Note: This note book is to provide a general structure of the project and responsibility distribution

In [2]:

```
from sklearn.base import BaseEstimator, RegressorMixin
from scipy.optimize import minimize
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import GridSearchCV, PredefinedSplit
from sklearn.model_selection import ParameterGrid
from sklearn.metrics import mean_squared_error, make_scorer
import pandas as pd
import random
from collections import Counter

from random import shuffle
```

Calcification Train

In [4]:

```
cal_train_desc = pd.read_csv('calc_case_description_train_set.csv')
mass_train_desc = pd.read_csv('mass_case_description_train_set.csv')
```

In [7]:

```
cal_train_desc.sort_values(by = 'patient_id', inplace = True)
```

In [29]:

```
#cal_train_desc.head(n=10)
```

In [17]:

```
cal_train_desc.columns
```

Out[17]:

```
Index(['patient_id', 'breast density', 'left or right breast', 'image
view',
       'abnormality id', 'abnormality type', 'calc type', 'calc distri
bution',
       'assessment', 'pathology', 'subtlety', 'image file path',
       'cropped image file path', 'ROI mask file path'],
      dtype='object')
```

In [18]:

```
columns = ['image view', 'abnormality id', 'abnormality type', 'calc typ
      'subtlety']
```

In [20]:

```
for col in columns:
    print('---------------')
    print('Unique values of column {}'.format(col))
    print(cal_train_desc[col].value_counts())
```

```
---------------
Unique values of column image view
MLO    807
CC     739
Name: image view, dtype: int64
---------------
Unique values of column abnormality id
1    1172
2     219
3      88
4      35
5      20
6      10
7       2
Name: abnormality id, dtype: int64
---------------
Unique values of column abnormality type
calcification    1546
Name: abnormality type, dtype: int64
---------------
Unique values of column calc type
PLEOMORPHIC                                                 664
AMORPHOUS                                                   138
PUNCTATE                                                    106
LUCENT_CENTER                                                93
VASCULAR                                                     82
FINE_LINEAR_BRANCHING                                        77
COARSE                                                       35
ROUND_AND_REGULAR-LUCENT_CENTER                              31
PLEOMORPHIC-FINE_LINEAR_BRANCHING                            28
ROUND_AND_REGULAR-LUCENT_CENTER-PUNCTATE                     24
ROUND_AND_REGULAR-EGGSHELL                                   23
PUNCTATE-PLEOMORPHIC                                         21
DYSTROPHIC                                                   20
LUCENT_CENTERED                                              18
ROUND_AND_REGULAR                                            17
ROUND_AND_REGULAR-LUCENT_CENTERED                            14
AMORPHOUS-PLEOMORPHIC                                        12
LARGE_RODLIKE-ROUND_AND_REGULAR                              11
PUNCTATE-AMORPHOUS                                           10
COARSE-ROUND_AND_REGULAR-LUCENT_CENTER                       10
LUCENT_CENTER-PUNCTATE                                        8
VASCULAR-COARSE-LUCENT_CENTERED                               8
ROUND_AND_REGULAR-PLEOMORPHIC                                 7
EGGSHELL                                                      7
VASCULAR-COARSE                                               6
PUNCTATE-FINE_LINEAR_BRANCHING                                6
ROUND_AND_REGULAR-PUNCTATE                                    5
```

```
LARGE_RODLIKE                                                 4
SKIN-PUNCTATE-ROUND_AND_REGULAR                               4
SKIN-PUNCTATE                                                 4
COARSE-ROUND_AND_REGULAR-LUCENT_CENTERED                      4
PUNCTATE-ROUND_AND_REGULAR                                    4
AMORPHOUS-ROUND_AND_REGULAR                                   3
PUNCTATE-LUCENT_CENTER                                        3
MILK_OF_CALCIUM                                               2
ROUND_AND_REGULAR-PUNCTATE-AMORPHOUS                          2
COARSE-ROUND_AND_REGULAR                                      2
COARSE-LUCENT_CENTER                                          2
VASCULAR-COARSE-LUCENT_CENTER-ROUND_AND_REGULAR-PUNCTATE      2
COARSE-PLEOMORPHIC                                            2
ROUND_AND_REGULAR-LUCENT_CENTER-DYSTROPHIC                    2
SKIN                                                          2
SKIN-COARSE-ROUND_AND_REGULAR                                 1
ROUND_AND_REGULAR-AMORPHOUS                                   1
PLEOMORPHIC-PLEOMORPHIC                                       1
Name: calc type, dtype: int64
---------------
Unique values of column calc distribution
CLUSTERED              740
SEGMENTAL              168
REGIONAL                99
LINEAR                  90
DIFFUSELY_SCATTERED     37
CLUSTERED-LINEAR        25
CLUSTERED-SEGMENTAL      5
LINEAR-SEGMENTAL         5
REGIONAL-REGIONAL        1
Name: calc distribution, dtype: int64
---------------
Unique values of column assessment
4    753
2    482
5    159
3     89
0     63
Name: assessment, dtype: int64
---------------
Unique values of column pathology
MALIGNANT                  544
BENIGN                     528
BENIGN_WITHOUT_CALLBACK    474
Name: pathology, dtype: int64
---------------
Unique values of column subtlety
3    502
5    361
4    346
```

```
2    242
1     95
Name: subtlety, dtype: int64
```

