Exploratory Analysis of Cancerous vs. Non Cancerous Lungs

A Classification Problem

Team 8:
Reem Fashho, Amanda Nowacki, Shreya Shukla

https://www.npr.org/sections/health-shots/2015/04/13/398101515/why-some-doctors-are-hesitant-to-screen-smokers-for-lung-cancer

Introduction

- Lung Cancer is the **2nd most common form of cancer** in the United States
  - Leading cause of death from cancer

- Different Forms of Lung Cancer
  - Lung Nodules
  - **Non - Small Cell Lung Cancer (Most Common)**
  - Small Cell Lung Cancer
  - Mesothelioma (rare)

- Accurate assessment of disease state is critical for treatment approach

- Computed Tomography **(CT) scan is gold standard for lung cancer imaging**

Motivation

To aid radiologists in **detecting cancerous lung tissues** to **reduce the mortality rate** of Lung Cancer in the United States.

This will be achieved by developing computer - aided diagnostic (CAD) models that can **output a “second opinion” to complement physician diagnosis and treatment decisions.**

Data

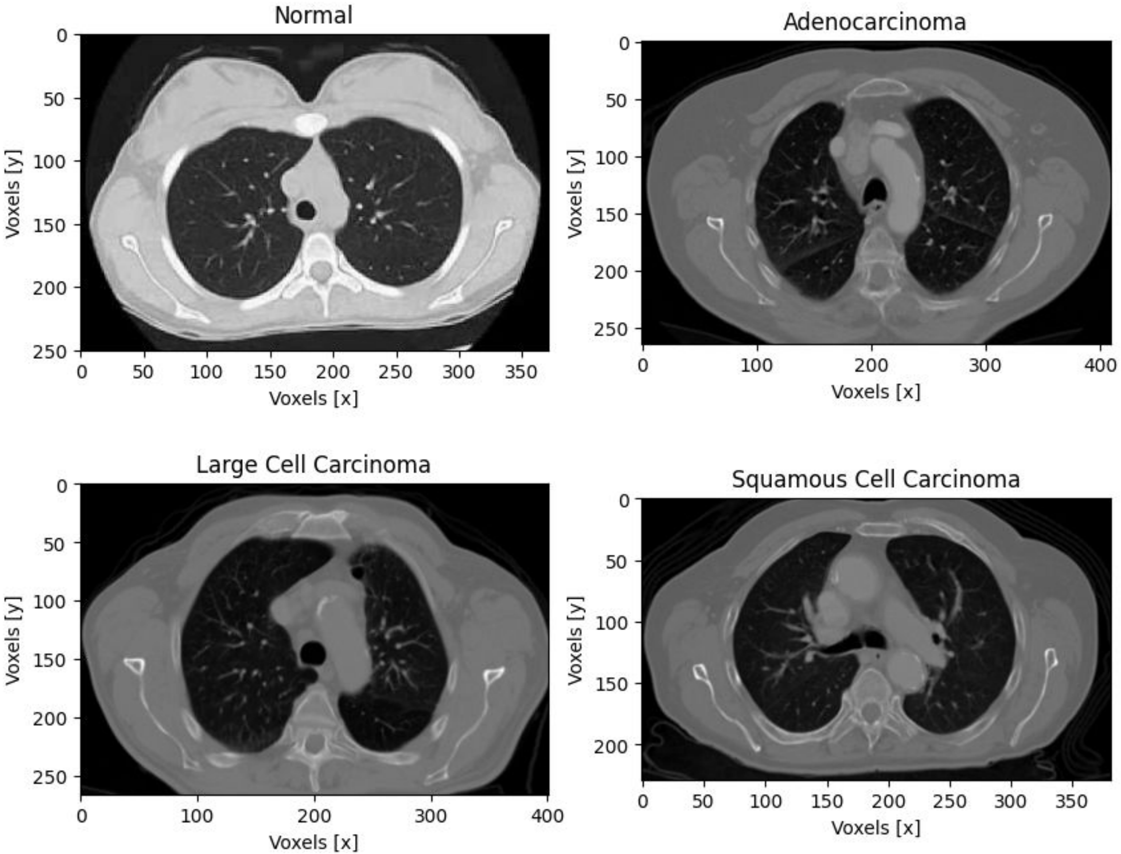
- **1000 Images of Lung CT Scans from Kaggle**
  - File Types: .jpg, .png

- Pre - Split Into: **70% Train, 20% Test, 10% Validation**

- 3 Non - Small Cell Lung Cancer Types:
  - Adenocarcinoma
  - Large Cell Carcinoma
  - Squamous Cell Carcinoma

- Limitations:
  - Small Set
  - Lacks Demographics Data
    - No variables other than image label, thus inhibiting extensive exploratory analysis

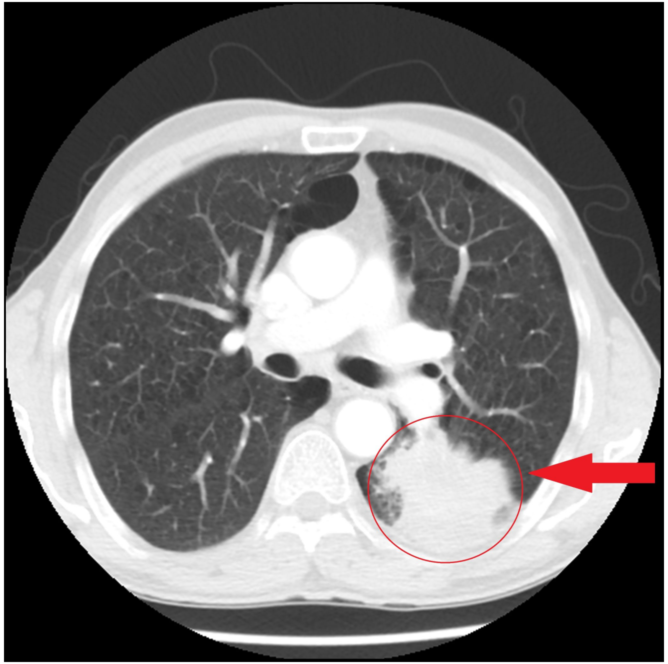
Data Image Visualization



Cancer Data Type 1

Adenocarcinoma

- ❖ Most Common Lung Cancer Type in the USA
- ❖ Strong Association with previous smoking
  - ➢ Yet, it's the most common form for nonsmokers
- ❖ Originates from the mucosal glands (hence the suffix adeno)
- ❖ Characterized by:
  - ➢ Chronic Inflammation
  - ➢ Scaring
  - ➢ Usually occurs in the periphery



https://www.wikidoc.org/index.php/Adenocarcinoma_of_the_lung_CT

Cancer Data Type 2

Large Cell Carcinoma (LCLC)

