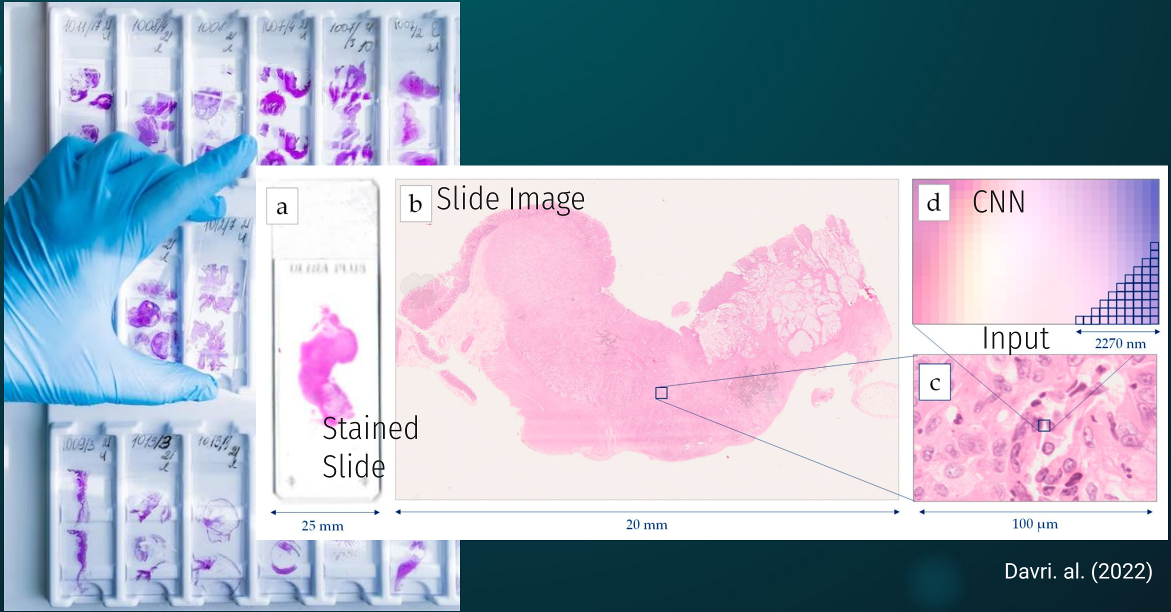
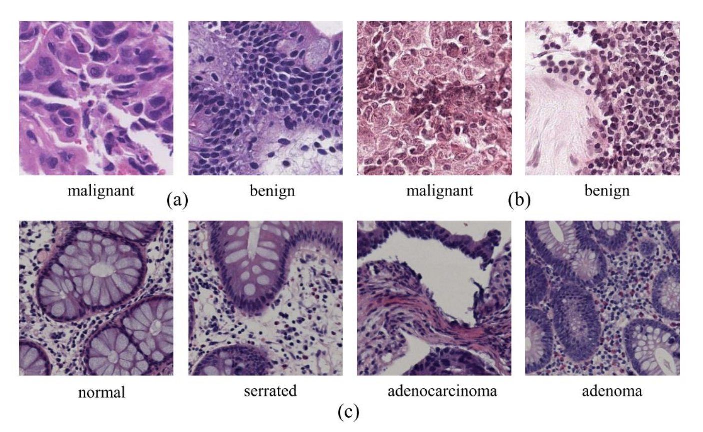
Exploring the Depths - Colon Tissue Image Classification

Sprint 2
Yukie Kuang

Sprint 1 Recap

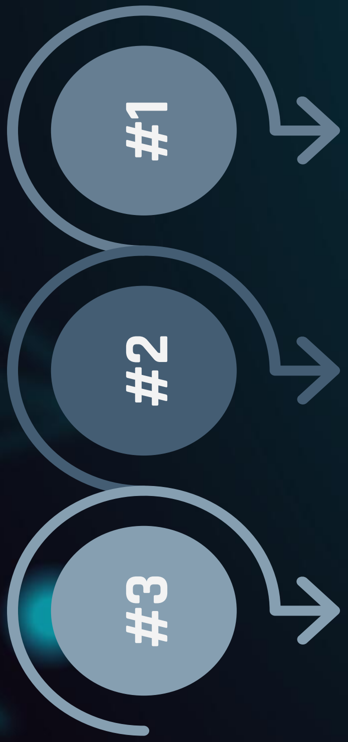


Davri. al. (2022)



