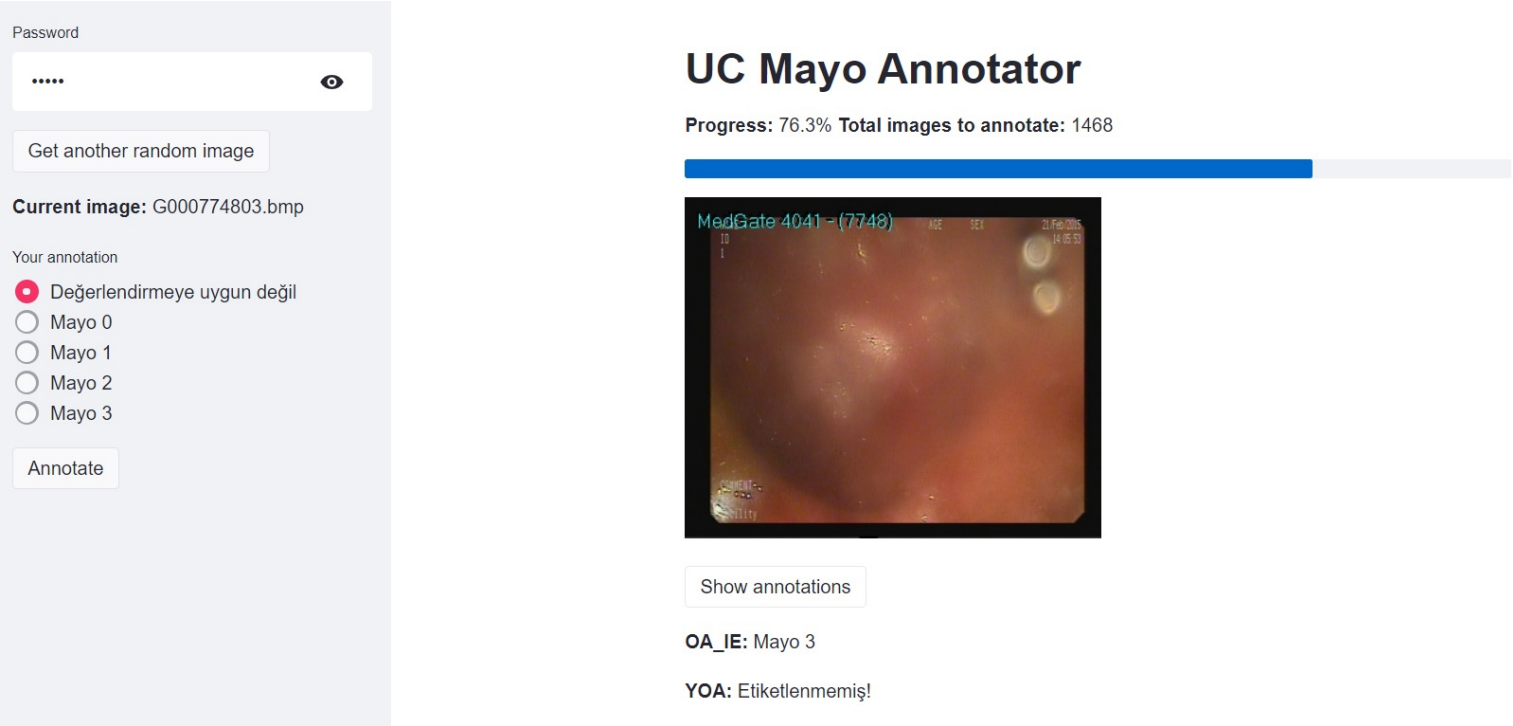
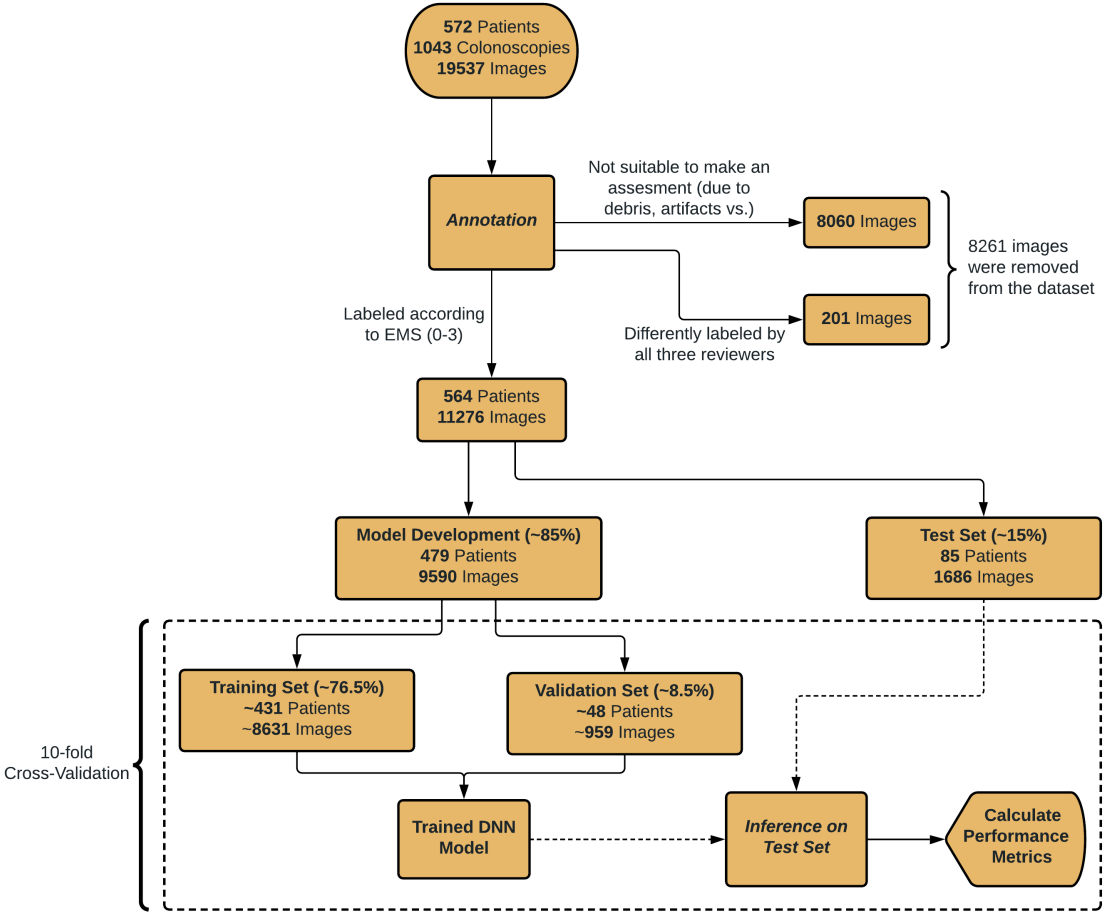
Machine Learning Systems Design