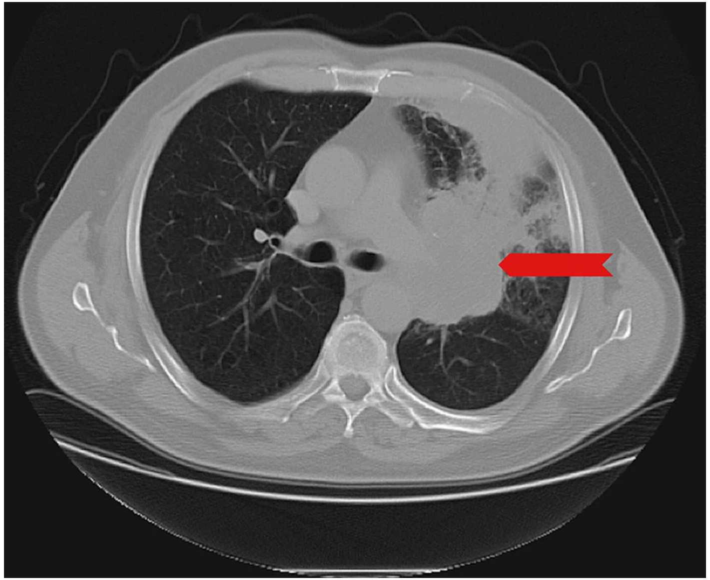
- ❖ Rapid Growth
- ❖ Can lead to fluid accumulation in chest cavity
- ❖ Characterized by:
  - ➢ Large Abnormal Cells
  - ➢ Usually occurs in the outer edge

https://radiopaedia.org/articles/large-cell-neuroendocrine-carcinoma-of-the-lung?lang=us

Cancer Data Type 3

Squamous Cell Carcinoma

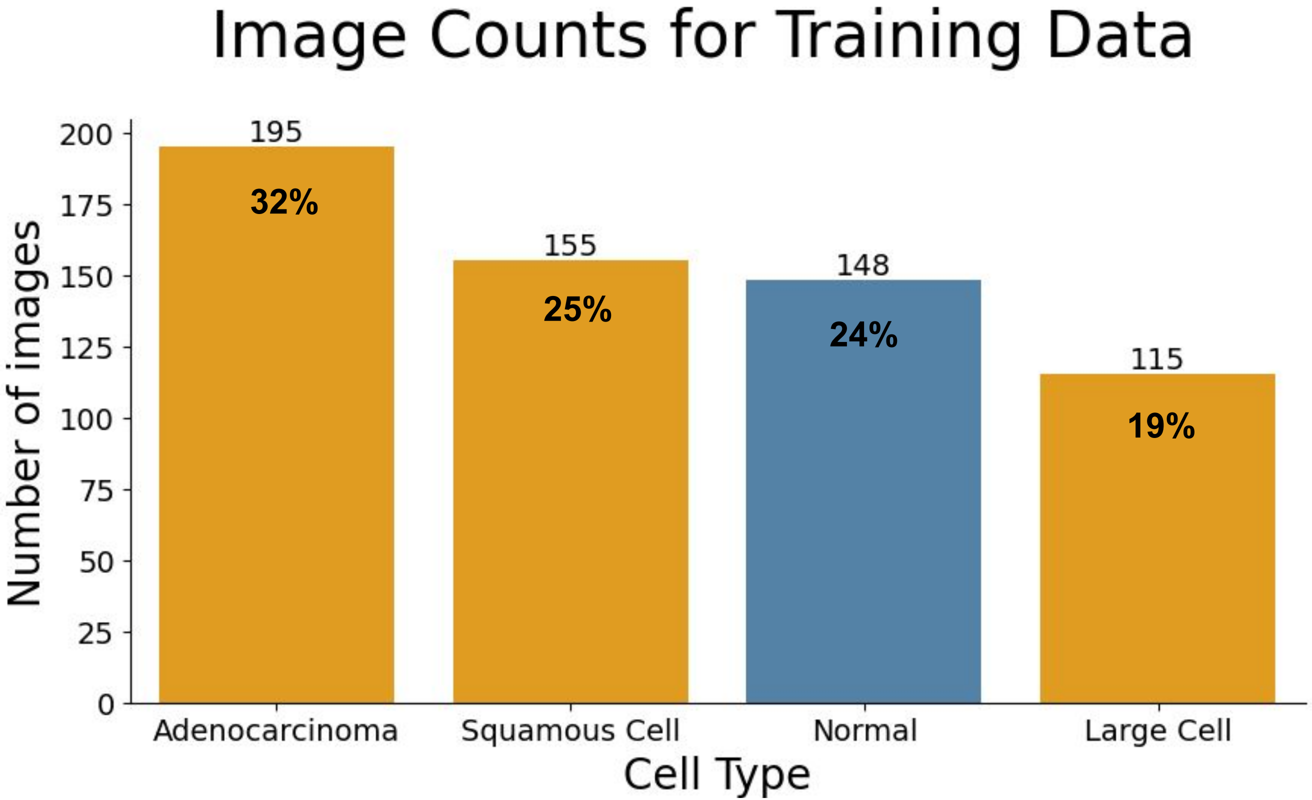
- ❖ Generally Linked to Smokers
- ❖ Slow Growing
- ❖ Develops on airways near the left/right bronchus
- ❖ Characterized by:
  - ➢ Found Centrally in the lung