Zhu et. al. (2022)

Journeying Through



Data Pre-processing

- importing image and respective annotated files
- Combining files and annotations
  - Create tuple

Image Pre-processing

- pixel value scaling
- noise detection
- data augmentation
- edge detection

Baseline Model

- Very barebones neural network

In the Thick of It - EDA

2 datasets from different sources:

- MHIST dataset is 224 x 224 - labels in .csv file
- Chaoyang dataset is 512 x 512 - labels in .json

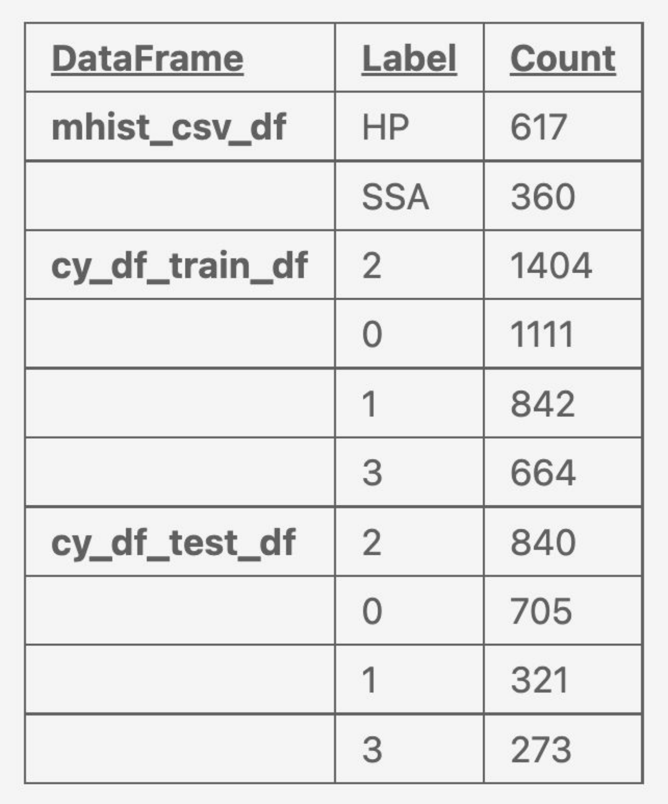
Labels:

Chaoyang dataset has 4 labels

- "0" means normal
- "1" means serrated
- "2" means adenocarinoma
- "3" means adenoma

MHIST dataset has 2 labels

- HP (Hyperplastic Polyp) aka normal polyps have no potential to become malignant
- SSA (Sessile Serated Adenoma)



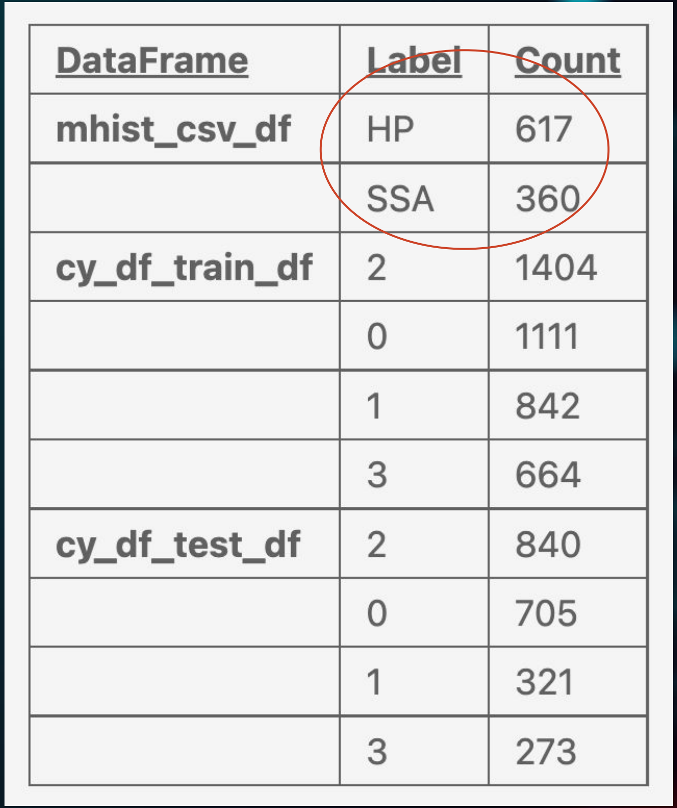
DataFrame	Label	Count
mhist_csv_df	HP	617
	SSA	360
cy_df_train_df	2	1404
	0	1111
	1	842
	3	664
cy_df_test_df	2	840
	0	705
	1	321
	3	273