Data Labelling by Subject Matter Experts



Interface Developed by Görkem Polat



https://zenodo.org/record/5827695#.Y0UhP-xByX0

Data Labelling by Subject Matter Experts

	Reviewer-1	Reviewer-2
Total images to evaluate	19537	
Not suitable to assign a Mayo score	7621	9207
Mayo score is assigned	11916	10330
Mayo-0	7398	4503
Mayo-1	2473	3796
Mayo-2	1190	1014
Mayo-3	855	1017

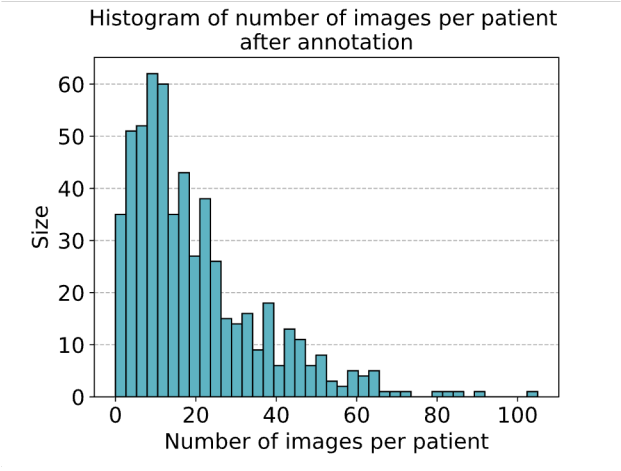
G. Polat, H.T. Kani, I. Ergenc, Y.O. Alahdab, A. Temizel, O. Atug, “Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Endoscopic Mayo Score by Using Regression-Based Deep Learning”, Inflammatory Bowel Diseases, 2022

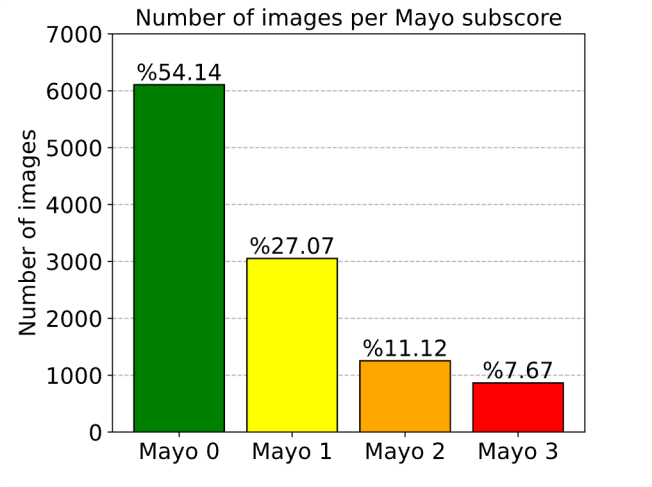
Data Labelling by Subject Matter Experts

	Reviewer-3	From Reviewer 1&2	Total
Total images to evaluate	7652	-	-
Not suitable to assign a Mayo score	1895	-	-
Mayo score is assigned	5757	-	-
All reviewers annotate differently	201	-	-
To join the final dataset (agreement by two reviewers)	5556	5720	11276
Mayo-0	2633	3472	6105
Mayo-1	1842	1210	3052
Mayo-2	784	470	1254
Mayo-3	297	568	865

G. Polat, H.T. Kani, I. Ergenc, Y.O. Alahdab, A. Temizel, O. Atug, “Improving the Computer-Aided Estimation of Ulcerative Colitis Severity According to Endoscopic Mayo Score by Using Regression-Based Deep Learning”, Inflammatory Bowel Diseases, 2022