https://www.cureus.com/articles

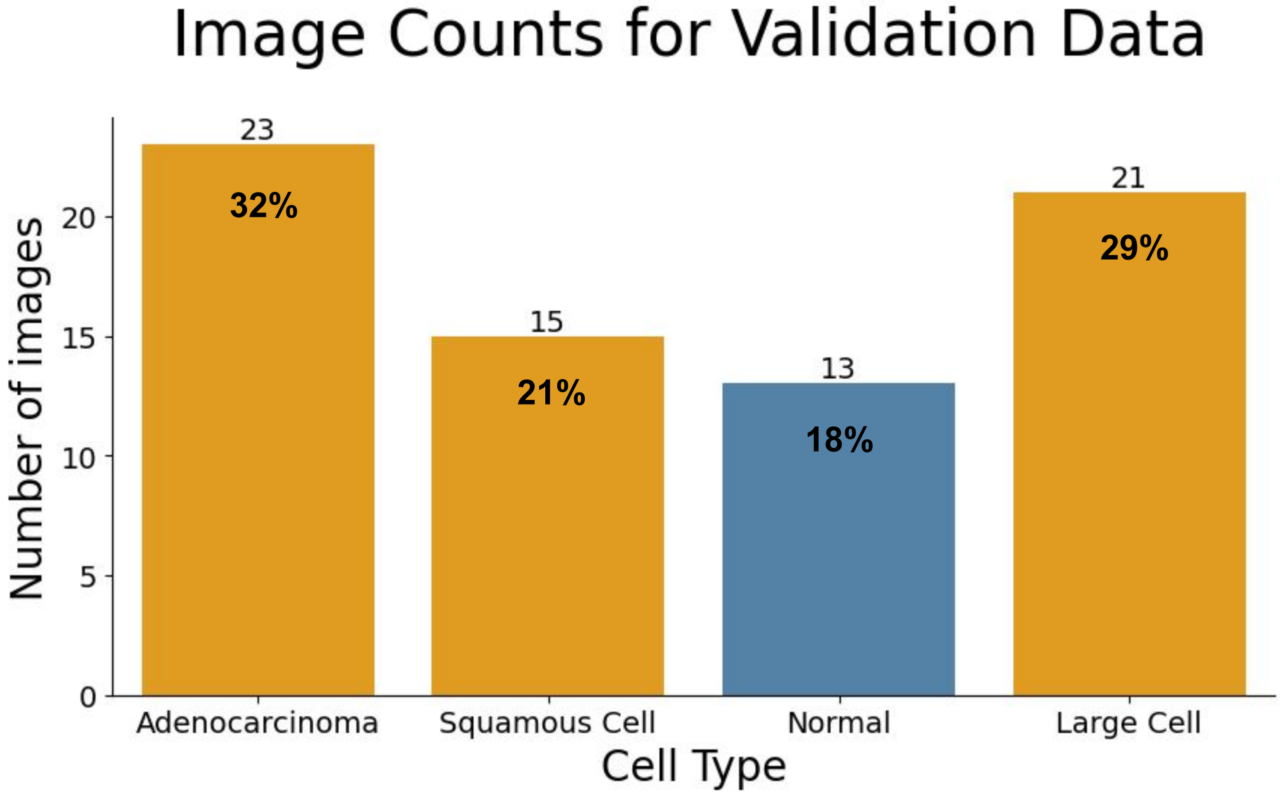
Data Distribution

613 Images



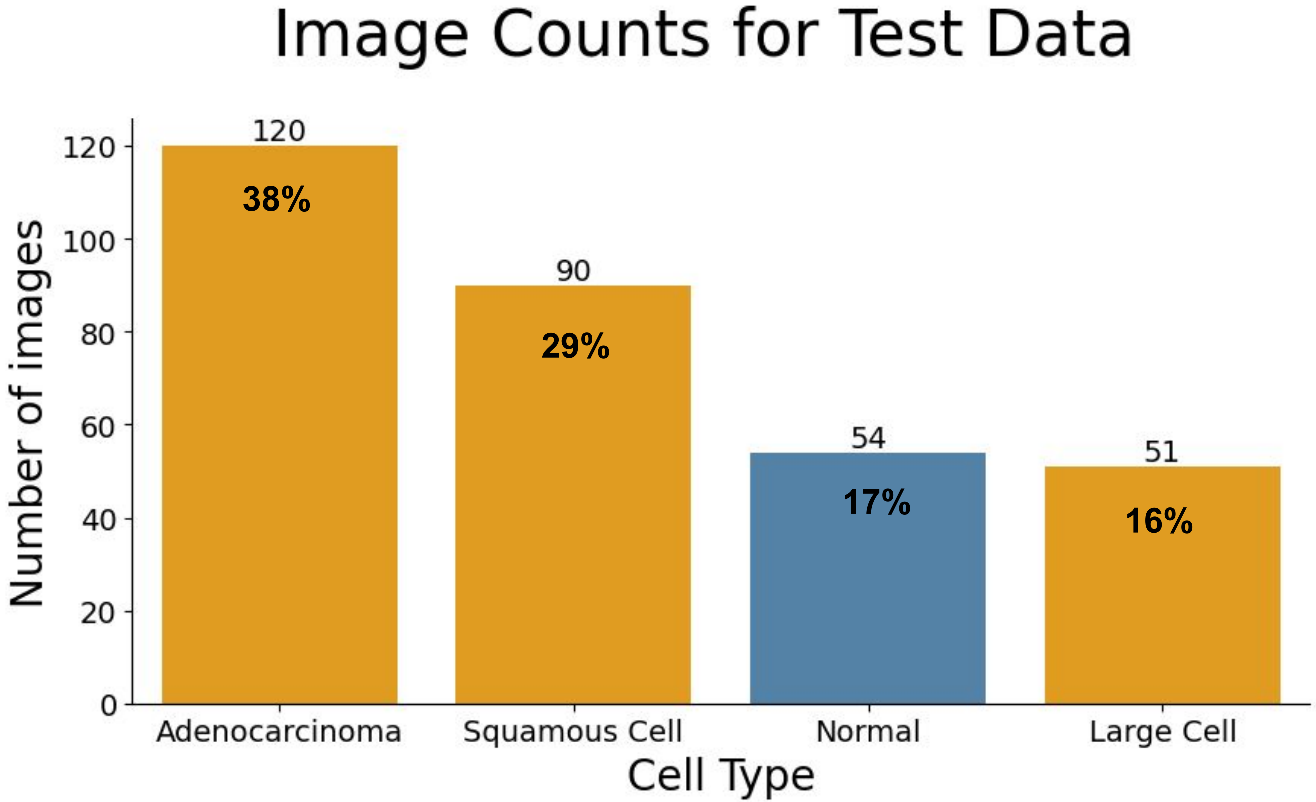
Data Distribution

72 Images



Data Distribution

315 Images



Problem Set Up

Aim

The models differentiate between cancerous and noncancerous lungs

Hypothesis

Can the models correctly locate the cancerous spots in the images?

Experimentation

How does adjusting the CT Scans images affect the model's classification (cancer/noncancer) performance?

Models

2D CNN	ResNet50	DenseNet201	VGG16
❖ 7 Layers : 2 Conv2D 2 MaxPooling2D 2 Dropout 1 Dense layer ❖ 70/20/20 split for training, testing, and validation ❖ 460 x 460 Image Size ❖ 10 epochs	❖ 50 Layers ❖ 70/20/20 split for training, testing, and validation ❖ 460 x 460 Image Size ❖ Weights from ImageNet dataset ❖ 10 epochs	❖ 201 layers ❖ 70/20/20 split for training, testing, and validation ❖ 460 x 460 Image Size ❖ Weights and biases from ImageNet ❖ 10 epochs	❖ 16 layers ❖ 70/20/20 split for training, testing, and validation ❖ 460 x 460 Image Size ❖ Weights and biases from ImageNet ❖ 10 epochs

Models Accuracy on Unaugmented Test Data

2D CNN	ResNet50	DenseNet201	VGG16
loss: 21.3331 acc: 0.8453	loss: 0.1494 acc: 0.9964	loss: 0.0616 acc: 0.9928	loss: 0.6152 acc: 0.9532