In the Thick of It - EDA

Size:

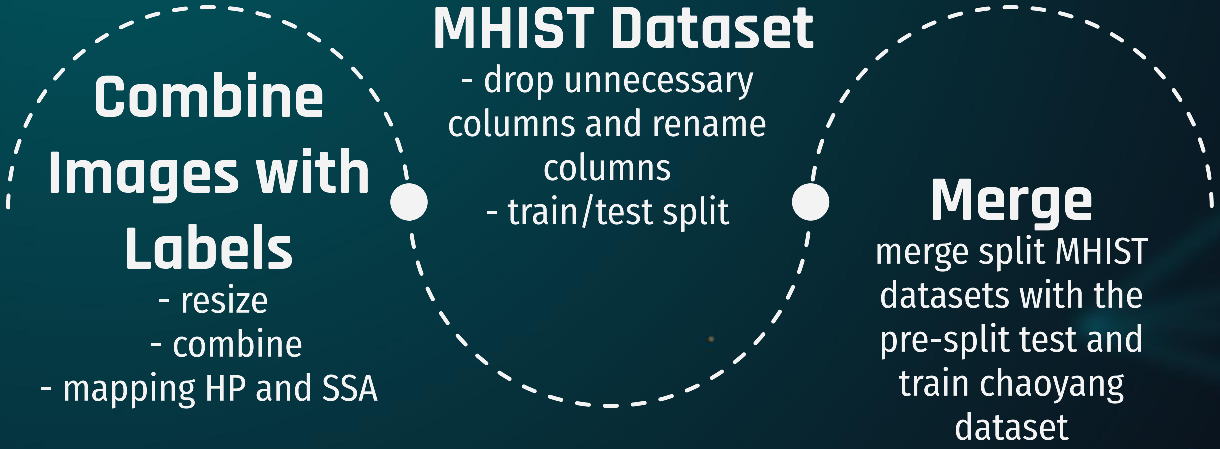
MHIST - 977*
Chaoyang - test: 2,139 train:4,021

It was thought that the MHIST dataset had provided annotations for 3,152. However, there was only one file that had annotations with only 977 entries!



DataFrame	Label	Count
mhist_csv_df	HP	617
	SSA	360
cy_df_train_df	2	1404
	0	1111
	1	842
	3	664
cy_df_test_df	2	840
	0	705
	1	321
	3	273

Data Pre-processing Steps



Combine Images with Labels

- resize
- combine
- mapping HP and SSA

MHIST Dataset

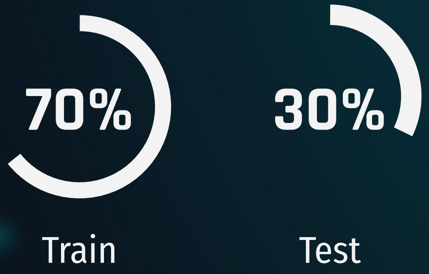
- drop unnecessary columns and rename columns
- train/test split

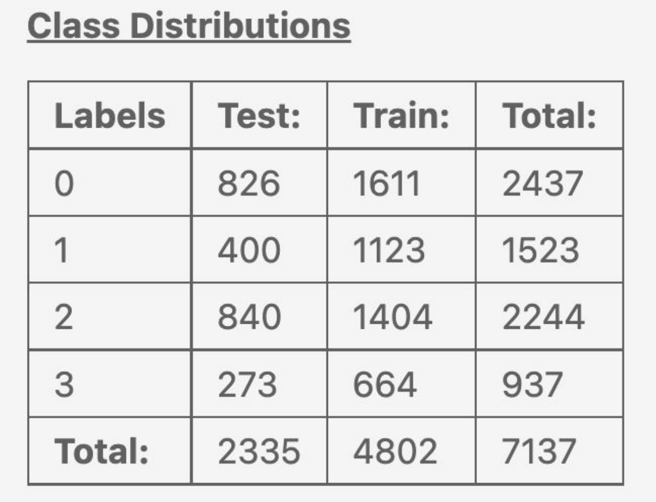
Merge

merge split MHIST datasets with the pre-split test and train chaoyang dataset

Combined Dataset

- There is around a 70/30 split





Class Distributions

Labels	Test:	Train:	Total:
0	826	1611	2437
1	400	1123	1523
2	840	1404	2244
3	273	664	937
Total:	2335	4802	7137

- The class distributions are showing label 2, 0, 1 to be the highest.