In [47]:
```
# Some patients have more than 1 pathology
multi_path_cal = cal_train_desc.groupby('patient_id').filter(lambda x: x
```

In [48]:
```
multi_path_cal
```

443	P_00557	2	RIGHT	CC	3	calcification	PLEOMORPHIC	CLUSTI
446	P_00557	2	RIGHT	MLO	2	calcification	PLEOMORPHIC	CLUSTI
441	P_00557	2	RIGHT	CC	1	calcification	PLEOMORPHIC	CLUSTI
440	P_00557	2	LEFT	MLO	1	calcification	PLEOMORPHIC	CLUSTI
439	P_00557	2	LEFT	CC	1	calcification	PLEOMORPHIC	CLUSTI
442	P_00557	2	RIGHT	CC	2	calcification	PLEOMORPHIC	CLUSTI
485	P_00600	3	LEFT	MLO	2	calcification	AMORPHOUS	CLUSTI
484	P_00600	3	LEFT	MLO	1	calcification	AMORPHOUS	CLUSTI
483	P_00600	3	LEFT	CC	2	calcification	AMORPHOUS	CLUSTI

In [49]:
```
multi_path_cal['patient_id'].nunique()
```

Out[49]:
```
14
```

```
In [50]: multi_path_cal.groupby('patient_id')['pathology'].nunique()
```

```
Out[50]: patient_id
         P_00418    2
         P_00467    2
         P_00557    2
         P_00600    2
         P_00858    2
         P_00937    2
         P_00992    2
         P_01156    2
         P_01200    2
         P_01276    2
         P_01284    2
         P_01409    2
         P_01582    2
         P_01819    2
         Name: pathology, dtype: int64
```

Note 1:

- There are 14 patients from Calcification train with more than 1 pathology so we can just leave these cases out.
- For these 14 patients, sometimes it is because they have biopsy for left and right breasts, each has a different pathology. Sometimes, on the same breast, some patient (e.g., P_00600) has both pathologies.

Mass Train

```
In [30]: mass_train_desc.sort_values(by = 'patient_id', inplace = True)
```

In [31]:
```
mass_train_desc.head(n=10)
```

Out[31]:

	patient_id	breast_density	left or right breast	image view	abnormality id	abnormality type	mass sha
0	P_00001	3	LEFT	CC	1	mass	IRREGULA ARCHITECTURAL_DISTORTI
1	P_00001	3	LEFT	MLO	1	mass	IRREGULA ARCHITECTURAL_DISTORTI
2	P_00004	3	LEFT	CC	1	mass	ARCHITECTURAL_DISTORTI
3	P_00004	3	LEFT	MLO	1	mass	ARCHITECTURAL_DISTORTI
4	P_00004	3	RIGHT	MLO	1	mass	OV
5	P_00009	3	RIGHT	CC	1	mass	OV
6	P_00009	3	RIGHT	MLO	1	mass	OV
7	P_00015	3	LEFT	MLO	1	mass	IRREGUL
9	P_00018	2	RIGHT	MLO	1	mass	OV
8	P_00018	2	RIGHT	CC	1	mass	OV

In [34]:
```
columns = ['image view', 'abnormality id', 'abnormality type', 'mass sha
       'subtlety']
```