LIME: Local Interpretable Model-Agnostic Explanations

Model Interpretability

Explains model predictions so the user can understand the underlying mechanisms of the black box technique

- Select a target instance for which we want to explain the prediction
- Generate a set of perturbed instances by making small changes to the features of the chose instance
- Evaluate the ML model on the set of perturbed instances
- Train an **interpretable model**, such as a linear model, on the **perturbed instances and their corresponding predictions.**
- Use the interpretable model to explain the prediction for the target instance by identifying the features that are most important for the prediction.

Model Explainability - LIME - Cancerous

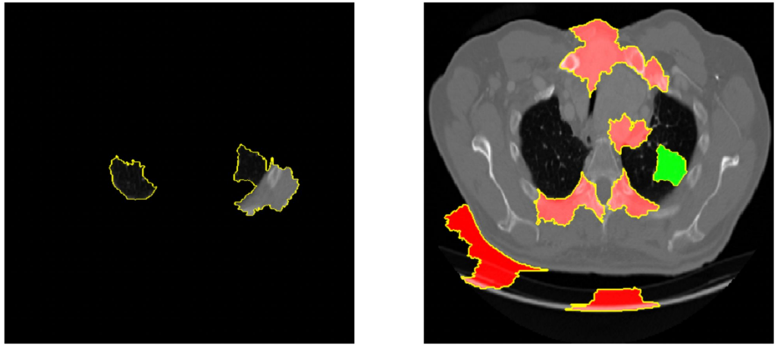
green are the features that positively contribute to the prediction

red are the features that negatively contribute to the prediction of the label

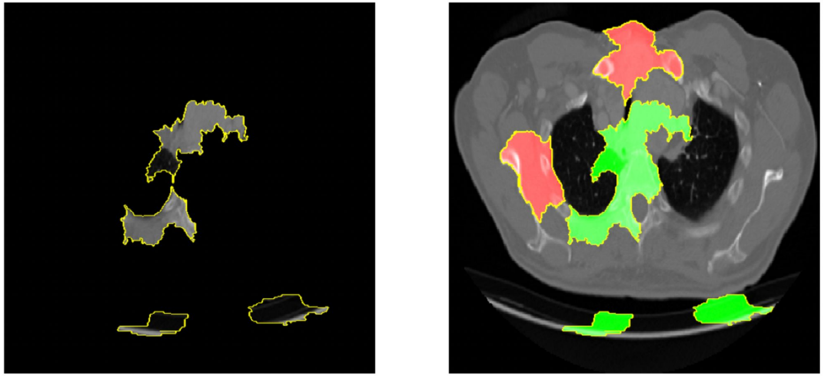
Model Explainability - LIME - Cancerous

green are the features that positively contribute to the prediction

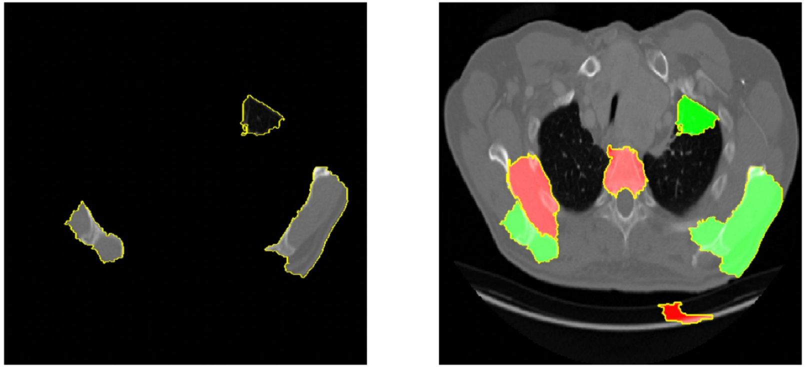
red are the features that negatively contribute to the prediction of the label