Data Labelling by Subject Matter Experts





Data Labelling by Subject Matter Experts

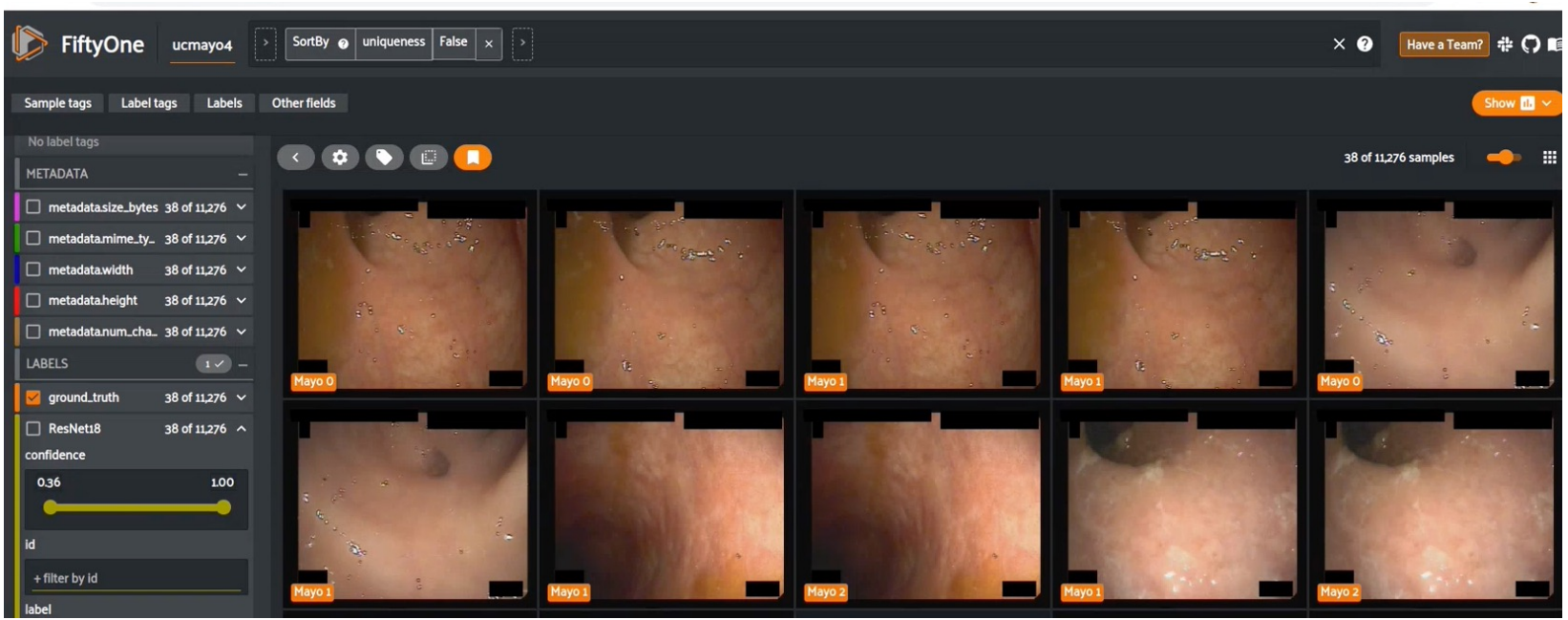


Image Credit: Oğuz Hanoğlu

Data Centric AI

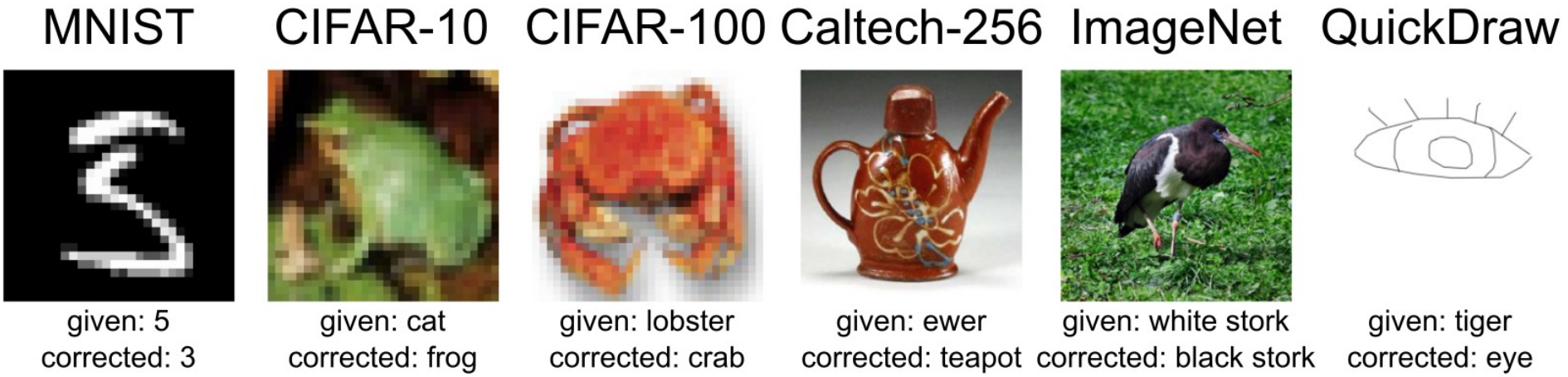
“Data-centric AI is the discipline of systematically engineering the data used to build an AI system.”
https://datacentricai.org/

Labelling Errors

- cleanlab finds issues in any dataset that a classifier can be trained on.
- works with any model by using model outputs (predicted probabilities) as input – it doesn't depend on which model created those outputs.

https://github.com/cleanlab/cleanlab

Labelling Errors



Northcutt, C., Jiang, L. and Chuang, I., 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, pp.1373-1411.