Labeled Datasets

- Chaoyang Train
- Chaoyang Test
- MHIST

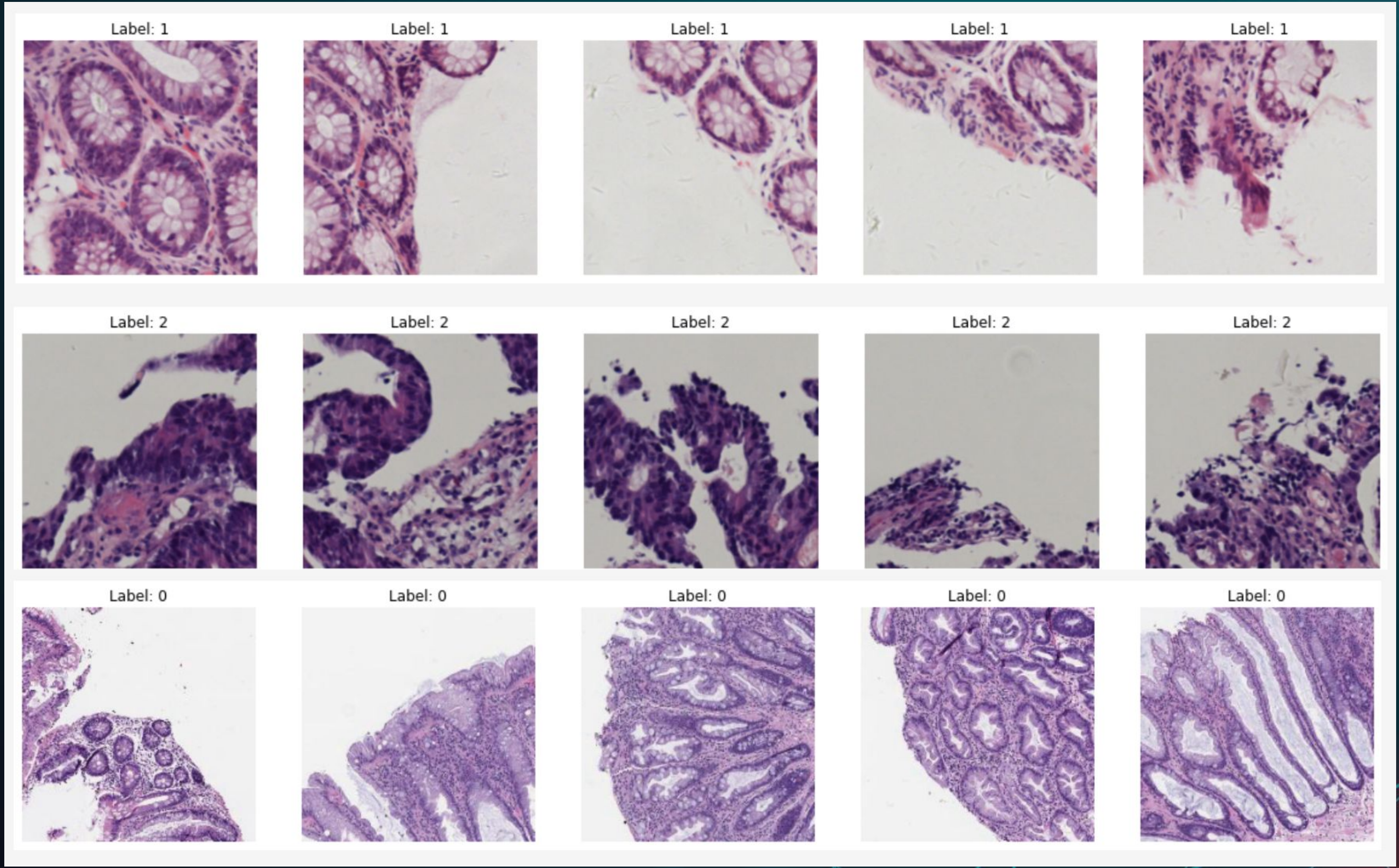
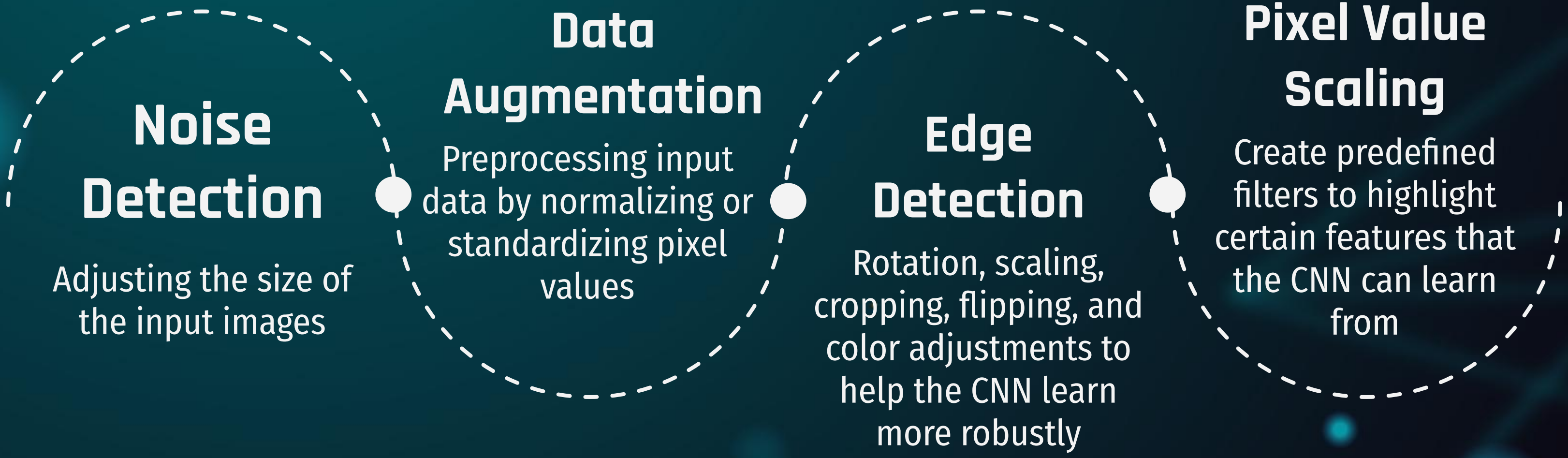


