Explainable AI

XAI

Laércio Pioli Junior

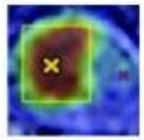
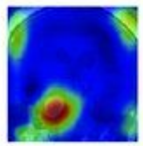
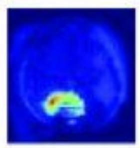
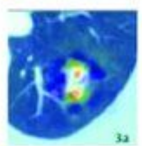
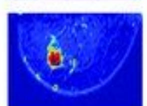
Sumário

1 Post-Hoc explanation techniques for cancer classification
2 State of the Art Taxonomy
3 XAI- State of the art

- Feature-oriented Methods
  - SHAP
  - Class Activation Maps (CAM)
  - Gradient-weighted class activation mapping (Grad-CAM)
- Global Methods
  - Global Attribution Mappings (GAMs) (2019)
  - Gradient-based Saliency Maps (2013)
  - Deep Attribute Maps (2018),
- Concept Models
  - Concept Activation Vectors (CAVs) - (2021)
- Surrogate Models
  - Local interpretable model-agnostic explanations (LIME) (2020)
- Local pixed-based Methods
  - Layer-wise relevance propagation (LRP) (2015)
- Human-centric Methods
  - Angelov and Soares (2020)

Sumário

1 Post-Hoc explanation techniques for cancer classification

Method	Results	Perturbation or Gradient	Local or Global	Model-Agnostic or Model-Specific
SHAP		Perturbation	Both	Model-Agnostic
CAM		Gradient	Local	Model-Specific
Grad-CAM		Gradient	Local	Model-Specific
PG-CAM		Gradient	Local	Model-Specific
Occlusion		Perturbation	Local	Model-Specific
Saliency Map		Gradient	Local	Model-Agnostic
Integrated Gradients		Gradient	Local	Model-Agnostic

3.1.4. Prostate Cancer

Fig 1: Post-Hoc explanation techniques for cancer classification [4]

2- State of the Art - Taxonomy

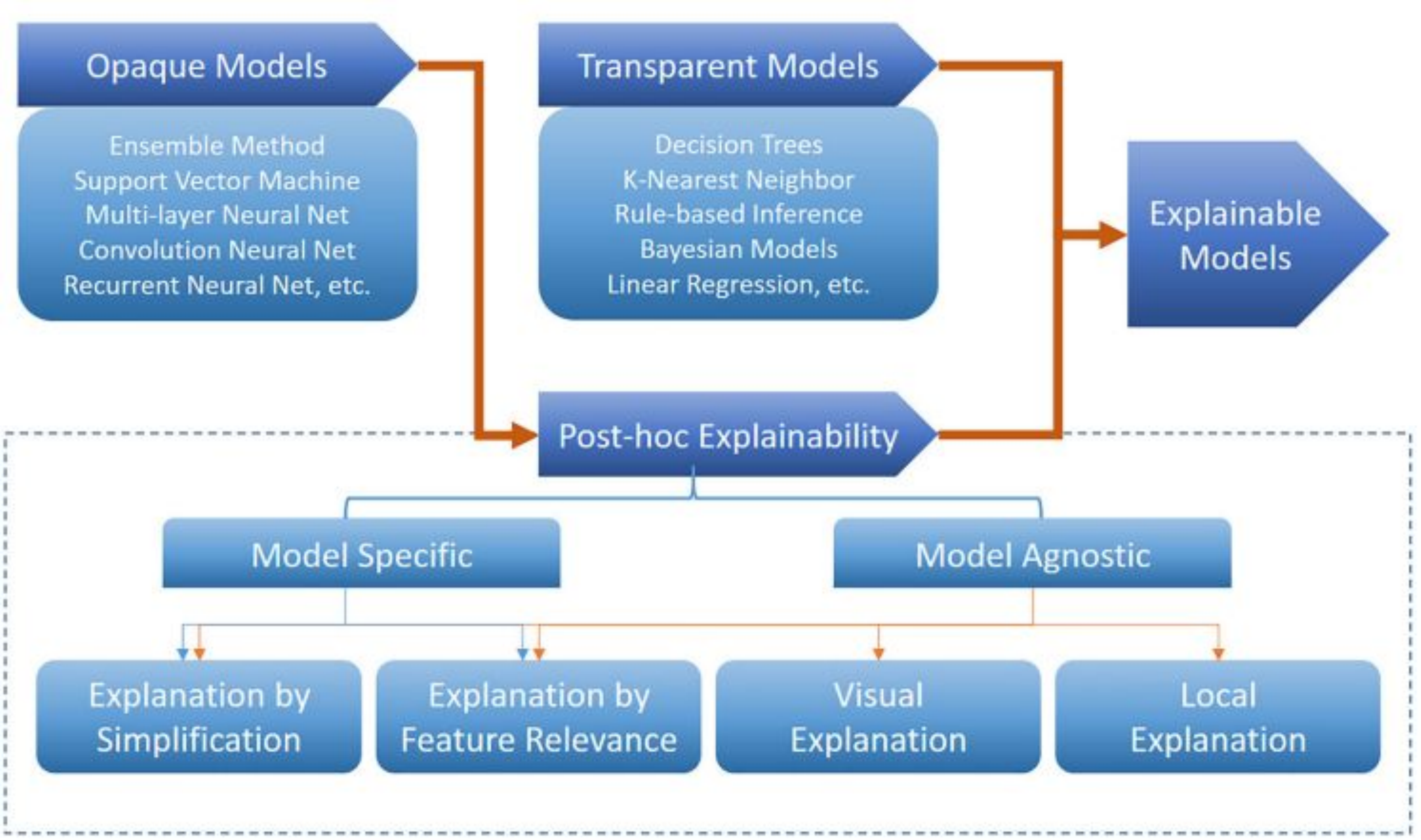


Figure 2: High-level XAI Taxonomy [1]

State of the Art - Taxonomy

- **Transparent model**: The decisions from these models are often transparent, although transparency, as a property, is not sufficient to guarantee that a model will be readily explainable.

- **Opaque model**: Although these models often achieve high accuracy, they are not transparent.

- **Model agnostic**: Model-agnostic XAI approaches are designed with the purpose of being generally applicable. As a result, they have to be flexible enough, so that they do not depend on the intrinsic architecture of the model, thus, operating solely on the basis of relating the input of a model to its outputs.

- **Model-specific**: Model-specific XAI approaches often take advantage of knowing a specific model and aim to bring transparency to a particular type of one or several models (Bach et al., 2015).

State of the Art - Taxonomy

- **Explanation** by **simplification**: By simplifying a model via approximation (Tritscher et al., 2020), we can find alternatives to the original models to explain the prediction we are interested in. For example, we can build a linear model or a decision tree around the predictions of a model, using the resulting model as a surrogate to explain the more complex one

- **Explanation** by **feature relevance**: This idea is similar to simplification. Roughly, this type of XAI approaches attempts to evaluate a feature based on its average expected marginal contribution to the model's decision, after all possible combinations have been considered

- **Visual explanation**: This type of XAI approach is based on visualization (Chattopadhay et al., 2018). As such, the family of data visualization approaches can be exploited to interpret the prediction or decision over the input data.

- **Local explanations** (Selvaraju et al., 2017) approximate the model in a narrow area, around a specific instance of interest, and offer information

XAI State of The art

- **Feature-oriented Methods**
  - **SHAP - to explain ML predictions**
    - SHAP seeks to deduce the amount each feature contributed to a decision by representing the features as players in a coalition game
  - **Class Activation Maps (CAM) - specific to CNN`s**
    - CAMs represent the per-class weighted linear sum of visual patterns present at various spatial locations in an image. More formally, global average pooling is applied to the final convolutional feature map in a network, before the output layer
  - **Gradient-weighted class activation mapping (Grad-CAM)**
    - A heatmap representation of the Grad-CAM indicates which regions of the input image were most important in the CNN's decisions
      - Output: However, GradCAM produces only coarse-grained visualizations and cannot explain multiple instances of the same object in an image

Feature oriented methods provide insights into where a decision is taking place in terms of the input, but fall short of a human level explanation of how and why the model came to those decisions

Taxonomy - Solutions

○ Global Methods

- **Global Attribution Mappings (GAMs) (Ibrahim et al., 2019)** to **NN**
  - For features with precise semantic definitions, GAMs can explain a NN prediction on a global level, across subpopulations, by formulating attributions as weighted conjoined rankings.
  - GAMs find a pair-wise rank distance matrix between features and a K-medoids clustering algorithm used to group similar local feature importances into clusters
- **Gradient-based Saliency Maps (Simonyan et al., 2013)**
  - are a visualization technique which render the absolute value of the gradient (in respect to the input features) of the majority predicted class as a normalized heatmap.
  - The pixels with a high activation are highlighted and correspond to areas that are most influential (i.e., salient)
- **Deep Attribute Maps (Ancona et al. 2018),**
  - Technique for rendering the explainability of gradient-based methods
  - It illustrates evaluations between different saliency-based explanation models.

Taxonomy - Solutions

- **Concept Models**
  - **Concept Activation Vectors (CAVs)** - (Kim et al. 2021) - to **NN**
    - a technique to explain globally the internal states of a neural network by mapping human understandable features to the high-level latent features extracted by the neural network
- **Surrogate Models**
  - **Local interpretable model-agnostic explanations (LIME) (Dieber & Kirrane, 2020)** to **ML** models
    - it is a model-agnostic technique to create locally optimized explanations of ML models
    - LIME trains an interpretable surrogate model to learn the local behavior of a global "black box" model's predictions
    - For image classification, an input image is divided into patches of contiguous superpixels (i.e., an image object) and a weighted local model is then trained on a new set of permuted instances of the original image.

Taxonomy - Solutions

- **Local pixed-based Methods**
  - **Layer-wise relevance propagation (LRP) (Bach et al., 2015)**
    - uses predefined propagation rules to provide an explanation of a multilayered neural network's output in respect to the input.
    - The method renders a heatmap, thereby providing insight into which pixels contributed to the model's prediction and the extent to which they did.
    - provides only a simplified distillation of the features'
- **Human-centric Methods**

It as a human-centric (anthropomorphic) phenomena rather than reducing it to statistics.
Indeed, humans compare items (e.g., images, songs, and movies) in their entirety and not per feature or pixel

  - **Angelov and Soares (2020)**

Research - Robustness Analysis of Deep Learning-Based Lung Cancer Classification Using Explainable Methods

- **Robustness Analysis of Deep Learning-Based Lung Cancer Classification Using Explainable Methods**
  - **Focus: Visual Explanation**
    - **CAD Images**

- **Post-hoc Explainable Models (Agnostic Methods)**
  - **Saliency Maps** **(Local Method - Feature Oriented - Agnostic) - VISUALIZATION**
  - **Integrated Gradients** **(Local Method - Feature Oriented - Agnostic) - VISUALIZATION**
  - **LRP** **(Local Method- Feature Oriented - Agnostic)**
  - **Deep Lift** **(Global Method - Feature Oriented - Specific)**
  -
  - **LIME** **(Local Method- Agnostic)**
  - **SHAP** **(Both Method- Agnostic)**

Research - Robustness Analysis of Deep Learning-Based Lung Cancer Classification Using Explainable Methods

- **Robustness Analysis of Deep Learning-Based Lung Cancer Classification Using Explainable Methods**

"The post-hoc and agnostic models such as Local Interpretable Model-Agnostic Explanations(LIME) and SHapley Additive exPlanations (SHAP) **were used to identify the most relevant genes that allow the lung cancer type and subtype classification** [17], or the most important features for lung cancer survival prediction [18]"

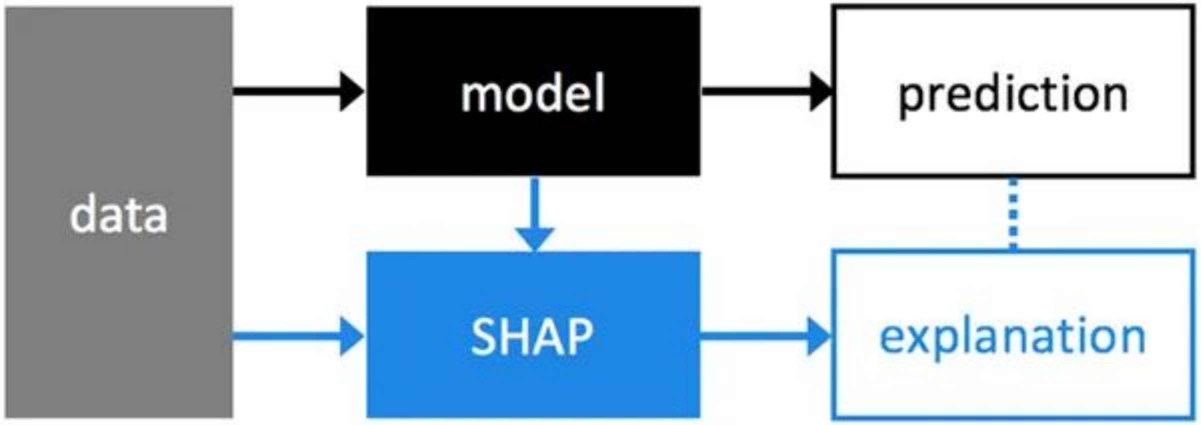
"**Our study focus broadens the spectrum of usability of these methods,** by implementing them for a different purpose and studying the coherence of explanations along models trained with different train-test split, consequently assessing the robustness of the developed architecture without compromising its predictive ability"

Pergunta: Qual foi o desempenho do modelo comparado com SHAP x LIME?

7 Python Frameworks for XAI

1. SHAP (SHapley Additive exPlanations)

SHAP is a model agnostic framework. It is a game theory based approach to explain the output of any ML model.



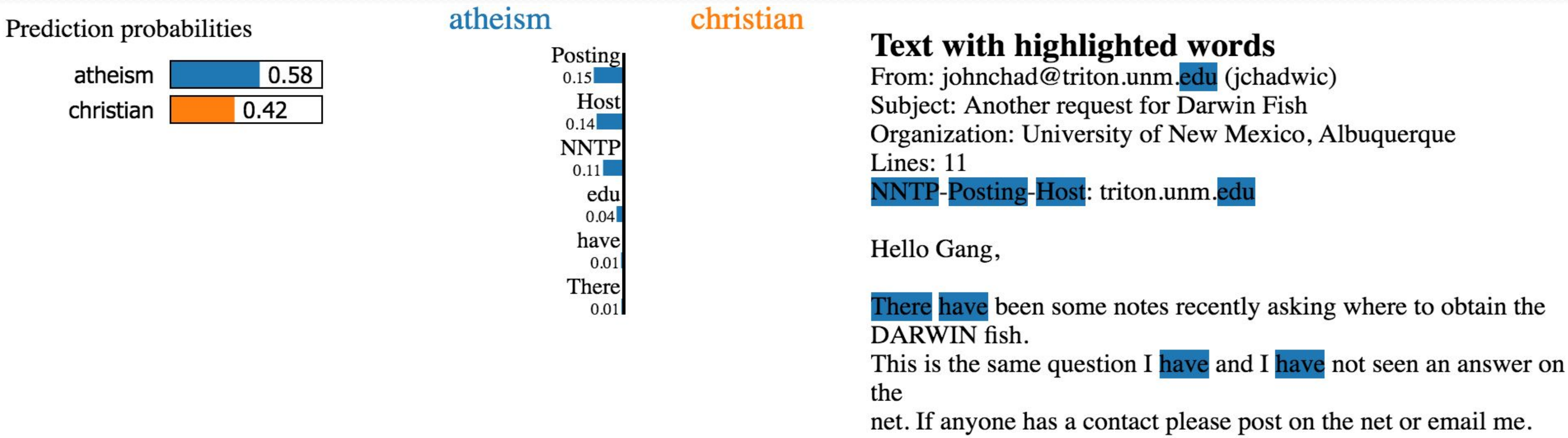
If the model is not additive then interpretation of the Shapley values is not always transparent, as predictive models may have non independent pay-off splits

2. LIME (Local Interpretable Model-Agnostic Explanations)

Lime basically tries to give a local linear approximation of the model's behaviour by creating local surrogate models which are trained to mimic the ML model's predictions locally.