https://github.com/cleanlab/cleanlab

Labelling Errors in Public Datasets

Dataset: CIFAR-10 Label: All classes with noise

CIFAR-10 given label:
cat

Cleanlab guessed: **frog**

MTurk consensus: **frog**

ID: 2405

CIFAR-10 given label:
airplane

Cleanlab guessed: **ship**

MTurk consensus: **ship**

ID: 6877

CIFAR-10 given label:
deer

Cleanlab guessed: **horse**

MTurk consensus: **horse**

ID: 8058

CIFAR-10 given label:
airplane

Cleanlab guessed: **automobile**

MTurk consensus: **Neither airplane nor automobile**

ID: 2532

CIFAR-10 given label:
bird

Cleanlab guessed: **horse**

MTurk consensus: **horse**

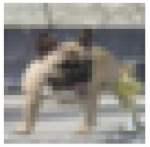
ID: 7657

CIFAR-10 given label:
automobile

Cleanlab guessed: **truck**

MTurk consensus: **Neither automobile nor truck**

ID: 1969



CIFAR-10 given label:
cat

Cleanlab guessed: **dog**

MTurk consensus: **dog**

ID: 2804

CIFAR-10 given label:
cat

Cleanlab guessed: **truck**

MTurk consensus: **Neither cat nor truck**

ID: 6792

https://labelerrors.com/

Labelling Errors in Public Datasets

Dataset: ImageNet Label: All classes with noise

ImageNet given label:
siamang

Cleanlab guessed: **baboon**

MTurk consensus: **baboon**

ID: 00047520

ImageNet given label:
red panda

Cleanlab guessed: **meerkat**

MTurk consensus: **meerkat**

ID: 00023365

ImageNet given label:
Irish Wolfhound

Cleanlab guessed: **Irish Water Spaniel**

MTurk consensus: **Irish Water Spaniel**

ID: 00022071

ImageNet given label:
macaque

Cleanlab guessed: **langur**

MTurk consensus: **langur**

ID: 00010077

ImageNet given label:
Bouvier des Flandres

Cleanlab guessed: **Siberian Husky**

MTurk consensus: **Both Bouvier des Flandres and Siberian Husky**

ImageNet given label:
tick

Cleanlab guessed: **scorpion**

MTurk consensus: **scorpion**

ImageNet given label:
Miniature Schnauzer

Cleanlab guessed: **Australian Silky Terrier**

MTurk consensus: **Australian Silky**

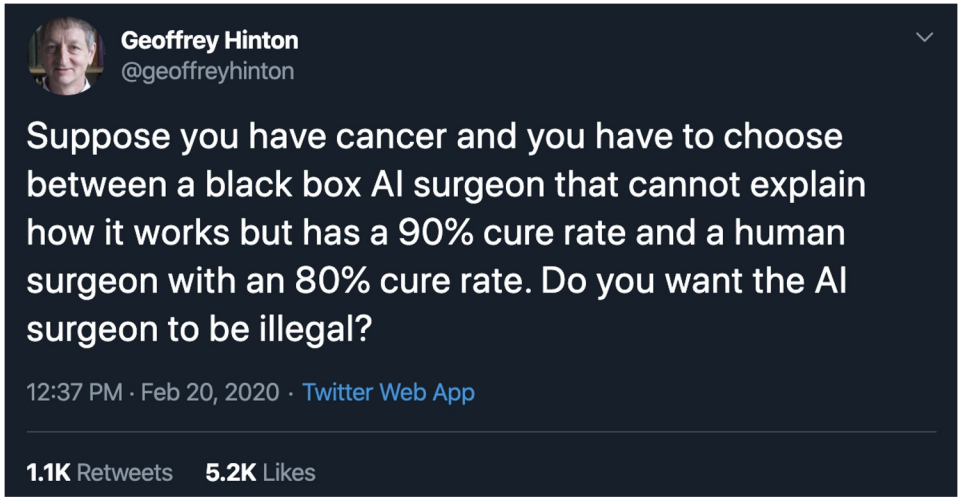
ImageNet given label:
southern black widow

Cleanlab guessed: **scorpion**

MTurk consensus: **scorpion**

https://labelerrors.com/

Explainable AI

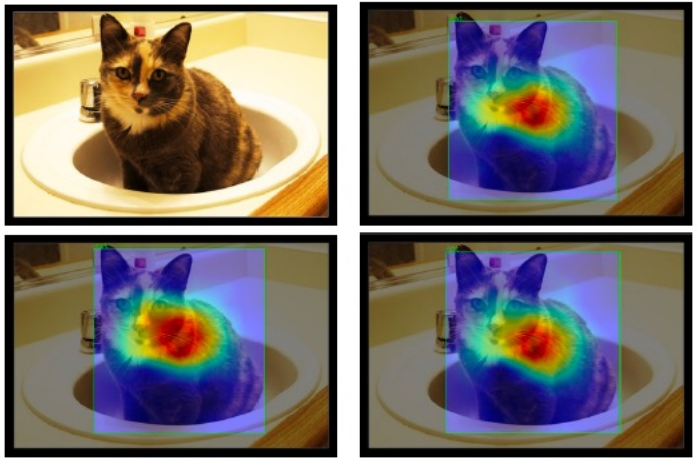


Explainable AI

When experts perform an assessment, they mainly look for the following three descriptors:

- **Vascular Pattern:**
  - How much obliteration exists in the vascular pattern?
  - Are capillaries clearly defined or is there a complete obliteration?
  - Is there an erythema in the tissue?
- **Bleeding:**
  - Is there a bleeding?
  - If yes, how prevalent and spontaneous is that?
- **Erosions and Ulcers:**
  - Is the mucosa normal?
  - If there are erosions, what are their sizes?
  - Are the ulcers superficial or deep?