VGG 16



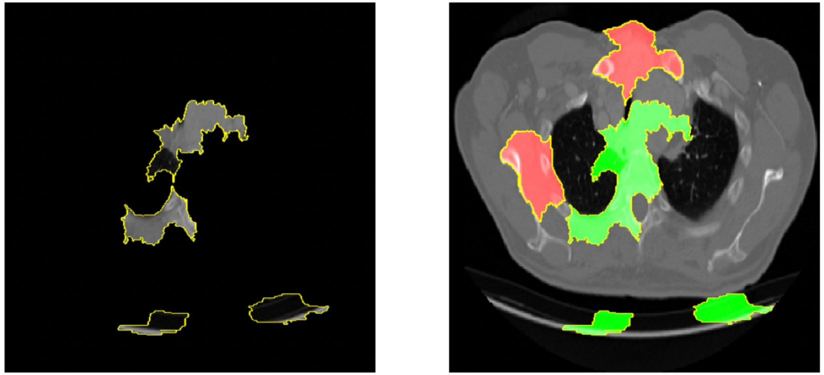
ResNet50



DenseNet201

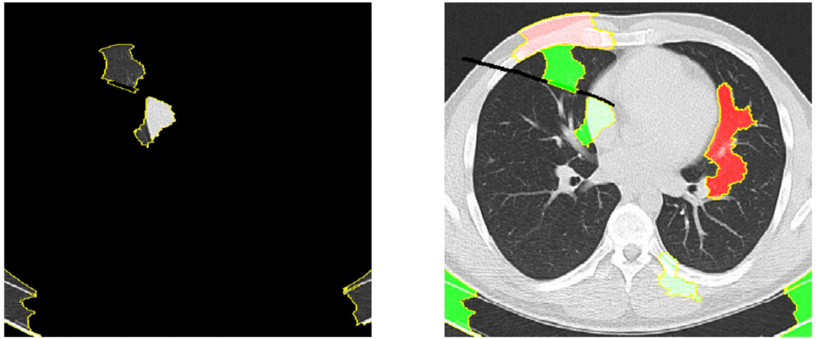
Model Explainability - LIME - Cancerous

green are the features that positively contribute to the prediction
red are the features that negatively contribute to the prediction of the label

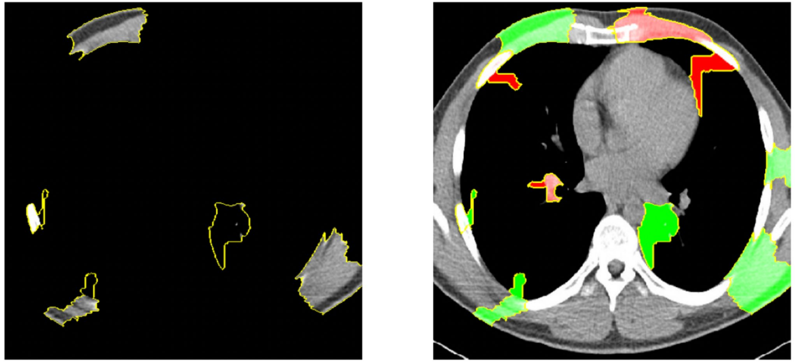


ResNet 50

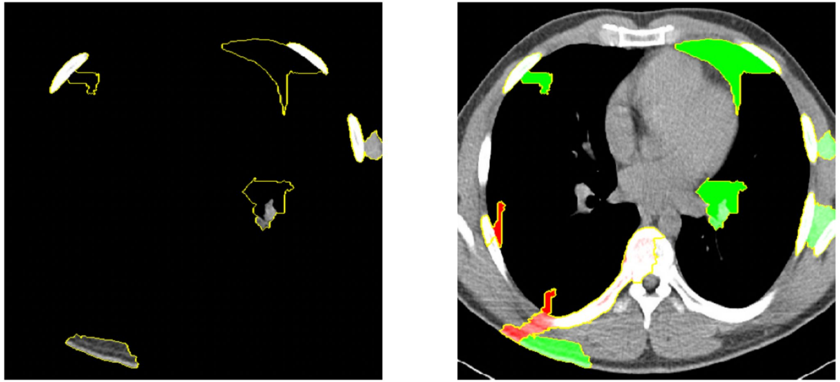
Model Explainability - LIME - Non Cancerous