Image Pre-processing Steps



Noise Detection

Adjusting the size of the input images

Data Augmentation

Preprocessing input data by normalizing or standardizing pixel values

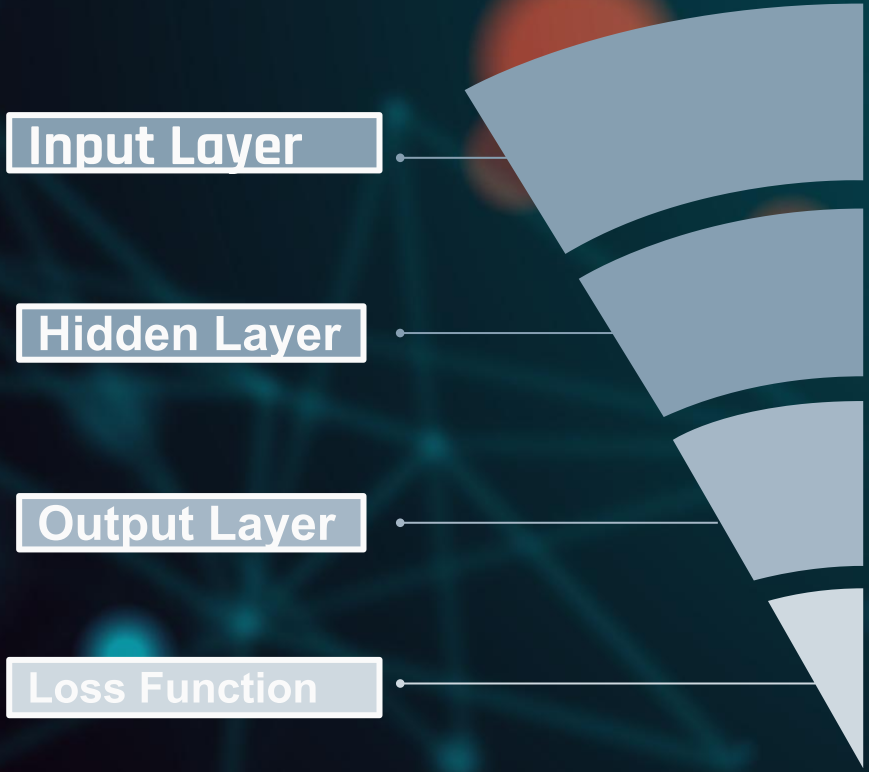
Edge Detection

Rotation, scaling, cropping, flipping, and color adjustments to help the CNN learn more robustly