In [35]:
```
for col in columns:
    print('---------------')
    print('Unique values of column {}'.format(col))
    print(mass_train_desc[col].value_counts())
```

```
---------------
Unique values of column image view
MLO    711
CC     607
Name: image view, dtype: int64
---------------
Unique values of column abnormality id
1    1216
2      68
3      23
4       7
```

```
6       2
5       2
Name: abnormality id, dtype: int64
---------------

Unique values of column abnormality type
mass    1318
Name: abnormality type, dtype: int64
---------------
Unique values of column mass shape
IRREGULAR                                  351
OVAL                                       321
LOBULATED                                  305
ROUND                                      123
ARCHITECTURAL_DISTORTION                    80
IRREGULAR-ARCHITECTURAL_DISTORTION          45
LYMPH_NODE                                  26
ASYMMETRIC_BREAST_TISSUE                    20
FOCAL_ASYMMETRIC_DENSITY                    19
OVAL-LYMPH_NODE                              6
LOBULATED-IRREGULAR                          5
LOBULATED-LYMPH_NODE                         3
ROUND-OVAL                                   3
IRREGULAR-FOCAL_ASYMMETRIC_DENSITY           2
LOBULATED-ARCHITECTURAL_DISTORTION           2
ROUND-LOBULATED                              1
ROUND-IRREGULAR-ARCHITECTURAL_DISTORTION     1
LOBULATED-OVAL                               1
Name: mass shape, dtype: int64
---------------
Unique values of column mass margins
CIRCUMSCRIBED                           305
SPICULATED                              281
ILL_DEFINED                             278
OBSCURED                                197
MICROLOBULATED                          108
CIRCUMSCRIBED-ILL_DEFINED                27
ILL_DEFINED-SPICULATED                   25
CIRCUMSCRIBED-OBSCURED                   19
OBSCURED-ILL_DEFINED                     19
OBSCURED-SPICULATED                       4
OBSCURED-ILL_DEFINED-SPICULATED           4
MICROLOBULATED-ILL_DEFINED                3
MICROLOBULATED-SPICULATED                 2
MICROLOBULATED-ILL_DEFINED-SPICULATED     2
CIRCUMSCRIBED-MICROLOBULATED              1
Name: mass margins, dtype: int64
---------------
Unique values of column assessment
4    533
```

```
5    299
3    279
0    129
2     77
1      1
Name: assessment, dtype: int64
---------------
Unique values of column pathology
MALIGNANT                  637
BENIGN                     577
BENIGN_WITHOUT_CALLBACK    104
Name: pathology, dtype: int64
---------------
Unique values of column subtlety
5    543
4    375
3    257
2    100
1     41
0      2
Name: subtlety, dtype: int64
```

In [53]:

```
multi_path_mass = mass_train_desc.groupby('patient_id').filter(lambda x:
```

In [56]:

```
multi_path_mass['patient_id'].nunique()
```

Out[56]:

```
13
```

Check if every patient has both mass and cal images

In [59]:

```
cal_patient = cal_train_desc['patient_id'].unique().tolist()
mass_patient = mass_train_desc['patient_id'].unique().tolist()
```

In [70]:

```
print('Number of patients with cal images: {}'.format(len(cal_patient)))
print('Number of patients with mass images: {}'.format(len(mass_patient)
print('Number of patients that are in one list but not other {}'.format(
print('Number of patients that are in both lists {}'.format(len(set(cal_
```

```
Number of patients with cal images: 602
Number of patients with mass images: 691
Number of patients that are in one list but not other 557
Number of patients that are in both lists 45
```

1. Abstract: Objective and main findings of the project

2. Problem motivation

3. Dataset

3.1. Data Source and description:

- Where we get the data?
- Main information of the data?

3.2. Train - Validation - Test split

- Rationale on how to get train / valid / test
- Volume of each set

3.3. Data Preprocessing

- 2D vs. 3D?
- Normalizing data
- Convert images into patches

4. Model building proces

4.1. Assumptions

4.2. Loss function

4.3. Evaluation metric

4.4. Machine Learning models: Traditional ML models. We can use the excel sheets of features related to each image and apply models such as Random Forest, SVM to classify the picture (i.e., cancerous vs. benign).

4.5. Deep learning models:

4.6. U-net

4.7. Dilated U-net

5. Experiments and Results

6. Discussion