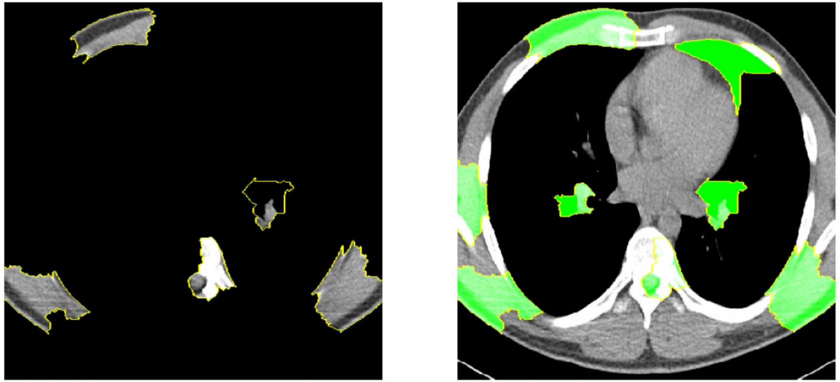
Baseline 2d CNN



VGG 16



ResNet50



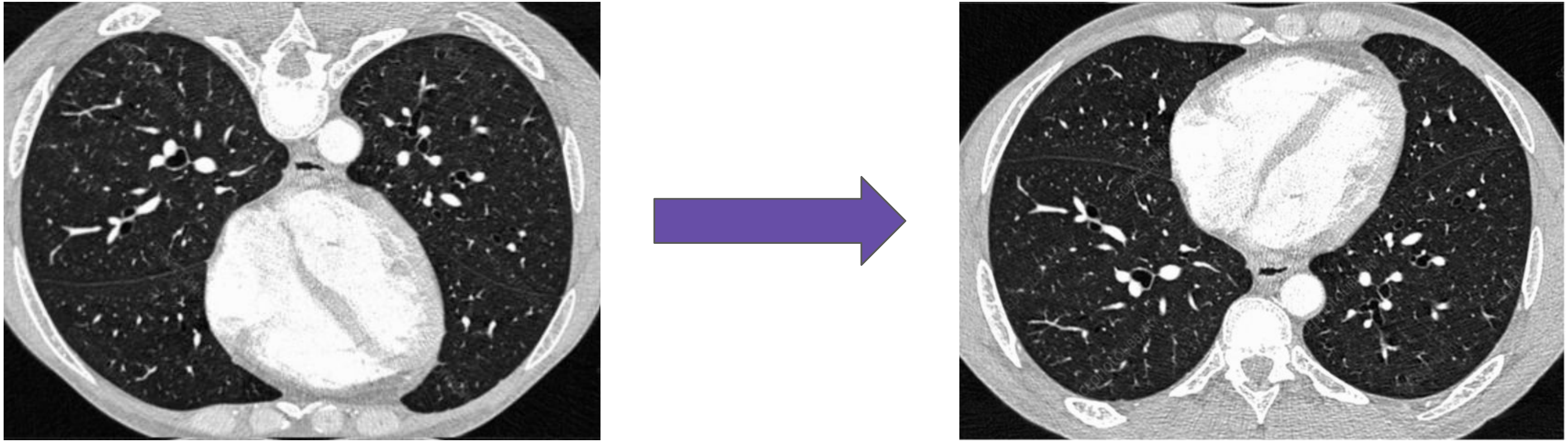
DenseNet201

Results After Augmentation

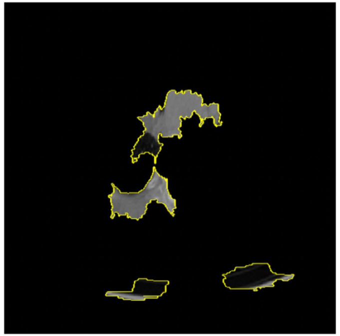
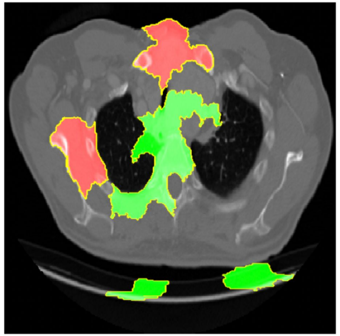
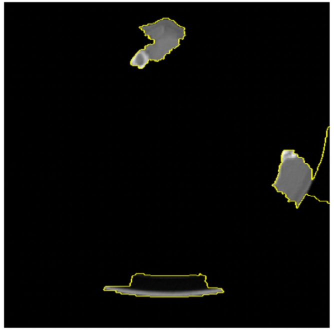
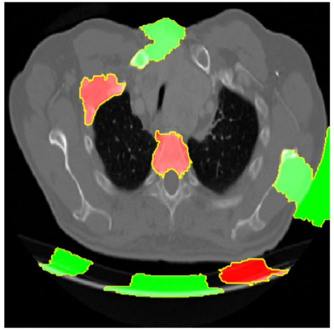
Data Augmentation

- Technique to artificially increase the size of the training set by creating or modifying copies of the original dataset
  Images - cropping, rotating, distortion, color distortions, blurring

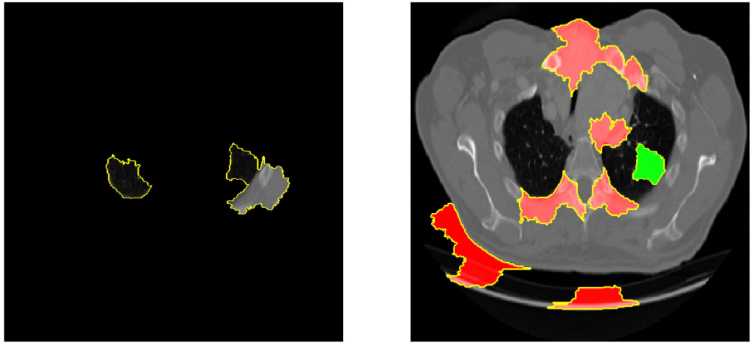
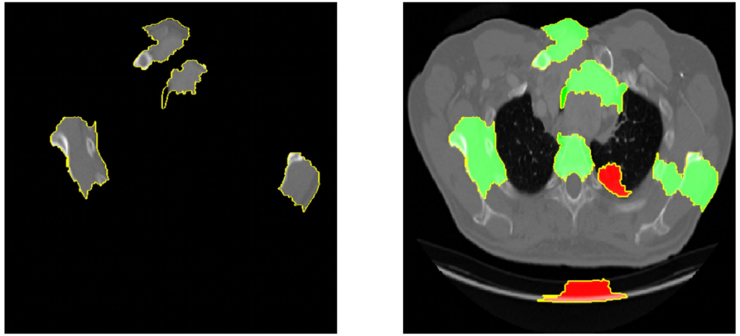
  Augmentation Applied to the CT Scans:
  - 25 Pixel Crop
  - Vertical flip of 50% of images
  - Gaussian Blur of Images



ResNet-50 Augmentation

Without Augmentation	With Augmentation
loss: 0.1494 - **acc: 0.9964**	loss: 0.1429 - **acc: 0.9964**
 	 

VGG-16 Augmentation

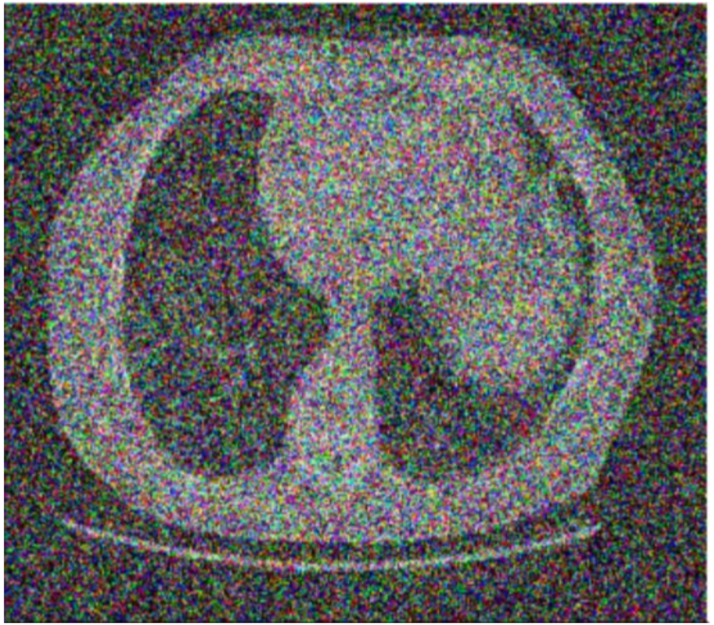
Without Augmentation	**With Augmentation**
loss: 0.6152 - **acc: 0.9532**	loss: 1.2348 - **acc: 0.2824**
	