Pixel Value Scaling

Create predefined filters to highlight certain features that the CNN can learn from

Baseline Model Creation - SSN



Input Layer

Door where data enters

Hidden Layer

Use special functions (like ReLU) to help the network learn complexity

Output Layer

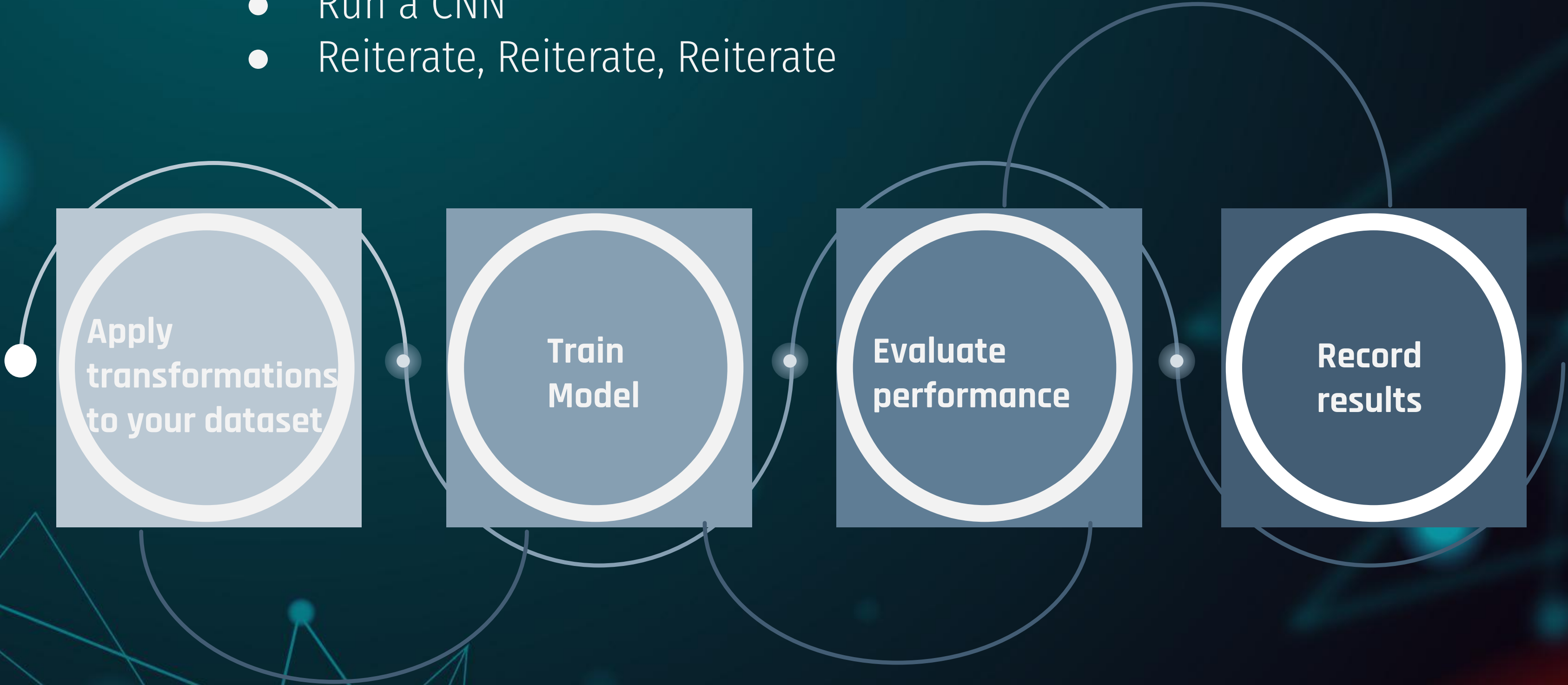
Gives you the final answer the network learned from the data

Loss Function

Scorecard - how far off predictions are from the actual answers.

Next Steps

- Complete a full run of the baseline model
- Run a CNN
- Reiterate, Reiterate, Reiterate



Thank you.

Image generated by dreamstudio.ai