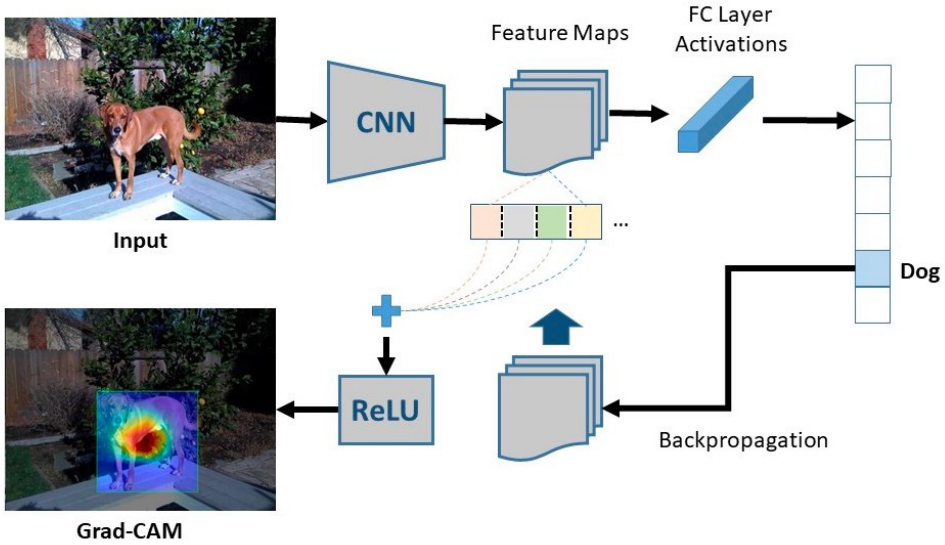
Gradient-weighted Class Activation Mapping (Grad-CAM)



- Grad-CAM uses the gradients and highlights the important regions in the image.
- Creates heatmaps using a trained network
- Also known as post-hoc attention - different from trainable attention, which entails learning specific parameters during training to build attention maps.

Image Credit: Ülkü Uzun

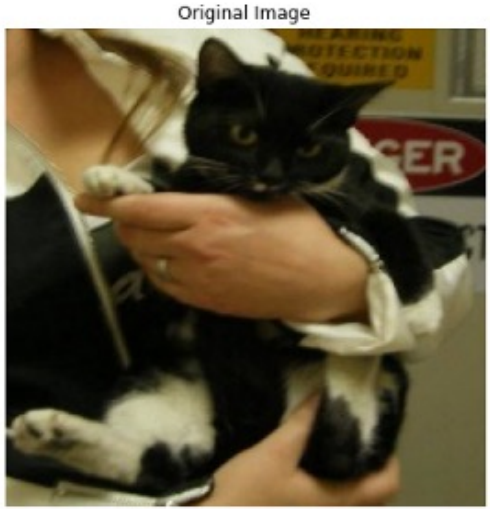
Gradient-weighted Class Activation Mapping (Grad-CAM)

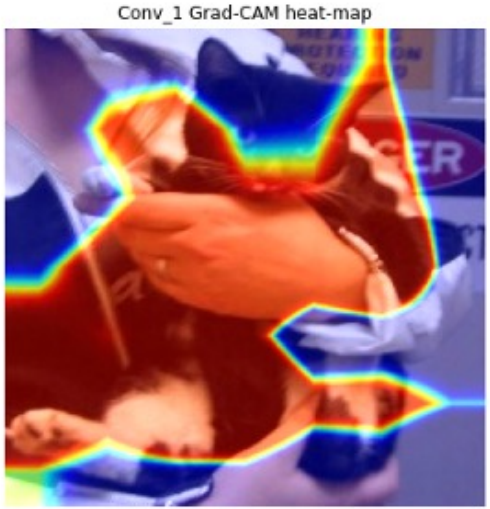


- A feature map show the location of features in the image.
- Each channel in the feature map array is multiplied by the gradient of the highest predicted class for the input image.
- The gradient is relative to the activations of the last convolution layer.
- By doing this, we increase the value of the fields that help find the class in the prediction and decrease the fields that don't contribute to the prediction.
- Then all channels are summed to get the heatmap class activation.