Gradient Based Performance

CT Scan Manipulation

Original Adenocarcinoma Image

Gaussian Noise



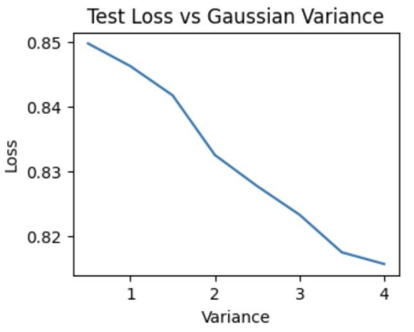
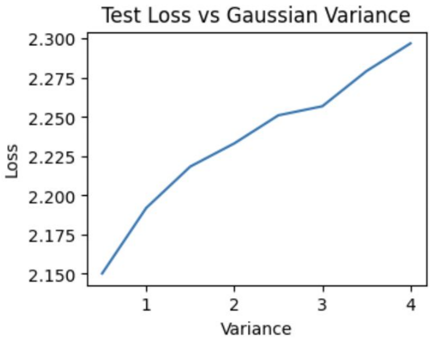
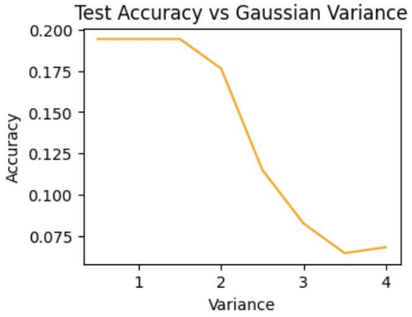
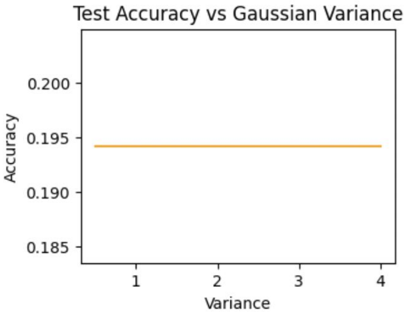
Noisy Adenocarcinoma Image

Gaussian Noise Variance

variances = [0.50, 1.00, 1.50, 2.00, 2.50, 3.00, 3.50, 4.00]

Gaussian Noise Variance

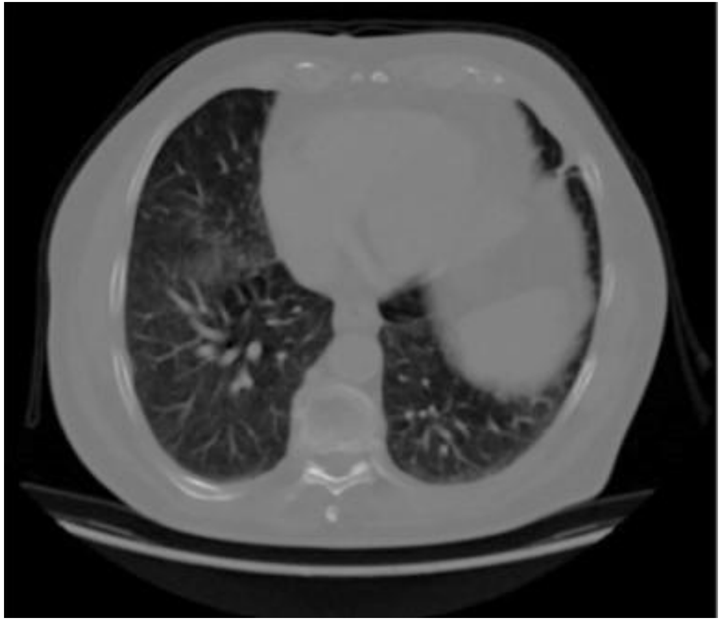
variances = [0.50, 1.00, 1.50, 2.00, 2.50, 3.00, 3.50, 4.00]

VGG -16 Unaugmented	ResNet-50 Augmented
	
	

CT Scan Manipulation

Original Adenocarcinoma Image

Contrast Reduction



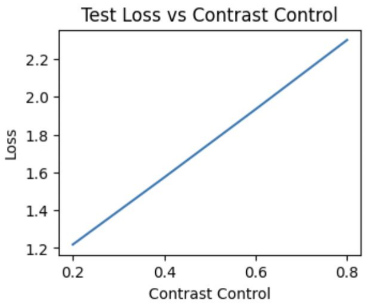
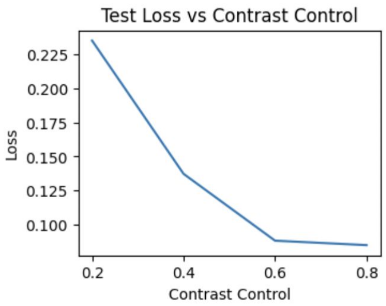
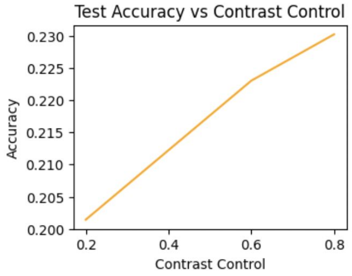
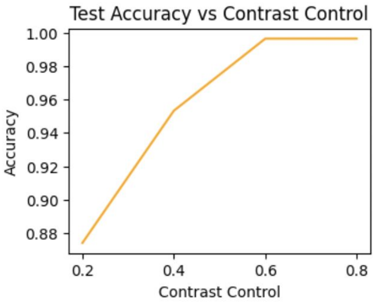
Lower Contrast Adenocarcinoma Image