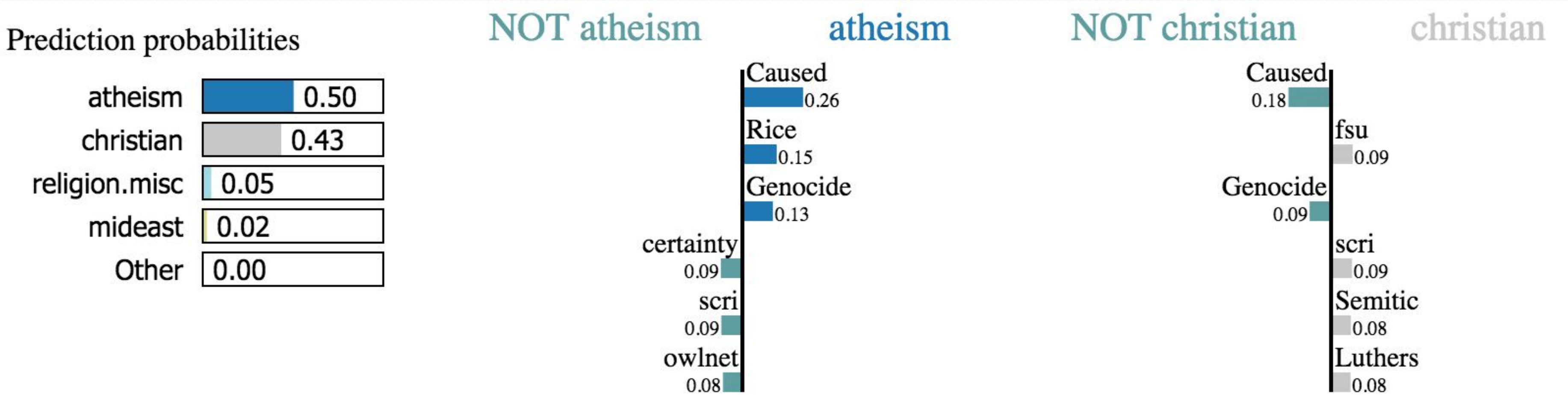
While global importance shows an average effect across the whole data set, a local-level observation may be influenced in different ways from each variable.

2 Class Case



2. LIME (Local Interpretable Model-Agnostic Explanations)

Multi Class Case



Tabular Data

Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

2. LIME (Local Interpretable Model-Agnostic Explanations)

Lime helps explain predictions for tabular data, images and text classifiers.

Lime basically tries to give a local linear approximation of the model's behaviour by creating local surrogate models which are trained to mimic the ML model's predictions locally.

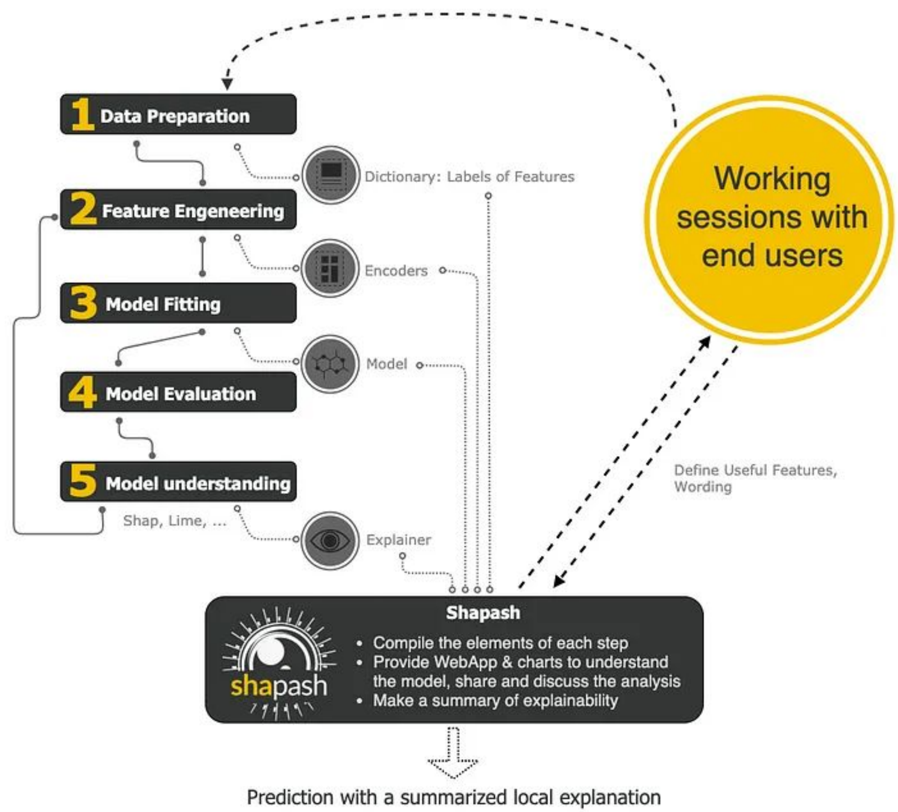
While global importance shows an average effect across the whole data set, a local-level observation may be influenced in different ways from each variable.

2 Class Case

3. Shapash

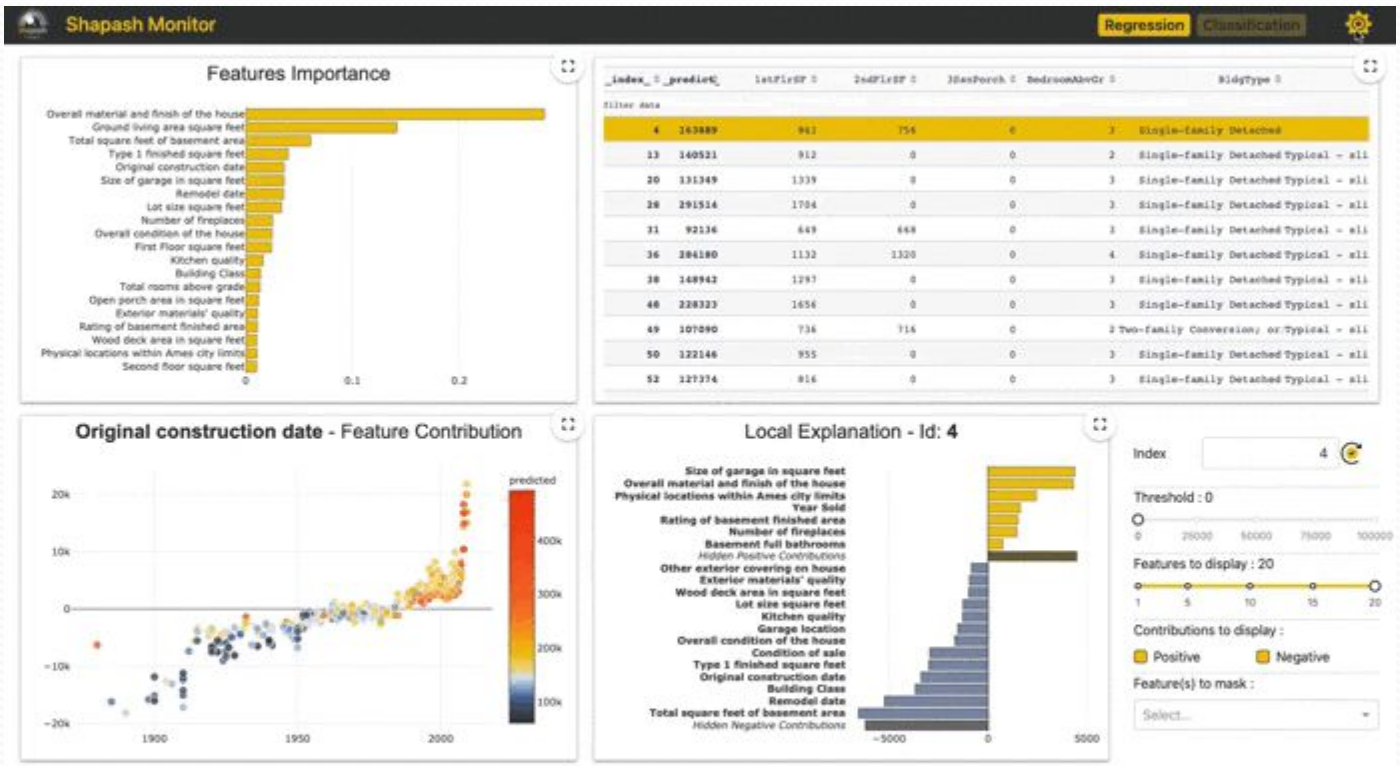
It is a Python library built by data scientists of a French insurer, MAIF.

It uses Shap or Lime backend to compute contributions. Shapash relies on the different steps necessary to build a ML model to make the results understandable.



3. Shapash

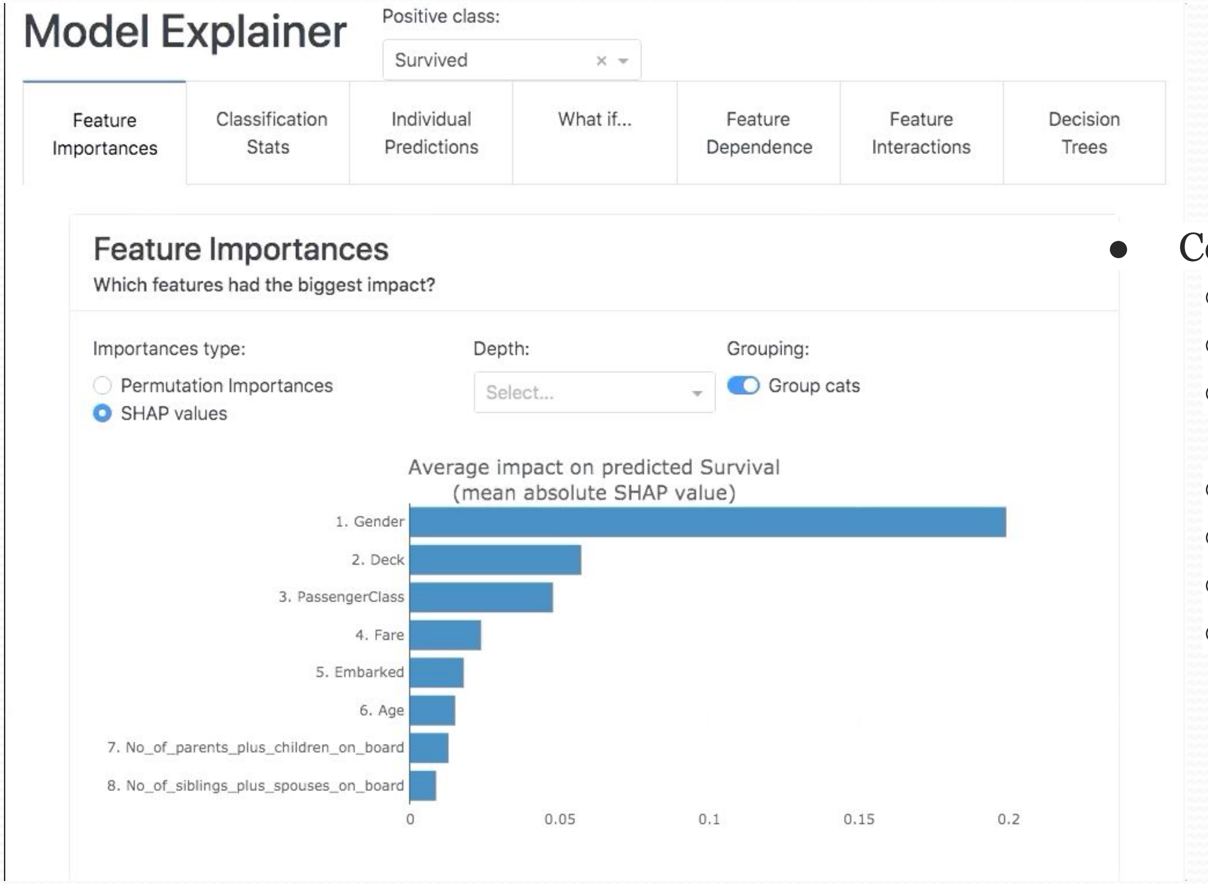
It works for regression, binary classification or multiclass problems and is compatible with many models: *Catboost*, *Xgboost*, *LightGBM*, *Sklearn Ensemble*, *Linear models* and *SVM*.



4. Explainer Dashboard

It is a an extensive and engaging interactive dashboard library to explain ML models across various spectrums and methodologies.

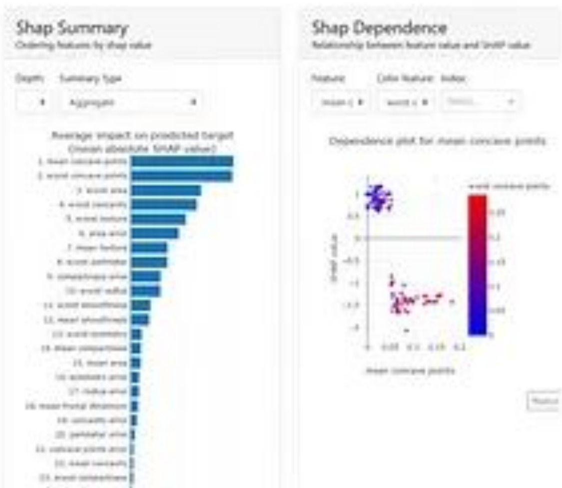
This is a lot more detailed compared to Shapash, i.e. not limited to just SHAP or Lime.