https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

Grad-CAM

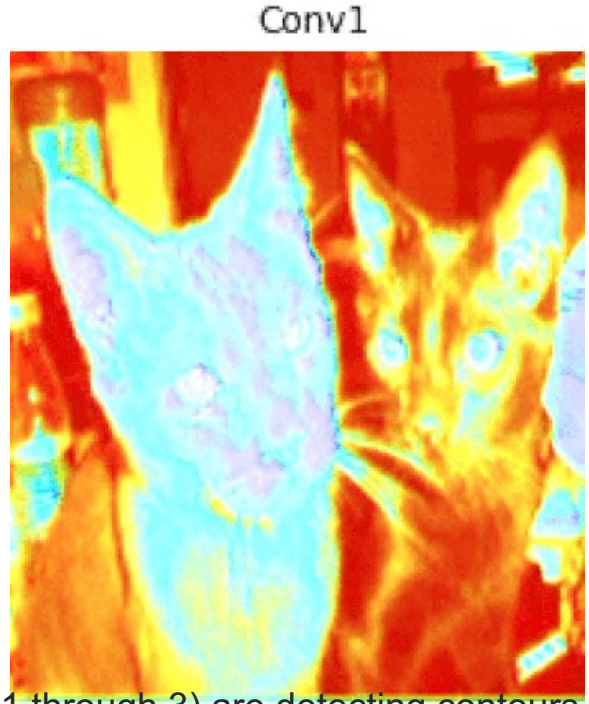




- Heat-maps for the last convolutional layer in our model, Conv_1. In theory, the heat-map for this layer should display the most accurate visual explanation of the object being classified by the model.
- In the gradient math, we are capturing the importances of all successive feature maps leading up to the last convolutional layer. We noticed that while the heat-map does emphasize the classified object, a cat in this case, it isn't very precise. The emphasized region (red) encapsulates regions of the cat but does not fit the cat with much precision.
- We know the model sees a cat, but we are not quite sure what it is about the cat that convinces the model that this is, in fact, a cat.

https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

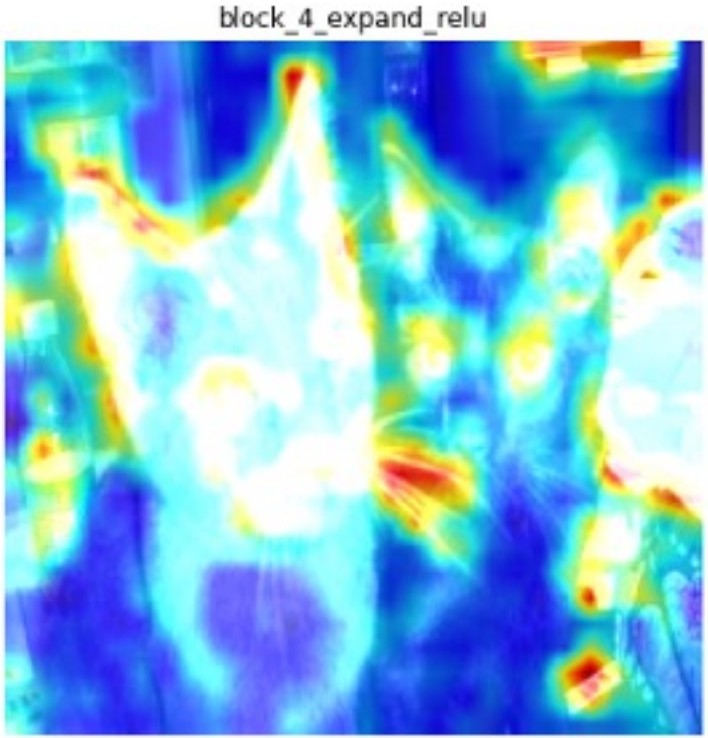
Grad-CAM



- The first ~10 layers (blocks 1 through 3) are detecting contours and borders in the image. Depthwise layers de-emphasize objects while project & expand layers de-emphasize contours. The next ~20 layers (blocks 4 through 11) are detecting concepts in the image.

https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

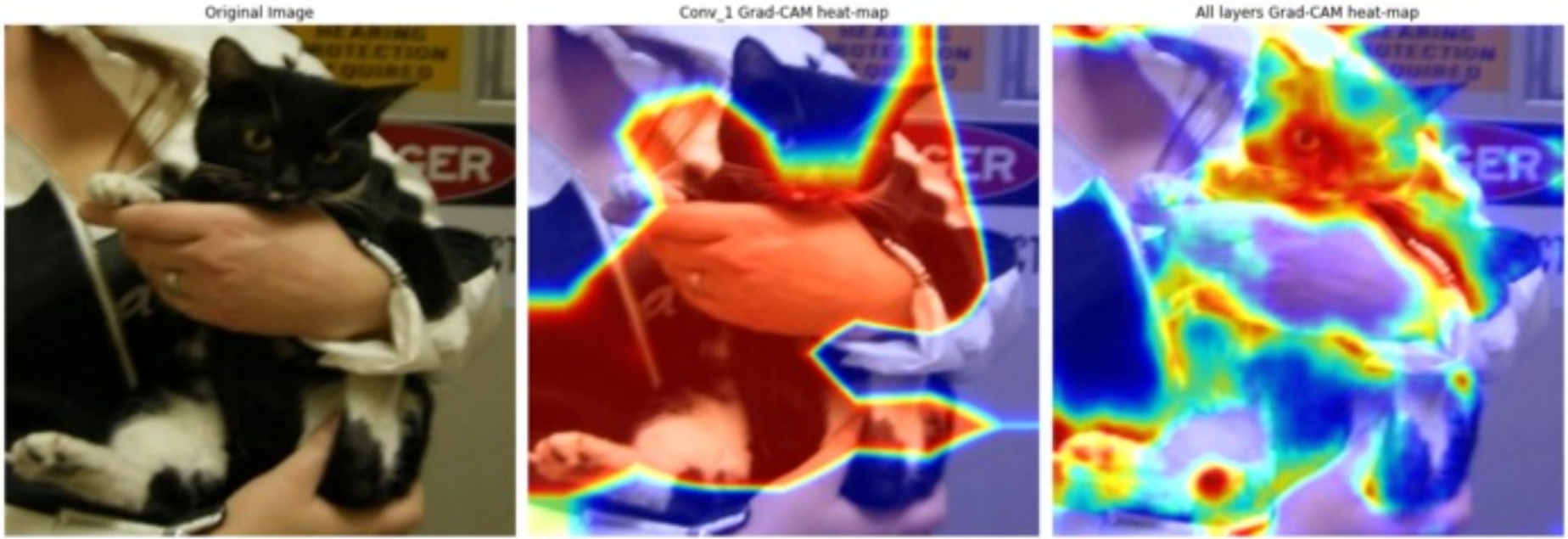
Grad-CAM



- Block 4 expand relu is a great example.

https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

Grad-CAM



- To incorporate earlier layers, we averaged together Grad-CAM heat-maps from all model layers.

