff