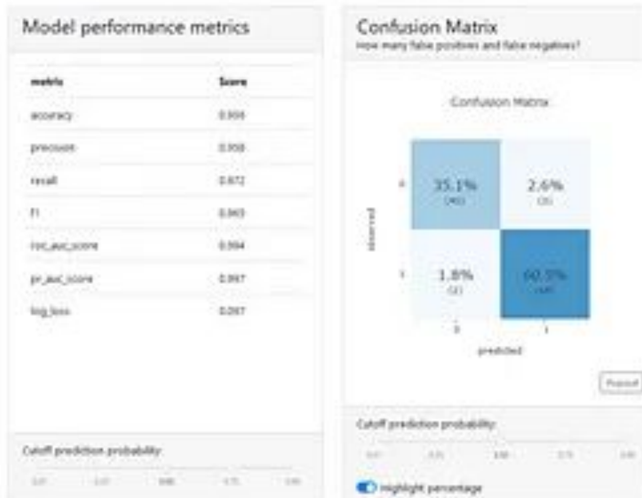
- Cover ML explainability aspects
  - Feature Importance
  - Metrics and evaluation
  - Local prediction explainability
  - What if analysis
  - Decision trees
  - Feature dependencies
  - Interactions

4. Explainer Dashboard

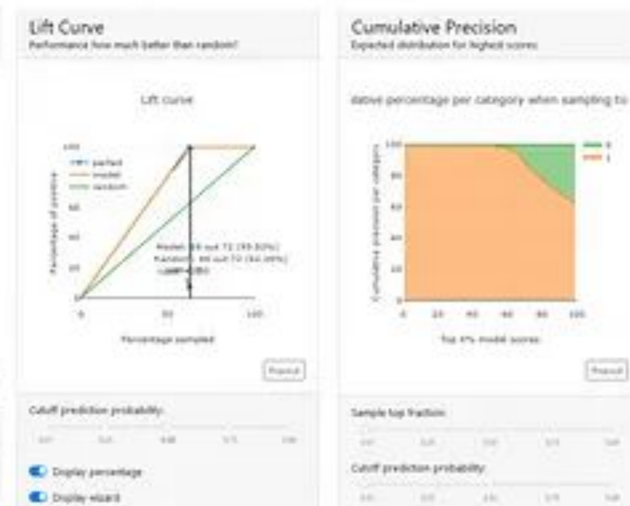
SHAP summary plots



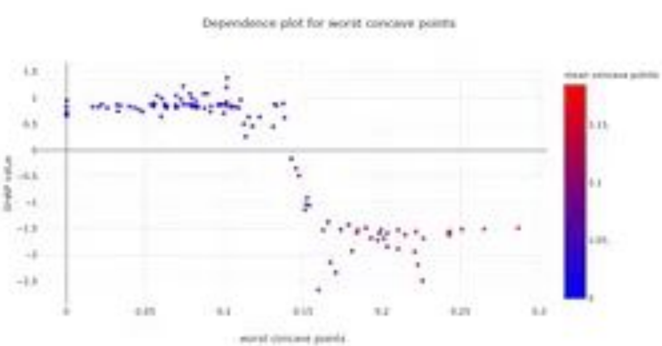
Model stats : Confusion matrix



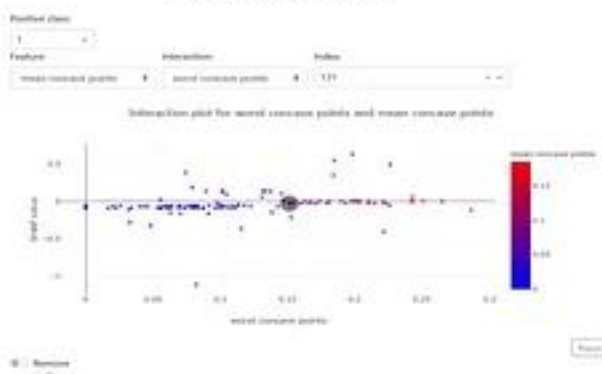
Model stats : Lift and PR



Dependence plots



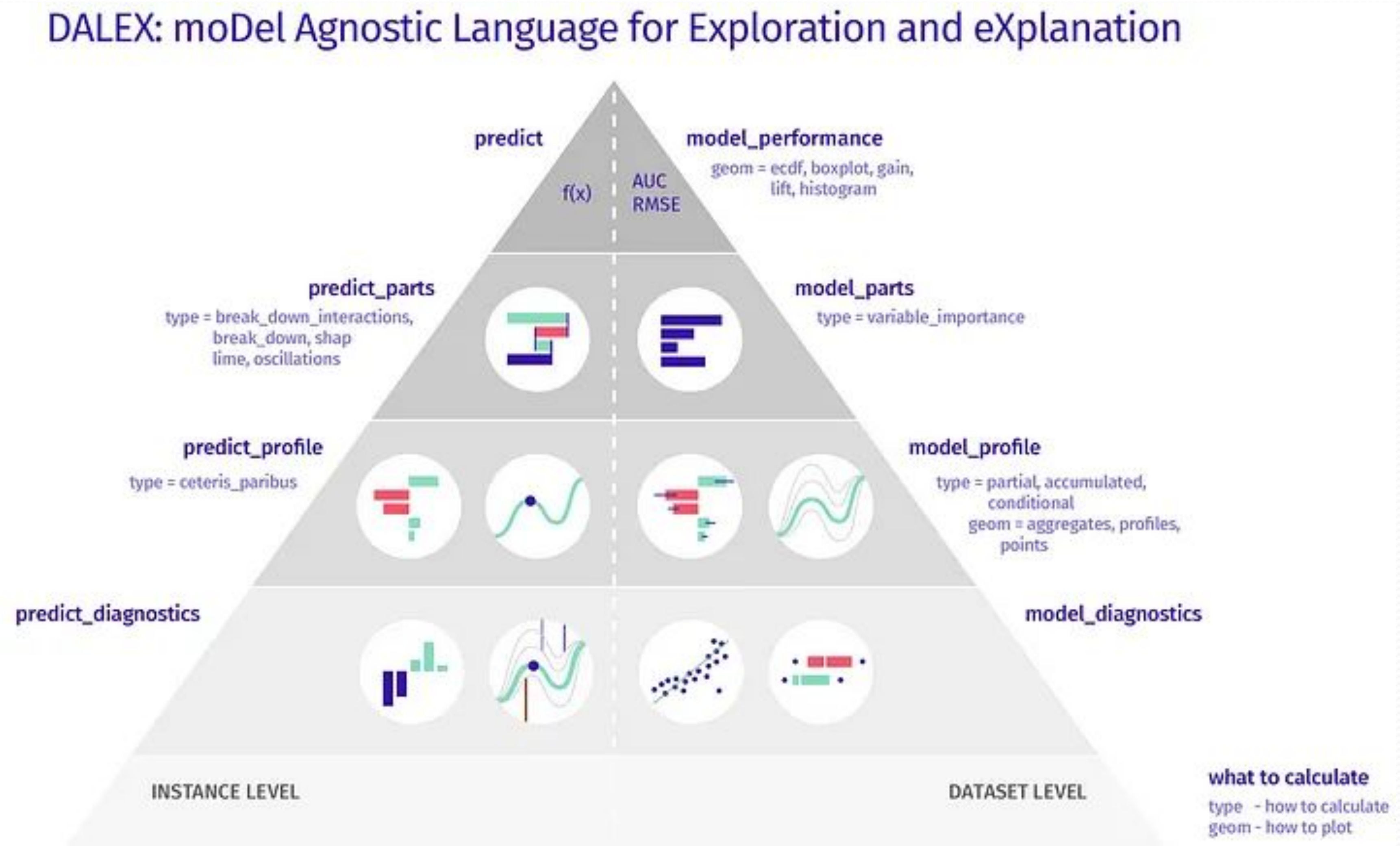
SHAP interactions



SHAP contributions

5. Dalex

It is a popular library which providers wrappers around various ML frameworks.



5. Dalex

The package available both in Python and R covers variable importance, PDP & ALE plots, Breakdown & SHAP waterfall plots. It also contains a neat wrapper around the native SHAP package in Python. This package works with various ML frameworks such as `scikit-learn`, `keras`, `H2O`, `tidymodels`, `xgboost`, `mlr` or `mlr3`.

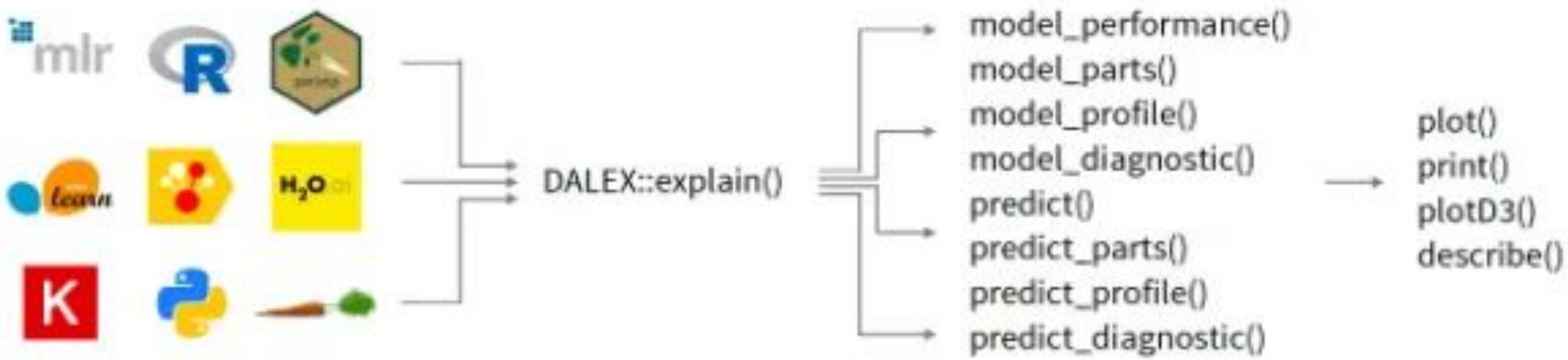
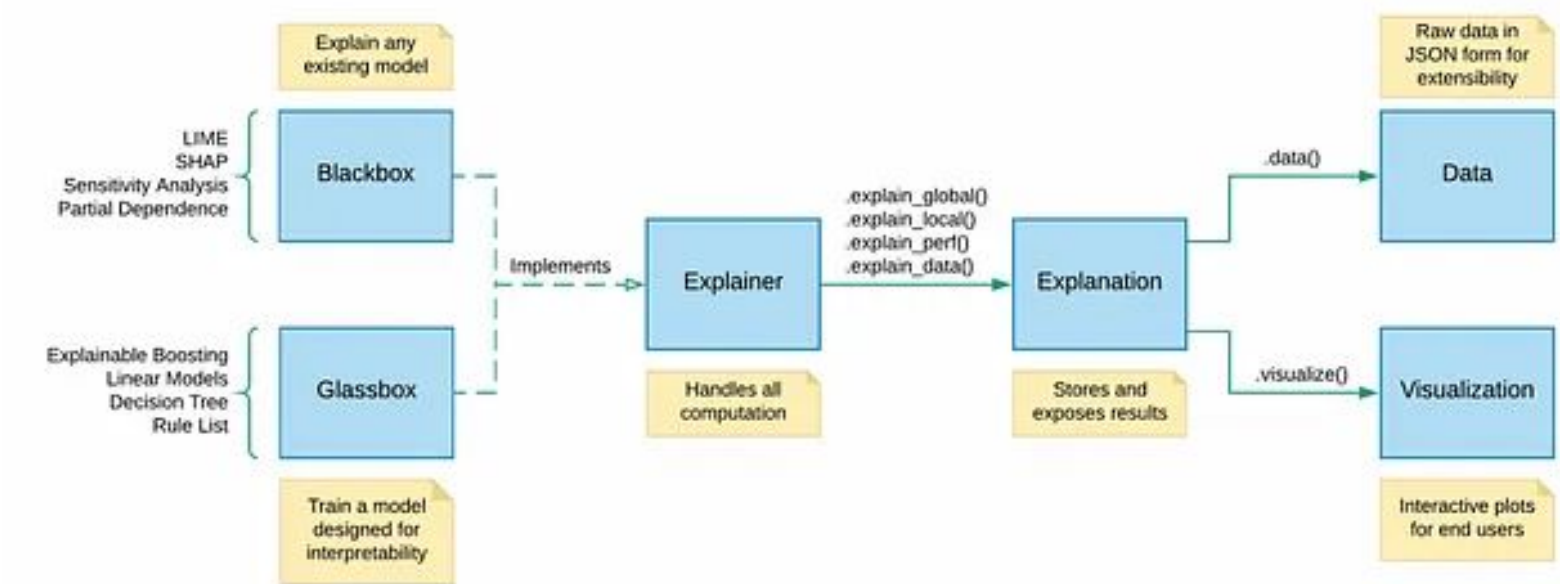


Figure 3.1: The `DALEX` package creates a layer of abstraction around models, allowing you to work with different models in a uniform way. The key function is the `explain()` function, which wraps any model into a uniform interface. Then other functions from the `DALEX` package can be applied to the resulting object to explore the model.

6. Explainable Boosting Machines (EBM)

It is an open source package from Microsoft that has a module of “glass-box” models which enable explainability.

This approach changes the training process from training a decision tree on all features to a decision tree with one feature at a time



Glassbox

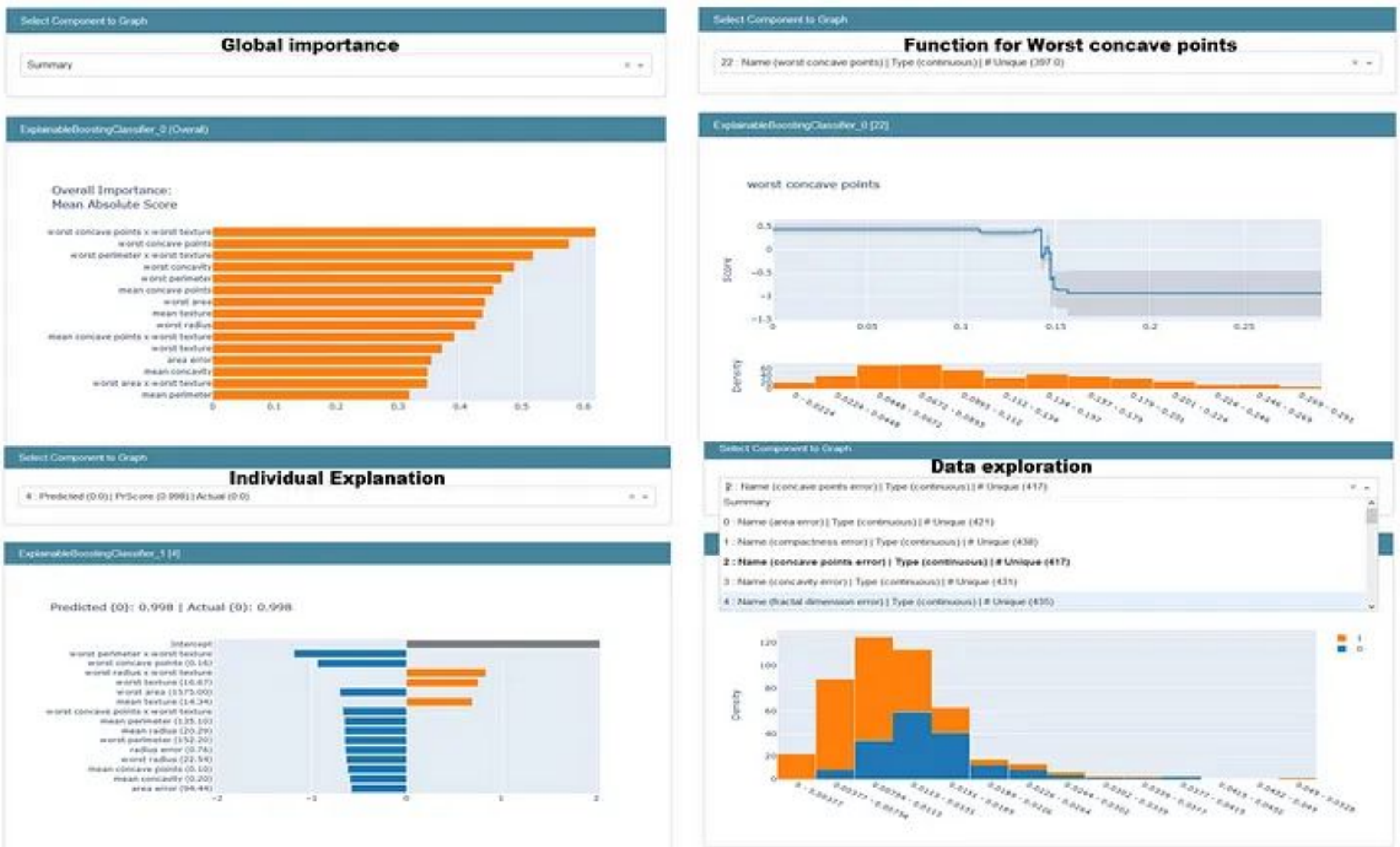
```python
from interpret import show
from interpret.glassbox import LogisticRegression
```

Blackbox

```python
from interpret import show
from interpret.blackbox import PartialDependence
from sklearn.neural_network import MLPClassifier
```

6. Explainable Boosting Machines (EBM)

Multiple trees are constructed per feature with small depth and in boosting fashion. Based on the summation of all its trees, we can estimate the function (*f*) of the input variable with the output variable.