Contrast Reduction

contrast_control = [0.2, 0.4, 0.6, 0.8]

Contrast Reduction

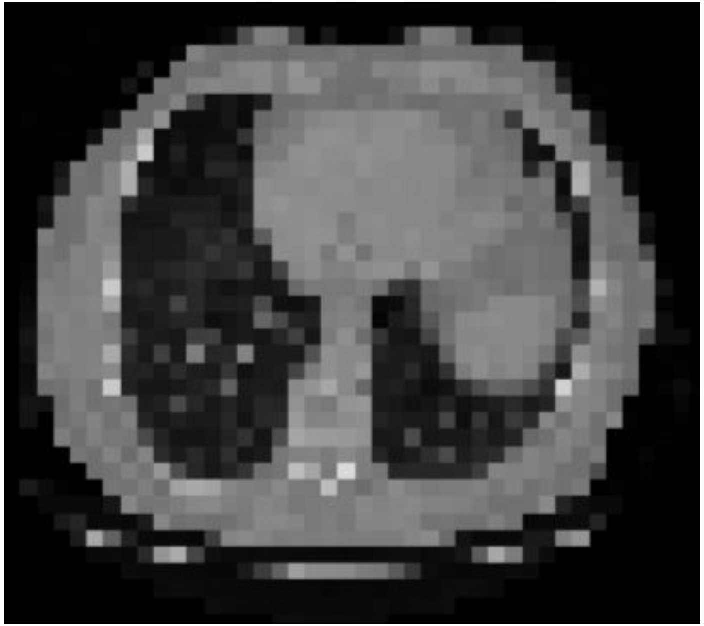
contrast_control = [0.2, 0.4, 0.6, 0.8]

VGG -16 Unaugmented	ResNet-50 Augmented
	
	

CT Scan Manipulation

Original Adenocarcinoma Image

Down Sampling



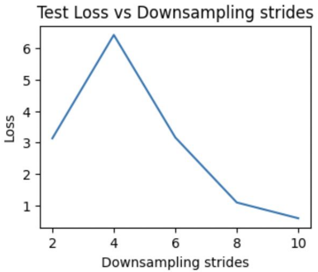
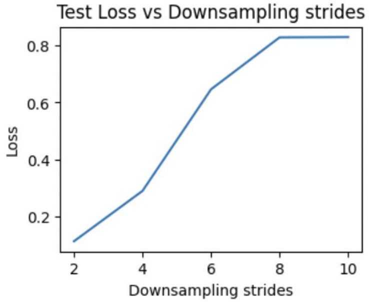
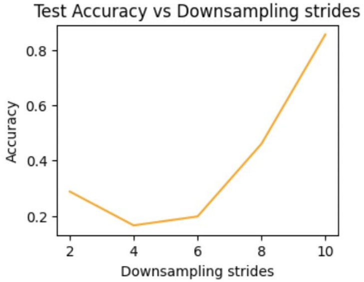
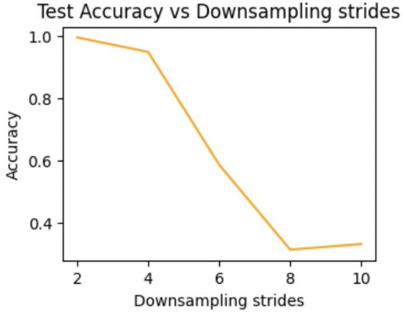
Undersampled Adenocarcinoma Image

Downsampling

strides = [2, 4, 6, 8, 10]

Downsampling

strides = [2, 4, 6, 8, 10]

VGG -16 Unaugmented	ResNet-50 Augmented
	
	

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

VGG-16 accuracy dropped significantly for augmented data, but LIME performed pretty well on the cancerous image

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

VGG-16 accuracy dropped significantly for augmented data, but LIME performed pretty well on the cancerous image

Performance Gradient for Unaugmented VGG16 and Augmented ResNet50.

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

VGG-16 accuracy dropped significantly for augmented data, but LIME performed pretty well on the cancerous image

Performance Gradient for Unaugmented VGG16 and Augmented ResNet50.

Gaussian Noise - Accuracy dropped significantly and remained same for different variances

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

VGG-16 accuracy dropped significantly for augmented data, but LIME performed pretty well on the cancerous image

Performance Gradient for Unaugmented VGG16 and Augmented ResNet50.

Gaussian Noise - Accuracy dropped significantly and remained almost same for different variances

Contrast (between 0 and 1) - VGG (decreased significantly) ResNet50 (stable), as value reaches 1, accuracy increases as expected

Summary & Conclusion

Limitation - Dataset was small

4 models - ResNet50 performed the best based on test accuracy and LIME.

Augmented data - To decrease the the possibility of memorizing data, we tested on augmented data

ResNet50 accuracy remained stable but didn’t perform well on LIME (LIME may not be best explainability model for this data or there can be some other version of lime that we can use)

VGG-16 accuracy dropped significantly for augmented data, but LIME performed pretty well on the cancerous image

Performance Gradient for Unaugmented VGG16 and Augmented ResNet50.

Gaussian Noise - Accuracy dropped significantly and remained almost same for different variances

Contrast (between 0 and 1) - VGG (decreased significantly) ResNet50 (stable), as value reaches 1, accuracy increases as expected

Down Sampling - VGG (unexpected increase in accuracy as stride increases) ResNet (more stable, accuracy decreases as stride increases)

Team Contribution

Student	Contribution
Reem Fashho	100
Amanda Nowacki	100
Shreya Shukla	100

Saliency Map

- Saliency Map of an image in the region in which a human's sight focuses initially.

- Main goal - highlight the importance of a particular pixel to the human visual perception.
  - Is the model using the correct information to classify the CT Scans?

- Brightness is directly proportional to the saliency of an image.

https://www.researchgate.net/figure/Some-examples-of-saliency-maps