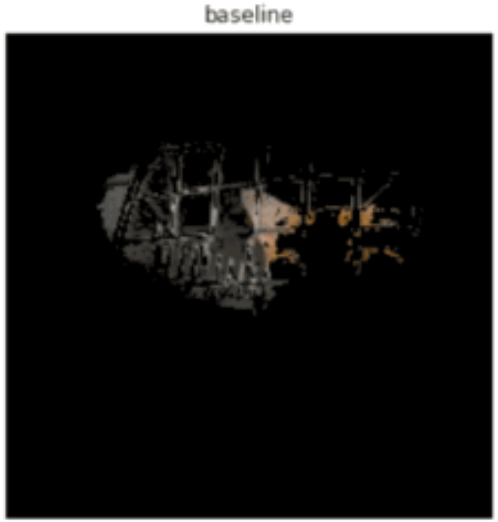
https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

Grad-CAM

The model is focusing on the transporter crate surrounding the cat rather than the cat itself and it is leading to a false classification.

https://towardsdatascience.com/understand-your-algorithm-with-grad-cam-d3b62fce353

Grad-CAM



- We created a supplemental dataset of cats & dogs in transporter crates from google images.
  Using albumentations, we beefed up the dataset with additional image augmentations.
- Armed with the data we needed, we began experimenting. We created a series of models with our original dataset and added 25%, 50%, 75% and 100% of the supplemental transporter crate dataset.
- At each iteration of our experiment, we went back to the Grad-CAM to see if the model adjusted its region of emphasis on the original image.

Explainable AI

- We have generated CAM visualizations using the technique in [11] for the ResNet18 model which is trained with both CE and CDW-CE loss (Figure 5).
- Our research question was “although CDW-CE has better performance, can it also generate *better* feature maps”.
- Better means more aligned with the experts’ decision making, i.e., better represents the symptoms on the tissue.

Explainable AI

Select ground-truth score:

Mayo-0

Analyze Random Image

Image Path

Analyze Selected Image

Activation Map

Mayo-0

Visualization mode

- Mod 1
- Mod 2
- Fixed

Blend ratio

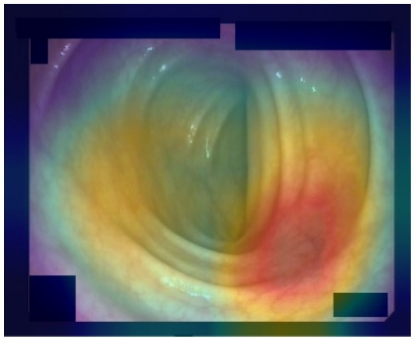
0.70

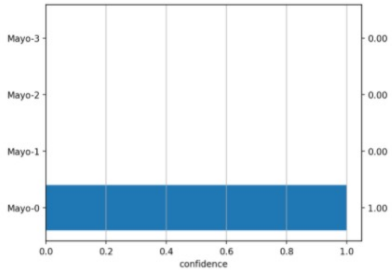
0.00 1.00

AI Mayo Score Analyzer

Image Path: Mayo 0/UC_patient_557_45.bmp

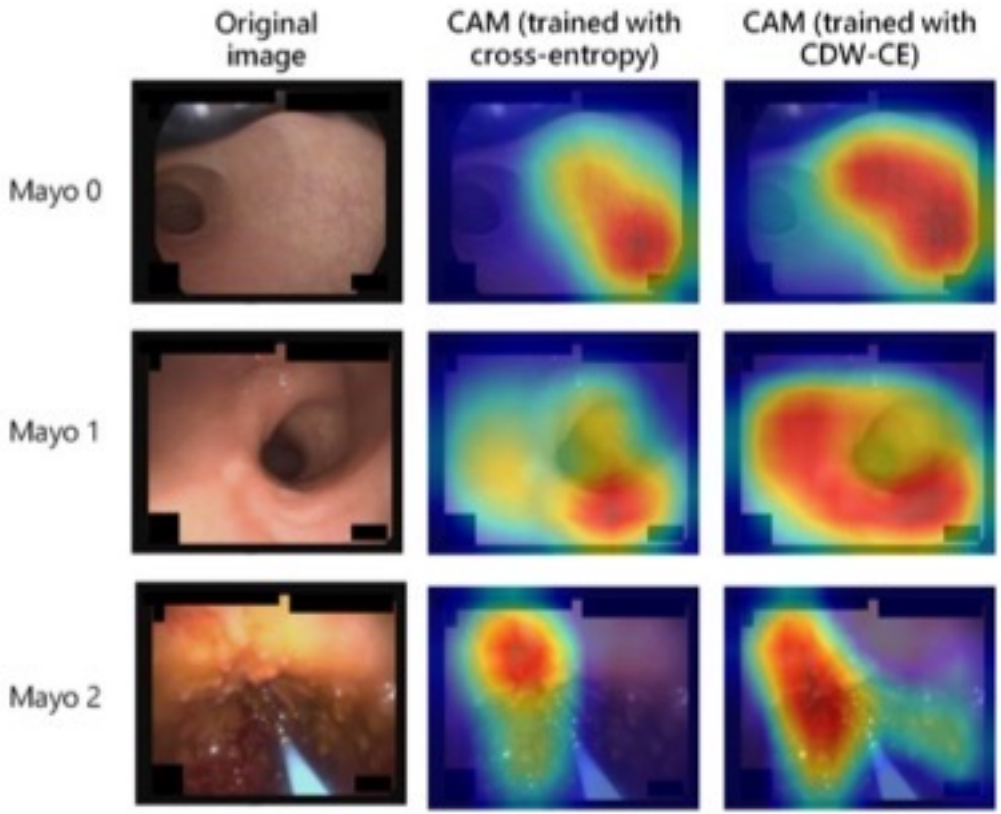
Result: ✔ **True**





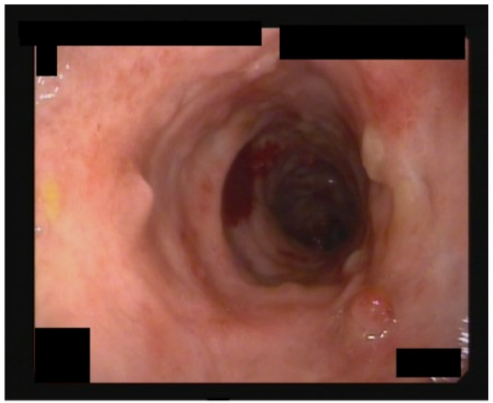
Interface Developed by Görkem Polat

Explainable AI

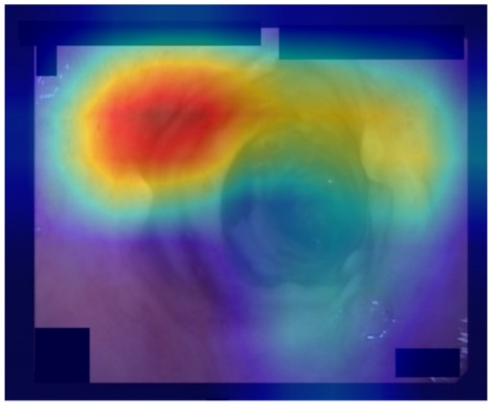


Explainable AI

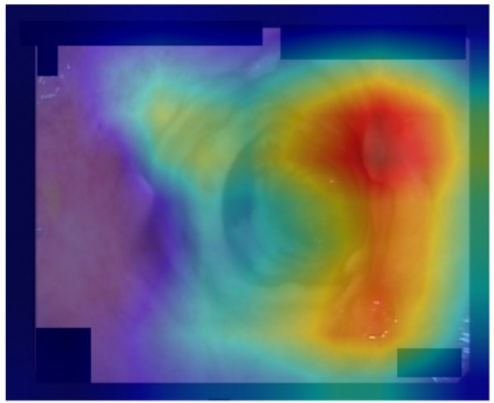
Both AI systems correctly predicted below image as **Mayo 2**.



AI-1 based it's decision on:



AI-2 based it's decision on:

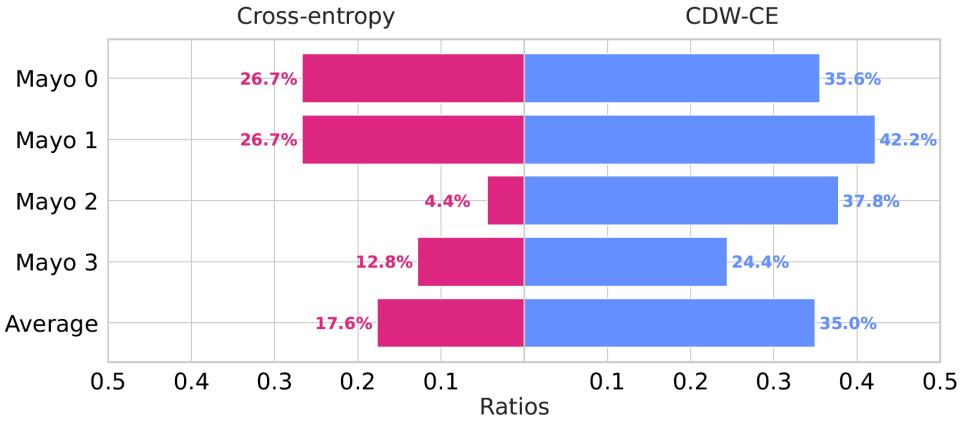


Which AI system is more reasonable?

- ○ AI-1 seems more reasonable than AI-2.
- ◉ Both AI systems seem equally reasonable.
- ○ AI-2 seems more reasonable than AI-1.

Confirm

Explainable AI



CAM comparison results. the percentage values experts found both visualizations equal are as follows: 37.7%, 31.1%, 57.8%, 62.8%, 47.4%, respectively.

Oguzhan (Ouz) Gencoglu • 2nd

Co-founder & Head of AI @ Top Data Science

2yr •

Honestly, best data scientists I have worked with had the following mindset: "It's not my fault but it's still my problem."

- Client data is messy? Not your fault.
- Paper you found relevant did not release their code? Not your fault.
- Colleague left the project without proper documentation? Not your fault.
- Now the model predictions are off because the world changed in unforeseeable ways? Not your fault.

But it is still YOUR problem!

Top experts accept the situation real quick, sit down & make a plan and communicate openly in a timely manner.

They are excellent at focusing only on the things that are in their control and serenely welcoming literally everything else. That's why they are comfortable finding themselves in uncomfortable situations. And that means A LOT in data science.