7. ELI5

ELI5 is yet another explainability package by MIT supporting ML frameworks and packages such as *Scikit-learn, Keras, Xgboost,* etc. This package neatly categorizes the explanation into two

- Global which is show_weights() in which you can provide required parameters to specify type of global importance you wish to see (e.g. ‘gain’, ’weight’, ‘gain’, ‘cover’, ‘total_gain’, ‘total_cover’)
- Local level explanations are given by eli5.show_prediction()which explain the local level predictions

There are no engaging visuals that I could spot in this package. It’s mostly the unified platform across various ML packages/frameworks that is its main highlight.

Global importance

Weight	Feature
0.3846	mean concave points
0.2668	worst concave points
0.0647	worst perimeter
0.0552	worst area
0.0537	worst texture
0.0488	worst radius
0.0349	mean texture
0.0268	worst concavity
0.0147	area error
0.0081	mean smoothness
0.0068	compactness error
0.0063	concavity error
0.0043	worst smoothness
0.0042	mean area
0.0039	radius error
0.0038	perimeter error
0.0033	worst symmetry
0.0020	mean fractal dimension
0.0018	smoothness error
0.0014	mean compactness
	... 10 more ...

Individual Explanation

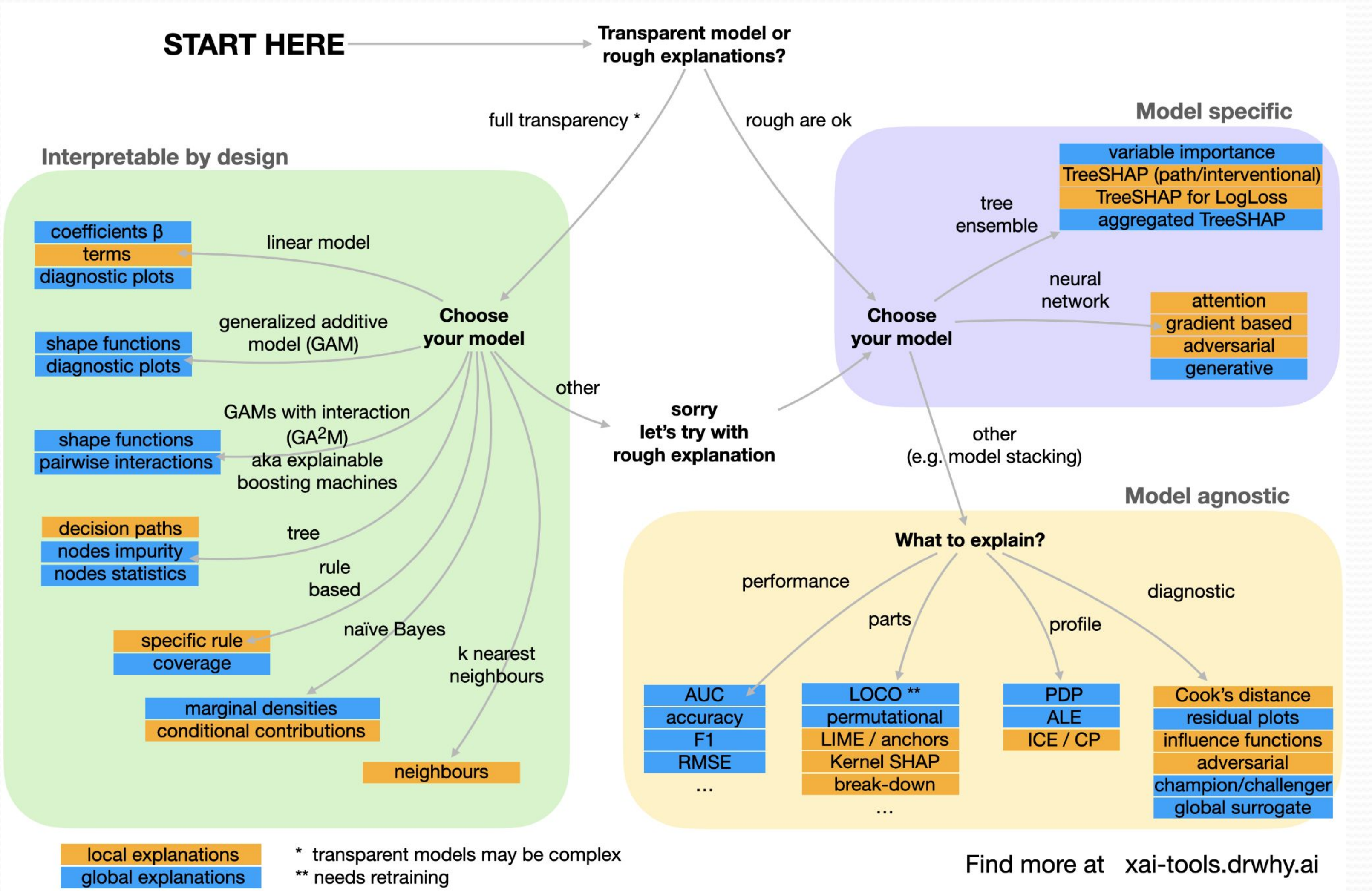
y=0 (probability **0.999**, score **-6.627**) top features

Contribution?	Feature	Value
+2.352	worst concave points	0.179
+1.776	mean concave points	0.080
+0.904	worst area	1866.000
+0.614	worst perimeter	165.900
+0.522	worst concavity	0.269
+0.388	worst texture	26.580
+0.258	area error	96.050
+0.191	mean smoothness	0.090
+0.162	compactness error	0.017
+0.120	mean area	1130.000
+0.103	mean texture	21.310
+0.102	radius error	0.789
+0.086	mean fractal dimension	0.055
+0.071	worst radius	24.860
+0.052	mean concavity	0.108
+0.040	symmetry error	0.014
+0.033	concavity error	0.023
+0.004	texture error	0.797
-0.016	mean compactness	0.103
-0.016	worst fractal dimension	0.066
-0.024	worst smoothness	0.119
-0.049	concave points error	0.014
-0.053	worst symmetry	0.255
-0.991	<BIAS>	1.000

Permutation importance

Weight	Feature
0.0115 ± 0.0126	worst concavity
0.0060 ± 0.0084	worst area
0.0058 ± 0.0049	worst radius
0.0058 ± 0.0049	mean concave points
0.0056 ± 0.0030	area error
0.0055 ± 0.0064	worst concave points
0.0040 ± 0.0032	worst perimeter
0.0033 ± 0.0030	worst texture
0.0021 ± 0.0013	worst smoothness
0.0011 ± 0.0014	mean smoothness
0.0005 ± 0.0013	mean area
0.0003 ± 0.0003	perimeter error
0.0002 ± 0.0003	concavity error
0.0001 ± 0.0010	radius error
0.0001 ± 0.0007	compactness error
0.0001 ± 0.0006	mean compactness
0.0000 ± 0.0006	mean fractal dimension
0 ± 0.0000	fractal dimension error
0 ± 0.0000	mean concavity
0 ± 0.0000	worst compactness
	... 10 more ...

Landscape of R packages for eXplainable Artificial Intelligence



ChatGPT

In the context of lung cancer classification, post-hoc explainability can help us understand how the model arrived at its predictions and which features or variables were most important in making those predictions.

Here are some post-hoc explainability techniques that can be used for lung cancer classification:

- Feature Importance: This technique involves identifying the most important features that the model uses to make its predictions. There are several methods to calculate feature importance, such as permutation importance and SHAP values.
- Partial Dependence Plots: This technique involves plotting the relationship between a feature and the predicted outcome while controlling for other variables. Partial dependence plots can help visualize how a feature affects the model's prediction.
- Local Interpretable Model-agnostic Explanations (LIME): LIME is a technique that explains the predictions of a model by approximating the model with a simpler, interpretable model. LIME generates local explanations for each prediction by perturbing the input data and observing how the output changes.
- Decision Trees: Decision trees are a visual representation of the decision-making process of a model. By examining the decision tree, we can see which features are most important in determining the model's predictions.

By using these post-hoc explainability techniques, we can gain insight into how the lung cancer classification model works and identify areas for improvement. This can be valuable for clinicians and researchers looking to understand the underlying factors contributing to lung cancer and to develop better diagnostic and treatment strategies.

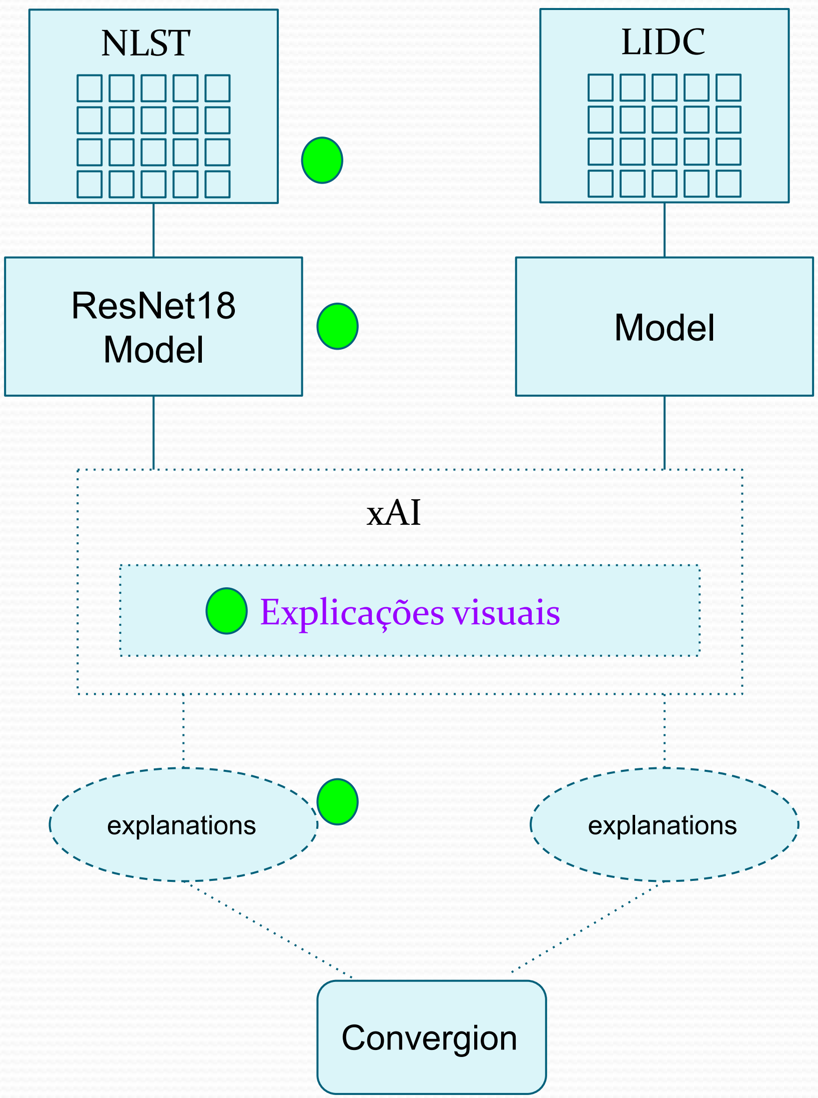
Proposal

- Keep studying the state-of-the-art interpretable ML approaches for image analysis (Python and R)
  - Avaliar Modelos Interpretáveis
    - Linear Regression
    - Logistic Regression
    - Decision Tree...
  - Model-Agnostic Methods
  - Global Model-Agnostic Methods
    - Partial Dependence Plot (PDP)
    - Accumulated Local Effects
    - Feature Interaction
    - Global Surrogate...
  - Local Model-Agnostic Methods
    - Local Surrogate (LIME)
    - SHAP
    - Counterfactual Explanations...
- Study the model and the Used Data.
- Compare the performance of our methodology (Gradient-based Saliency Maps) with Transparent models + LIME + SHAP
- Use them in conjunction with these methods to provide additional insights into model predictions.

There are many post-hoc explainability techniques that can be used for cancer classification, and the best one depends on the specific model and data being used. However, here are a few techniques that are commonly used for post-hoc explainability in cancer classification:

Solution Work

Post-Model



Feature Importance

Shap

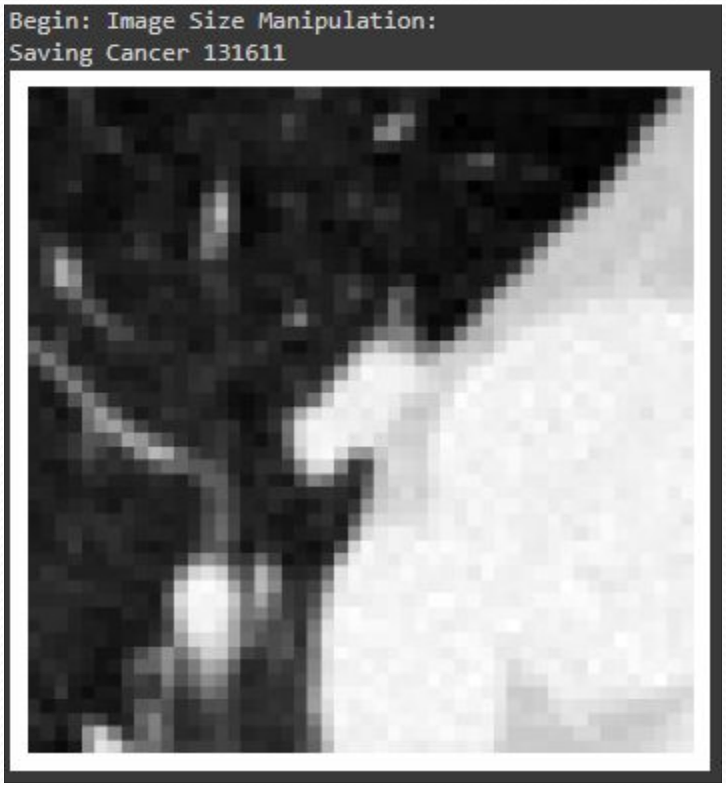
Importância de cada Feature
Árvores de decisão (xgboost)

Explicações visuais

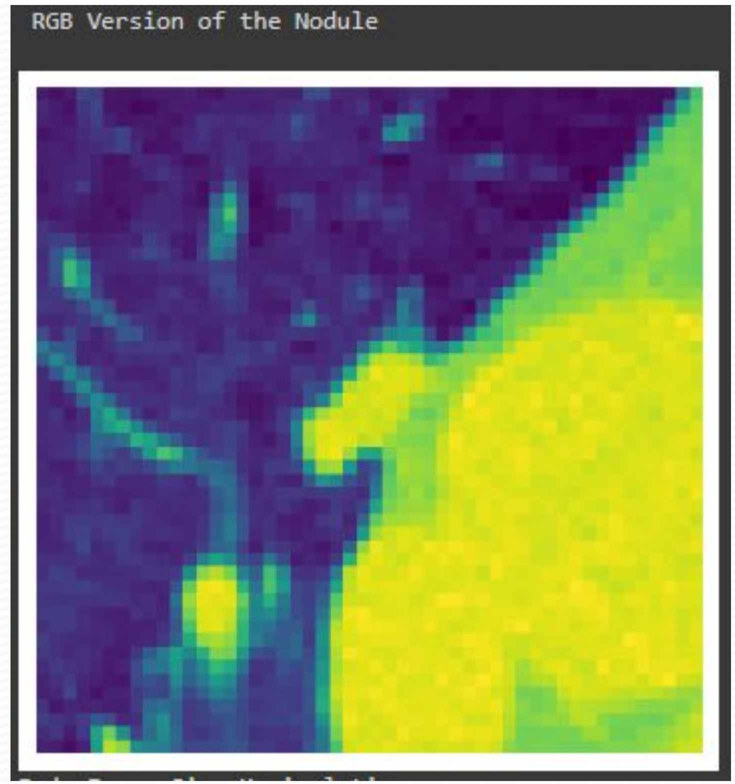
Vanilla Saliency Maps	Integrated Gradients
Image Gradients Generation (raw)	Image Gradients (Integral)
LRP	DeepLift

Visual Explanations

Original Nodule

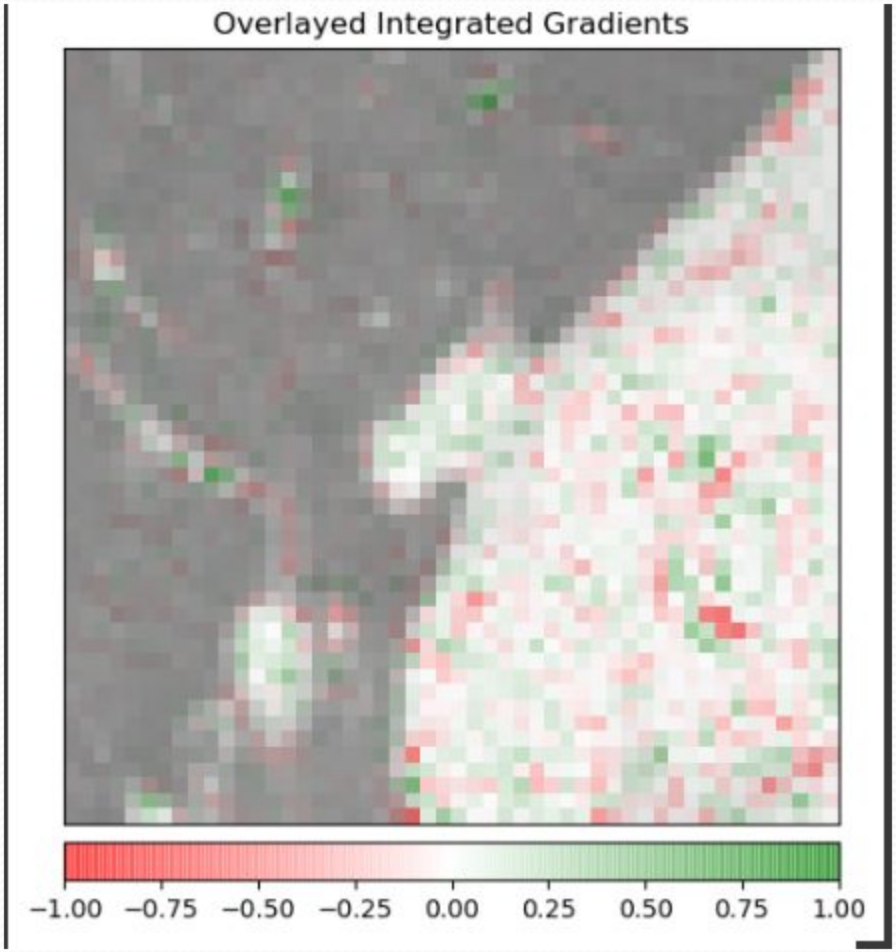


RGB Nodule

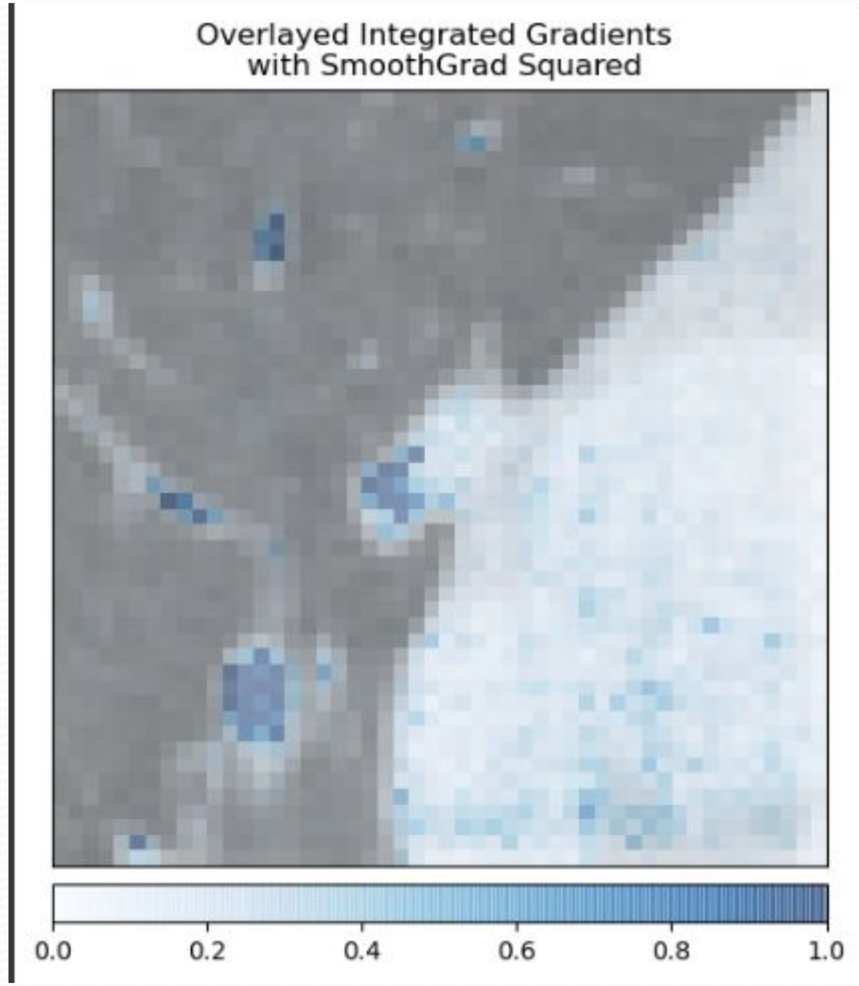


Visual Explanations

Integrated Gradients = All values

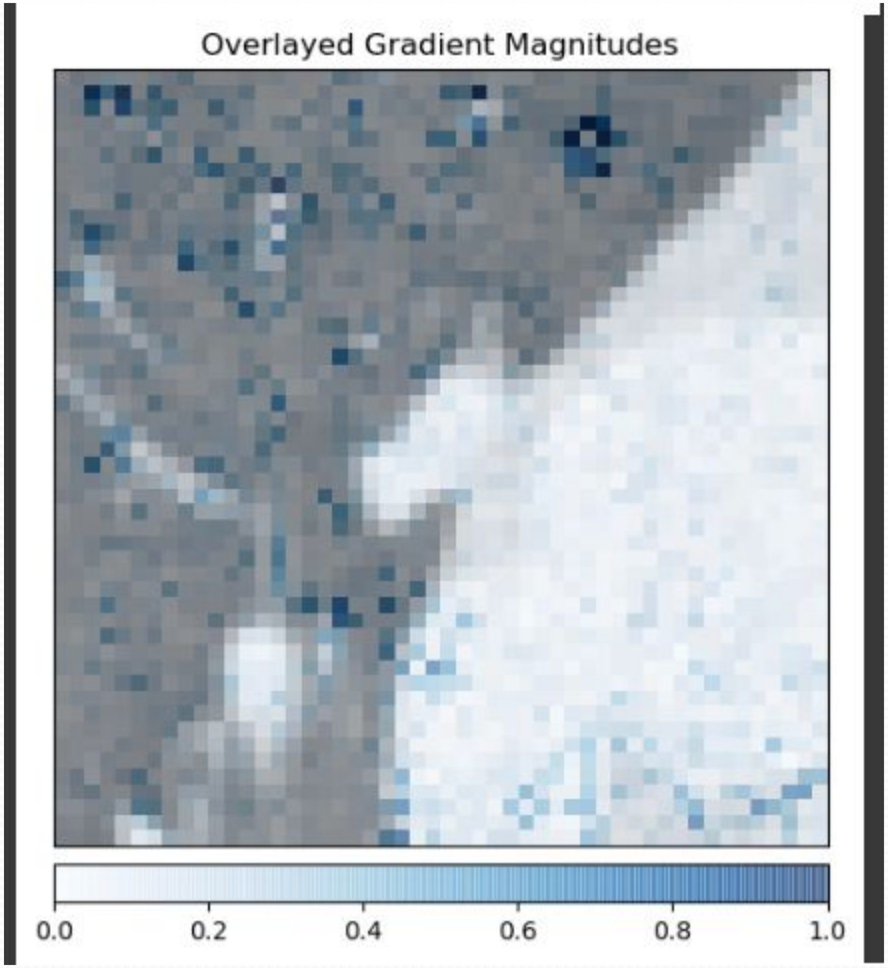


Integrated Gradients = Absolute Values

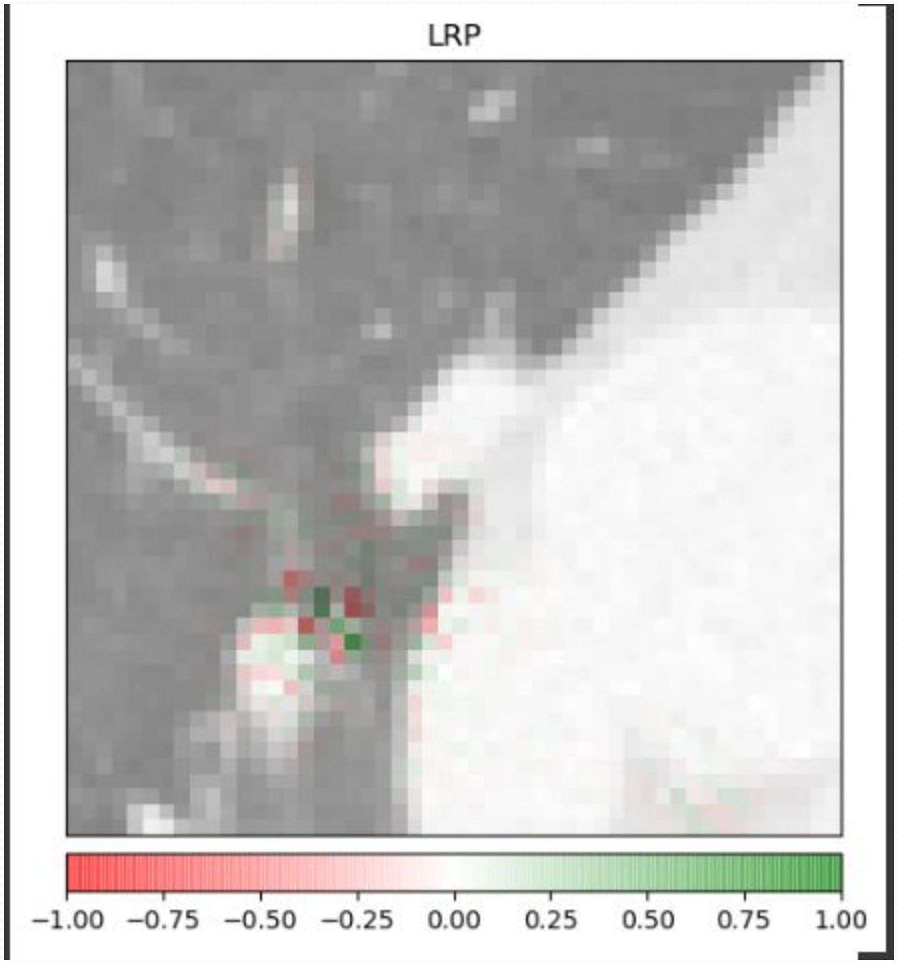


Visual Explanations

Saliency Maps



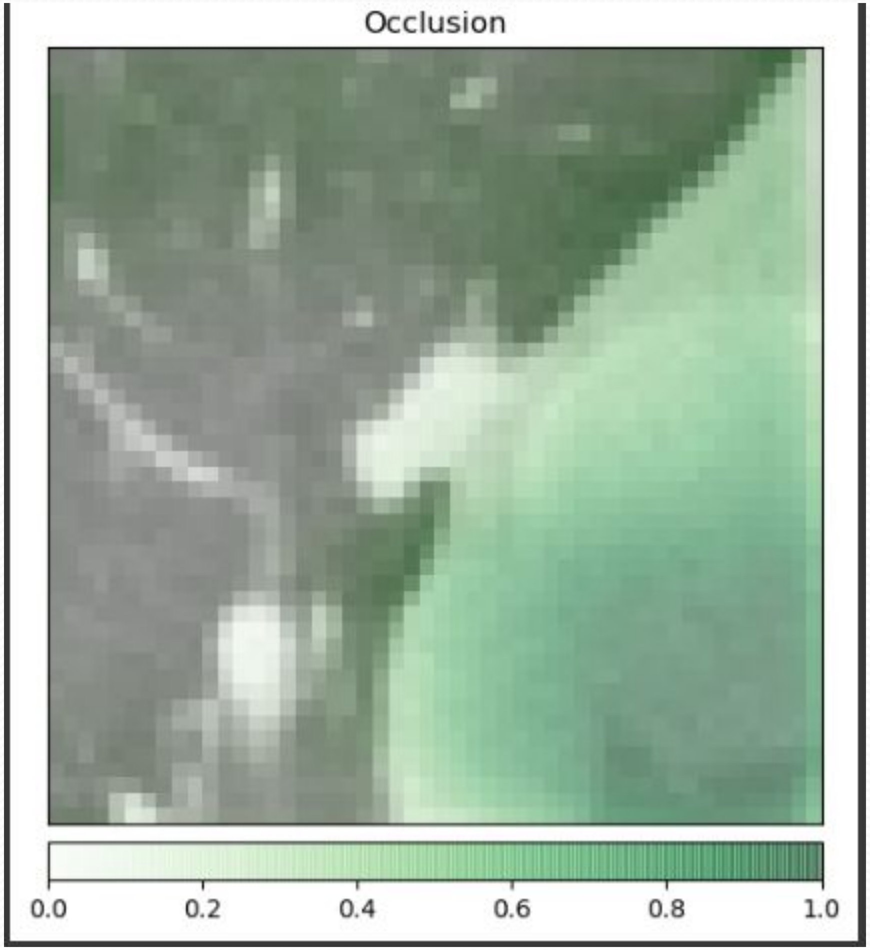
RRP



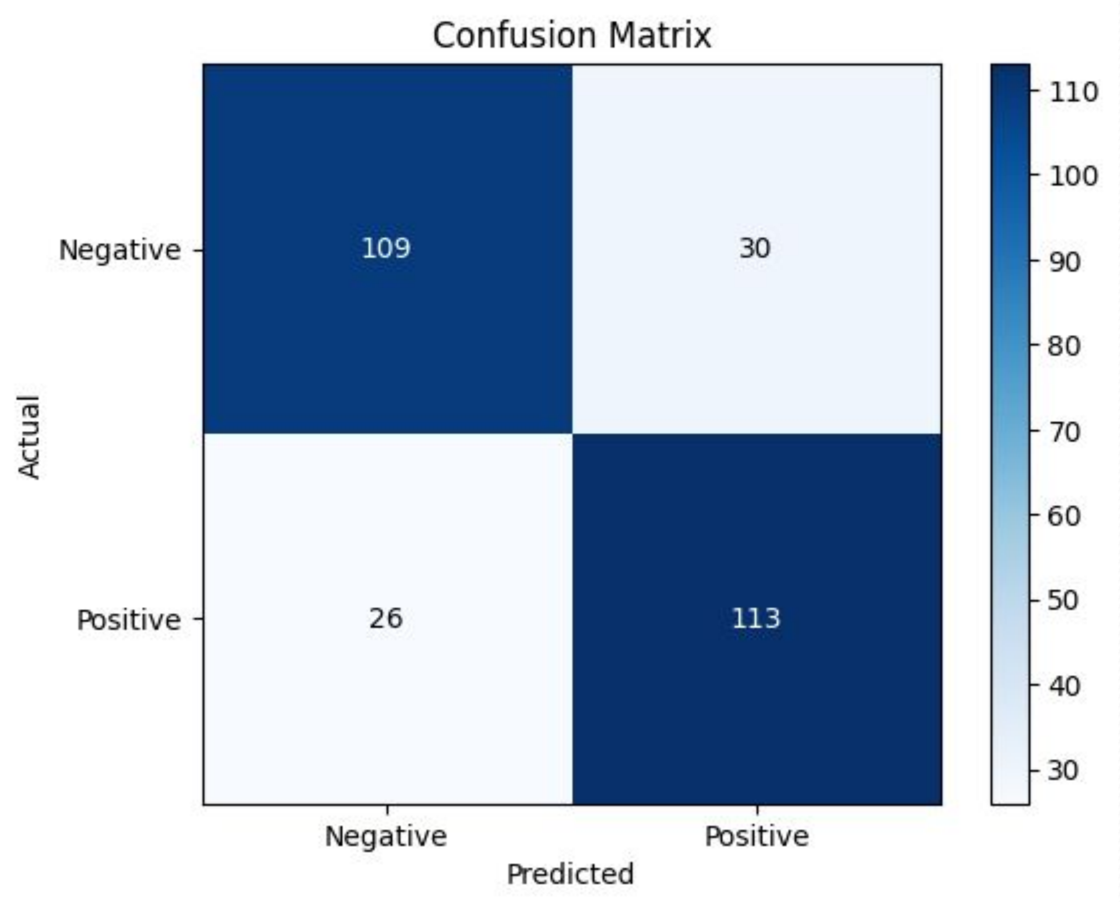
Visual Explanations

DeepLift

Oclusão

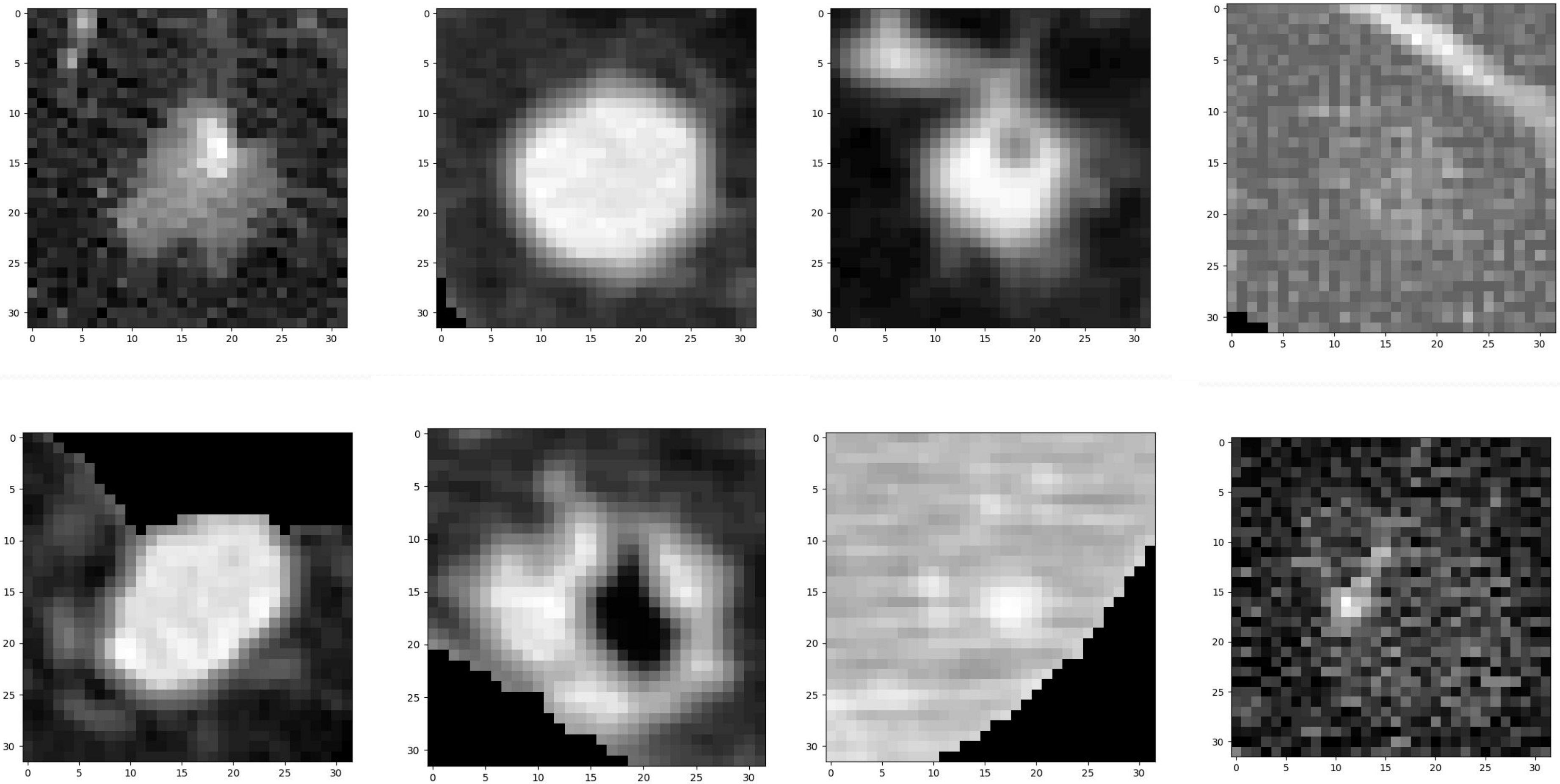


Visual Results
CNN 2D

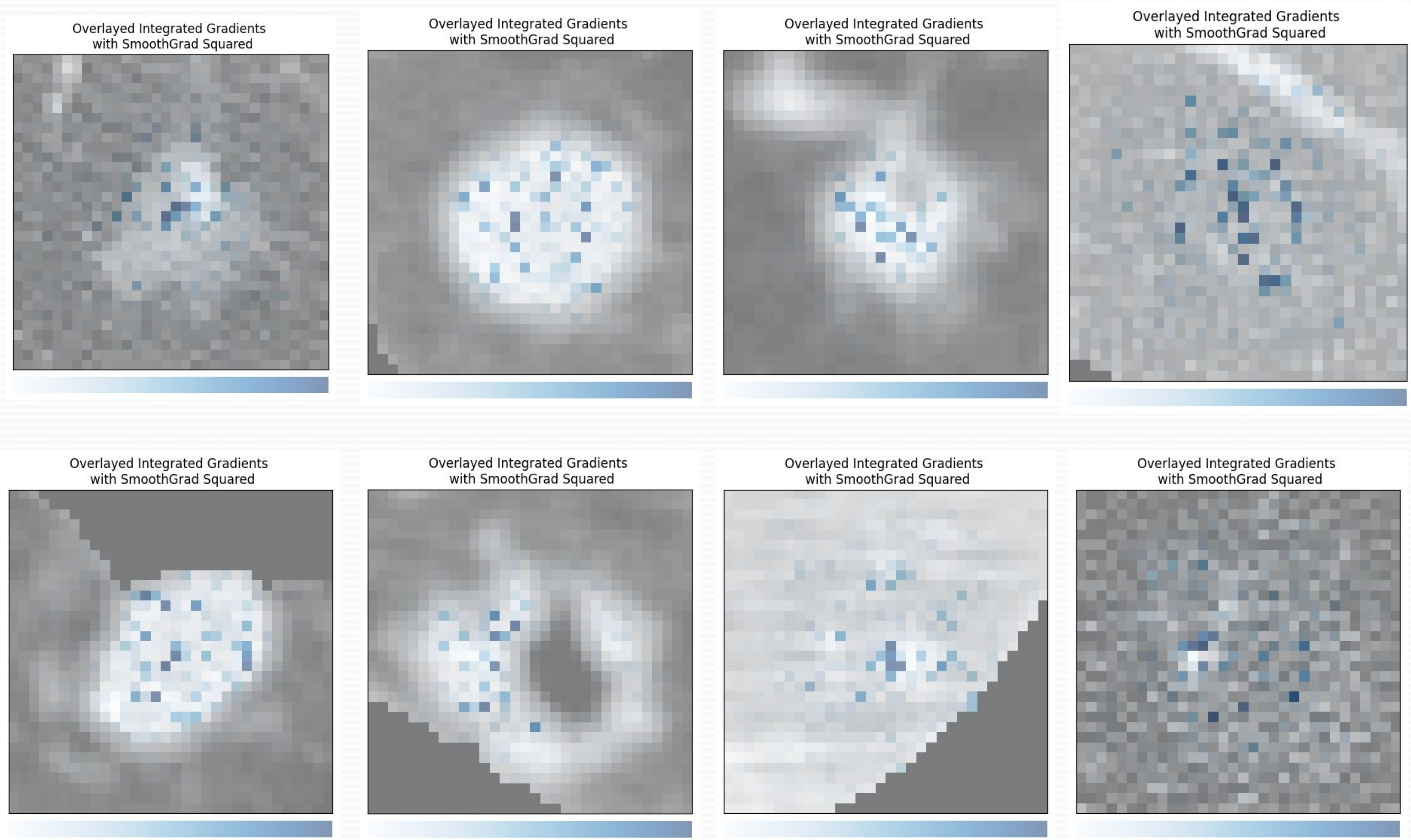


ACC: 79,14

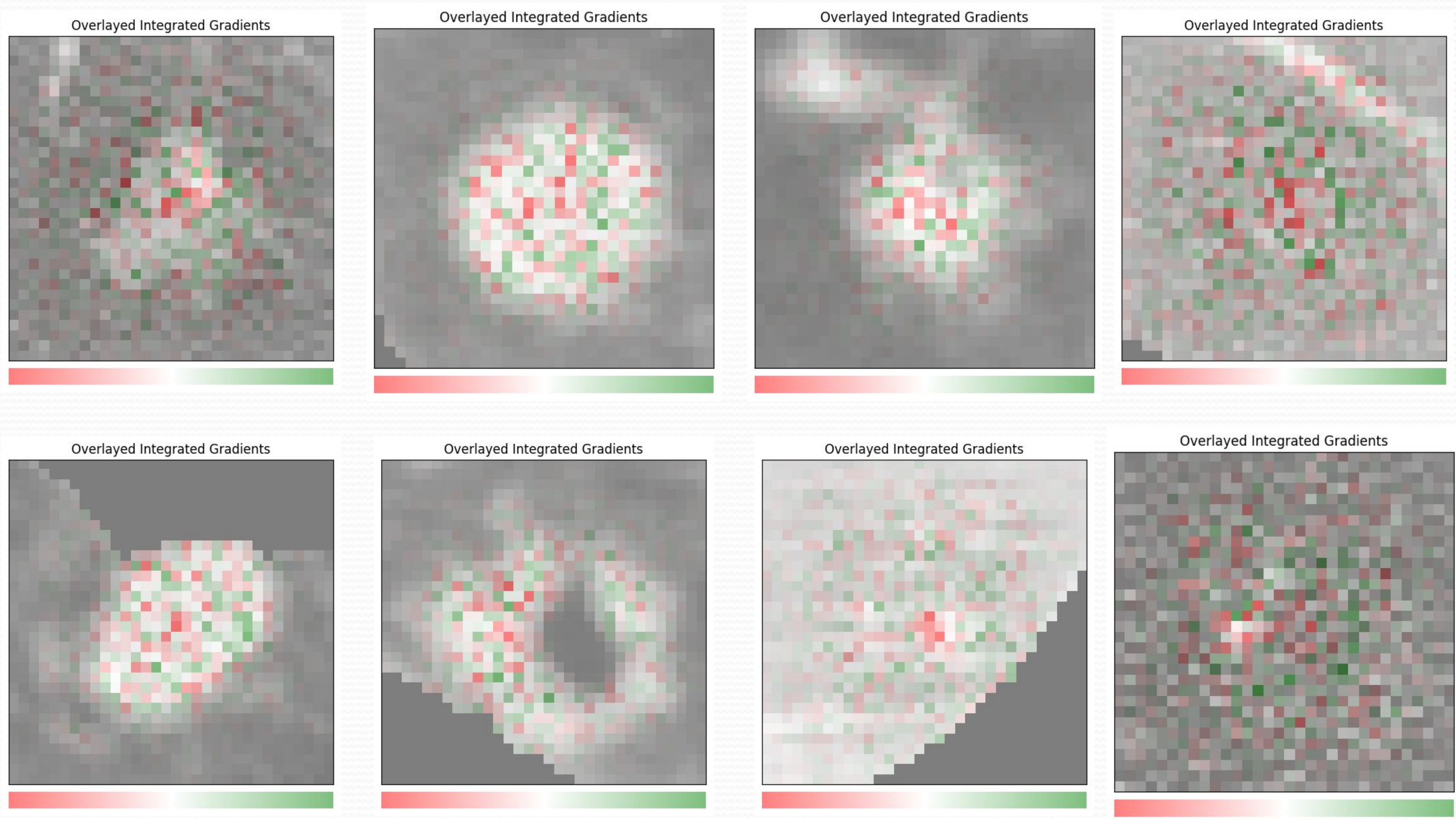
Y_true [True] & Y_pred = [True] -> Size= 113 Nodule



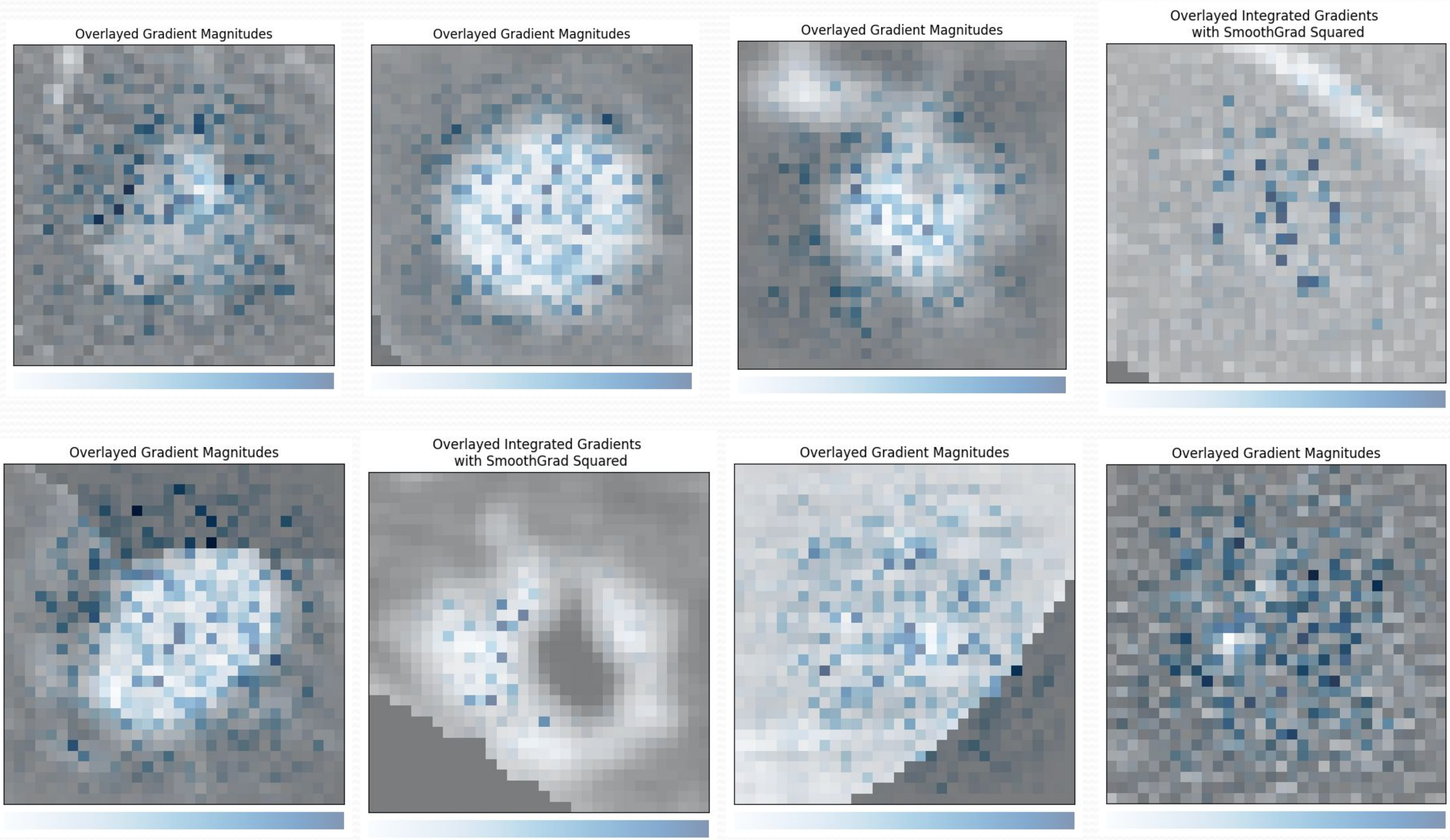
Y_true [True] & Y_pred = [True] -> Size= 113
IG_Abs



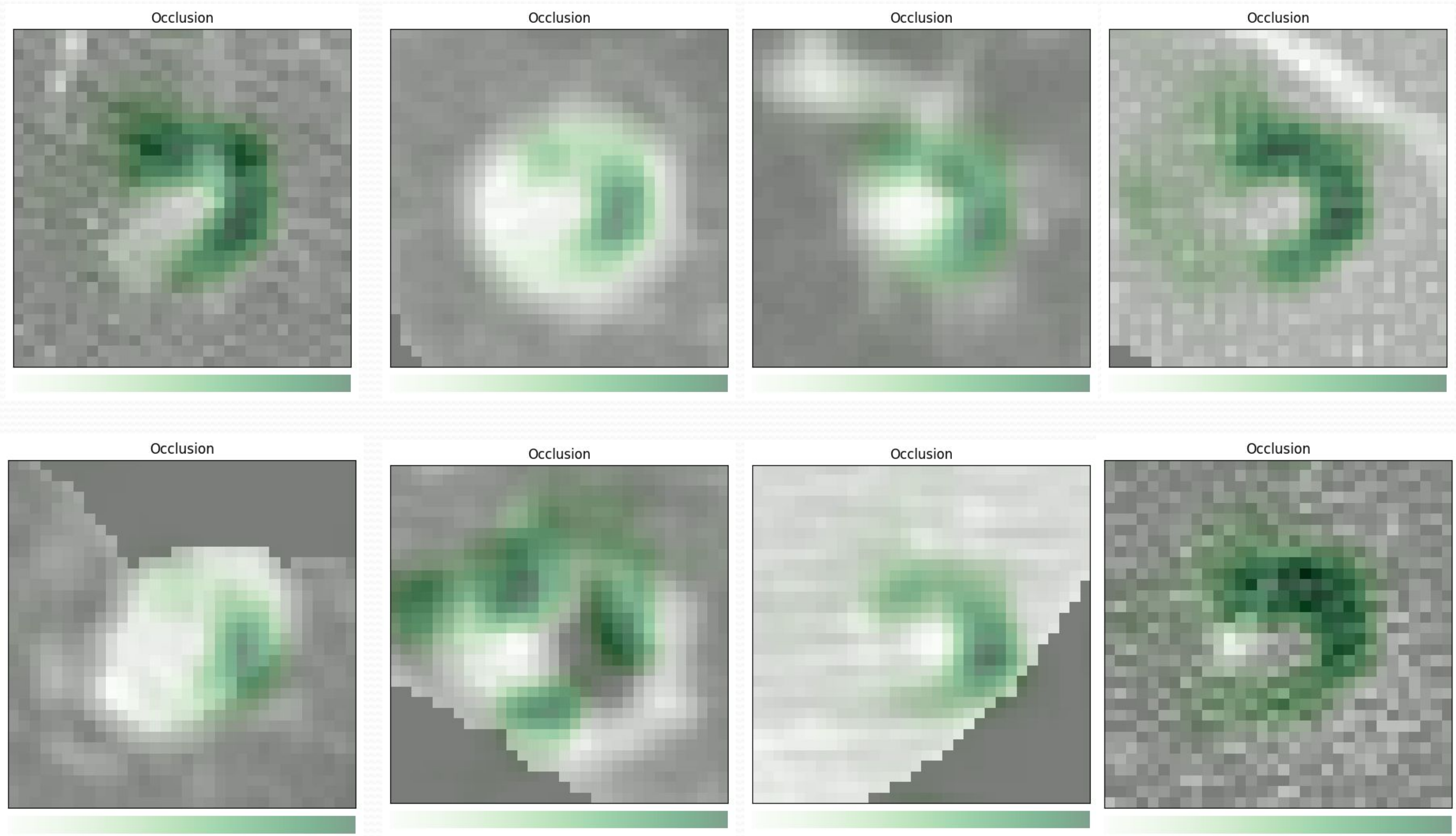
Y_true [True] & Y_pred = [True] -> Size= 113
IG_All



Y_true [True] & Y_pred = [True] -> Size= 113 SM

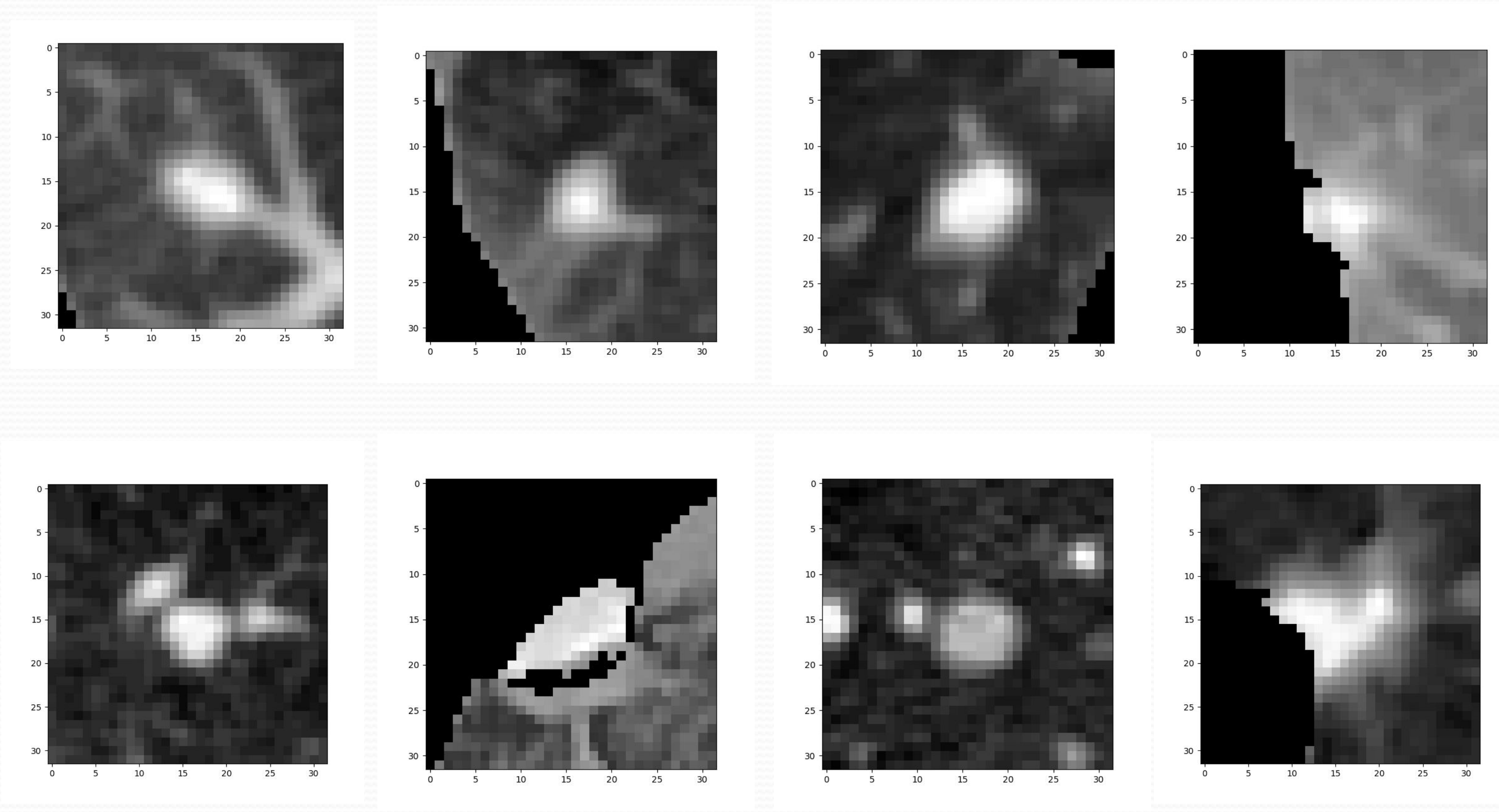


Y_true [True] & Y_pred = [True] -> Size= 113
OCC

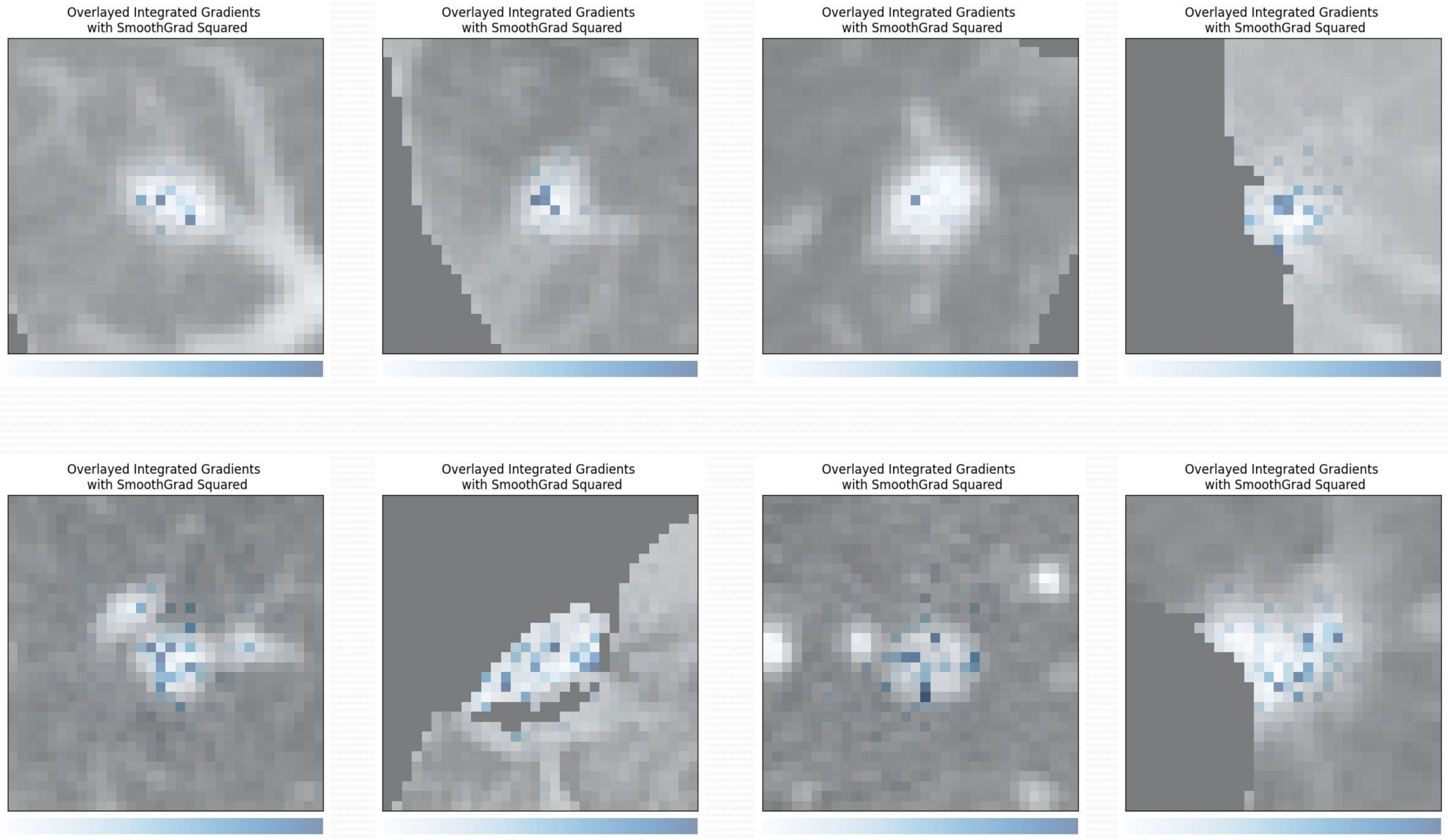


Y_true [True] & Y_pred = [False] -> Size= 26

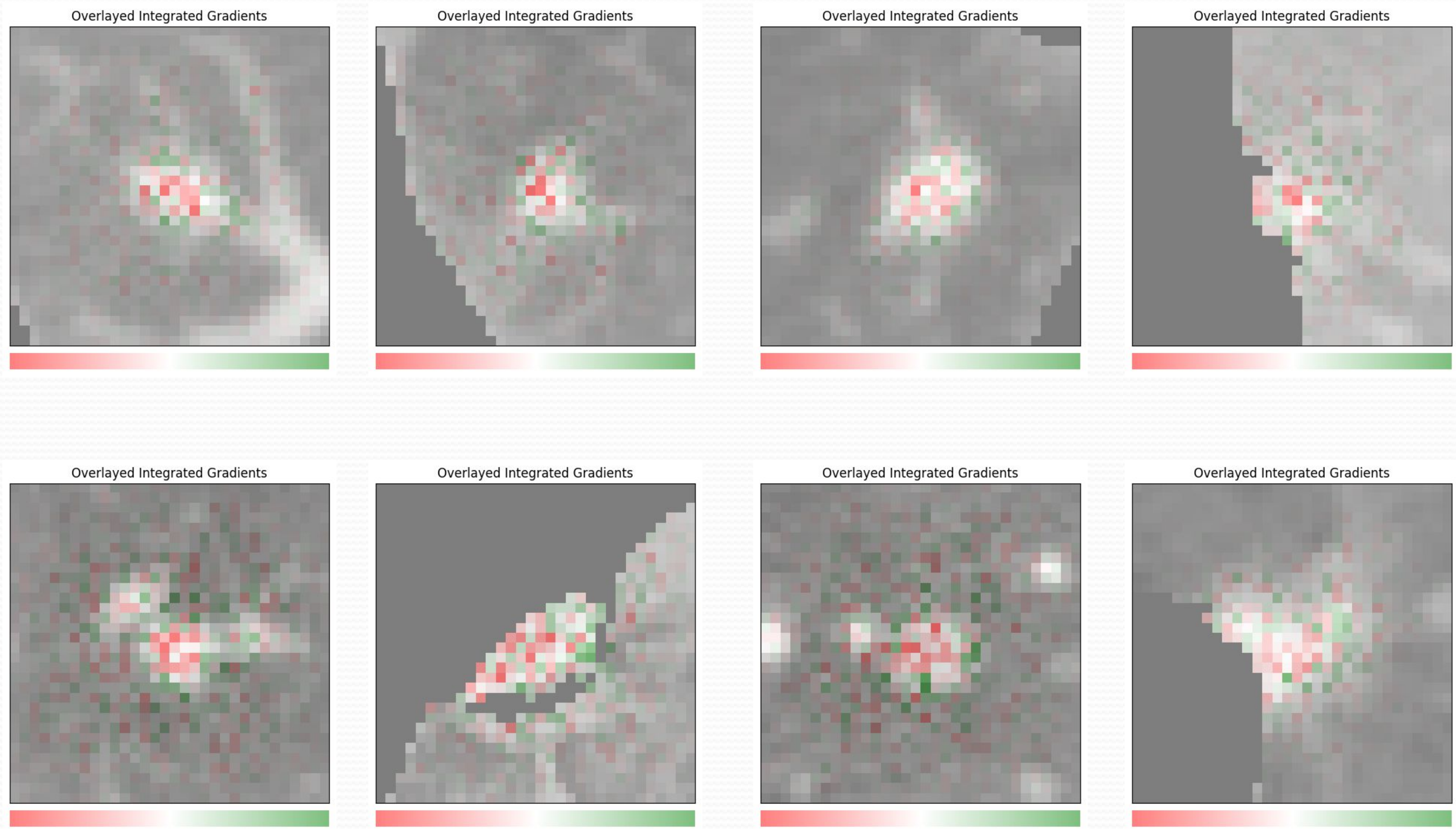
Nodule



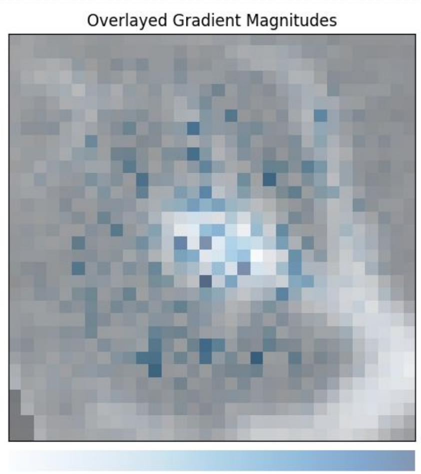
Y_true [True] & Y_pred = [False] -> Size= 26
IG_Abs

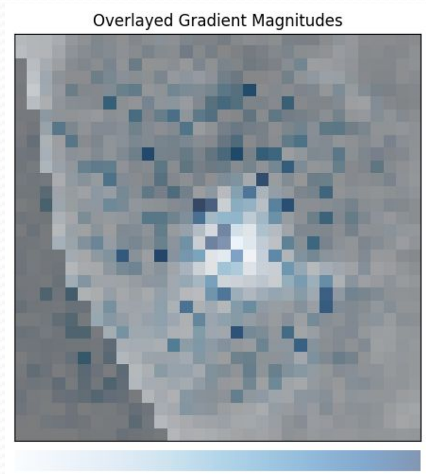


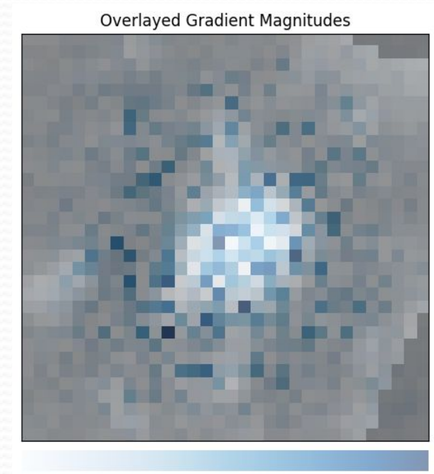
Y_true [True] & Y_pred = [False] -> Size= 26
IG_All

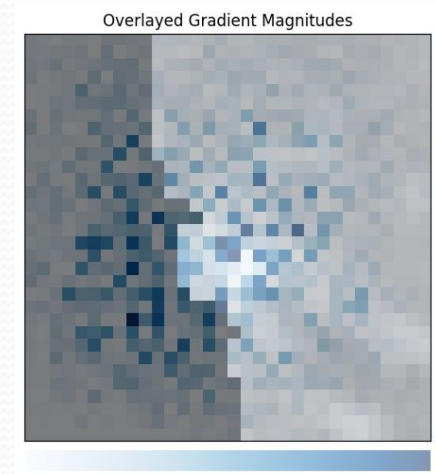


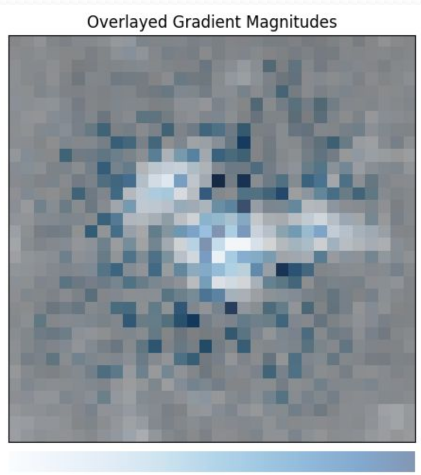
Y_true [True] & Y_pred = [False] -> Size= 26 SM

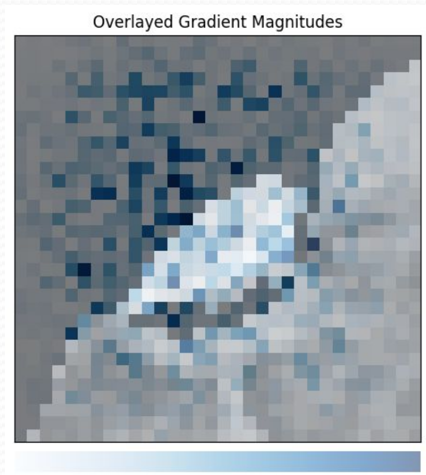


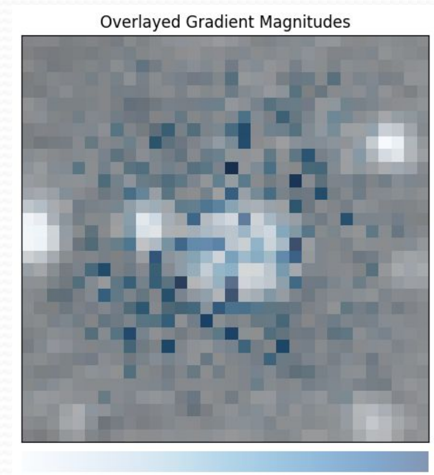


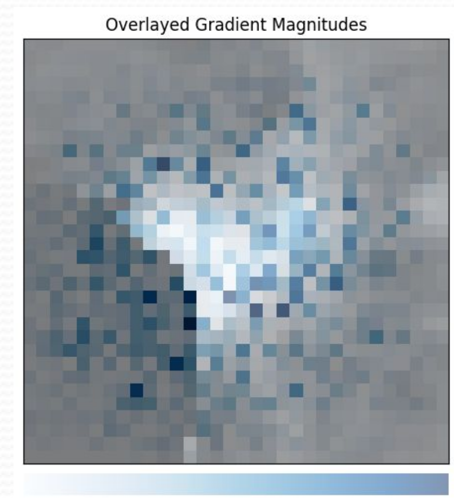




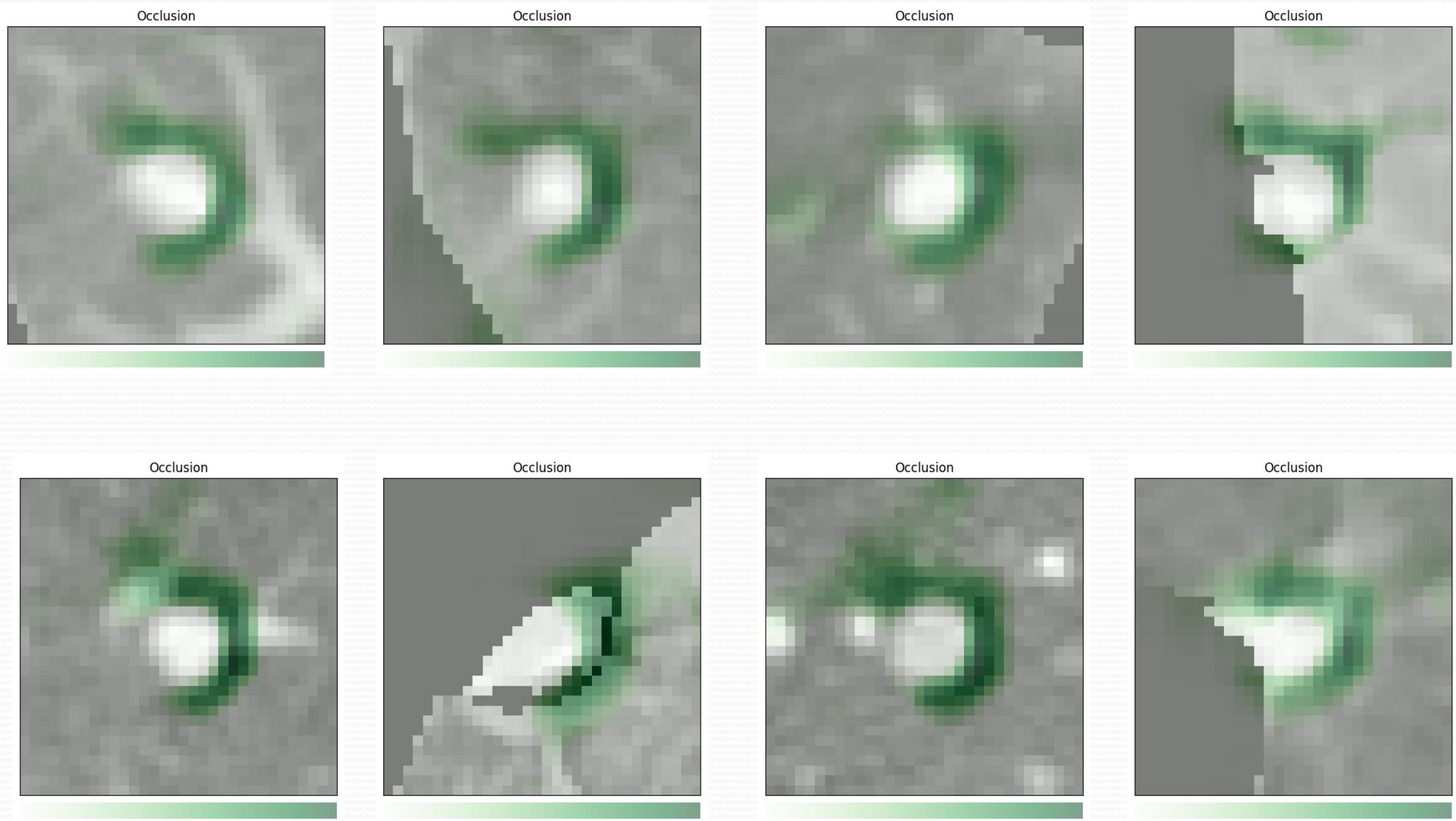




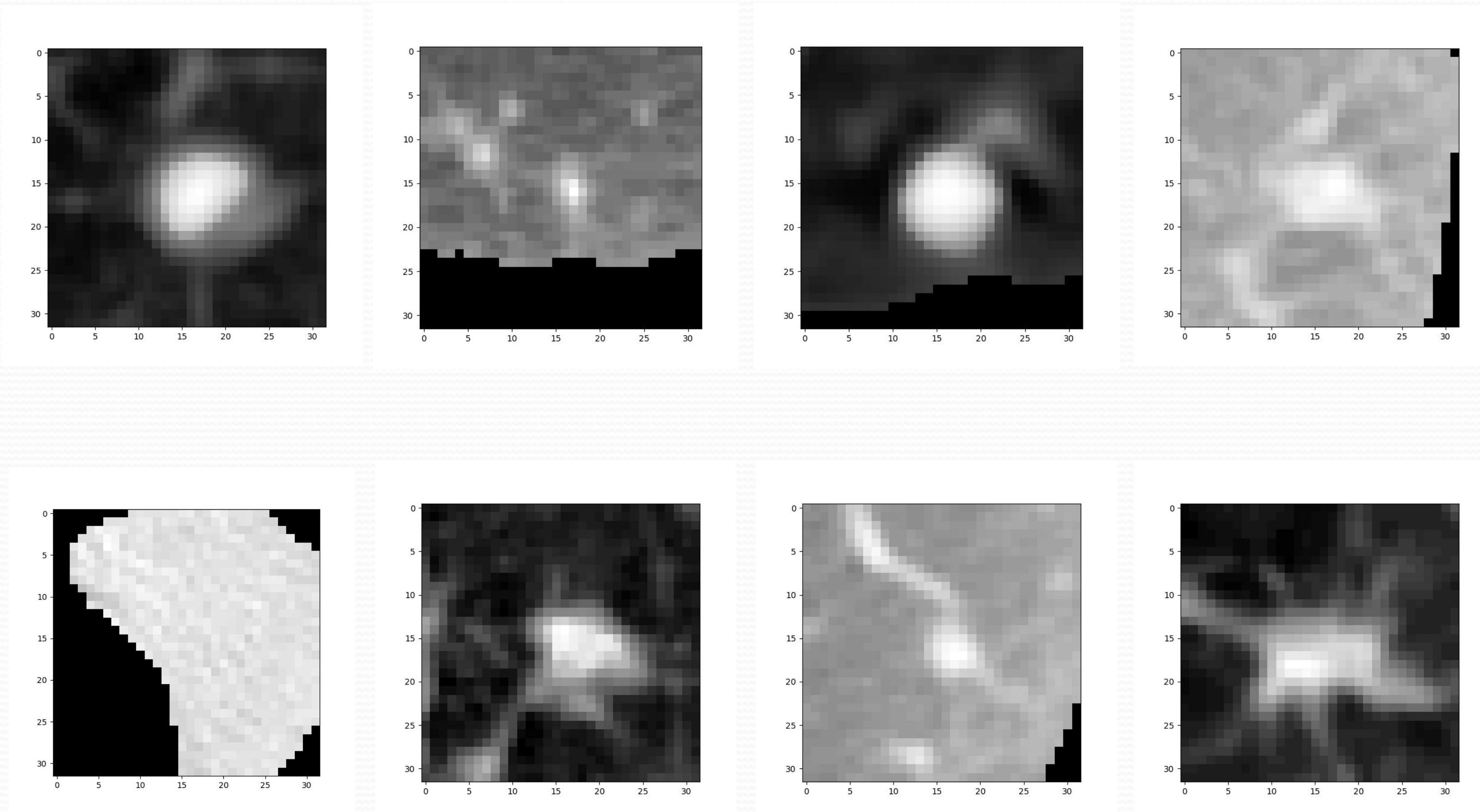




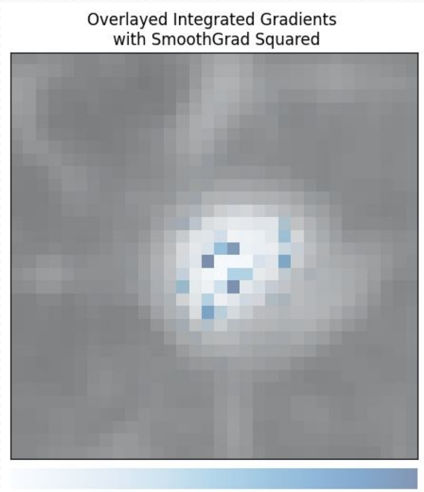
Y_true [True] & Y_pred = [False] -> Size= 26
OCC

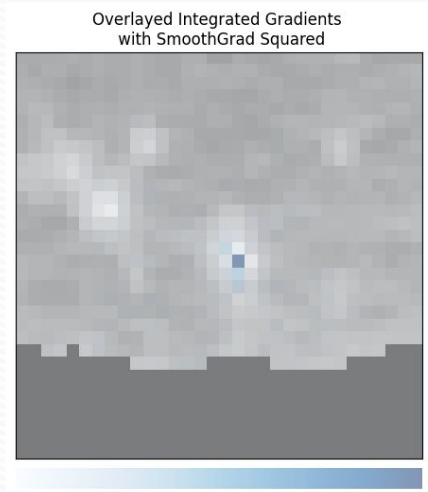


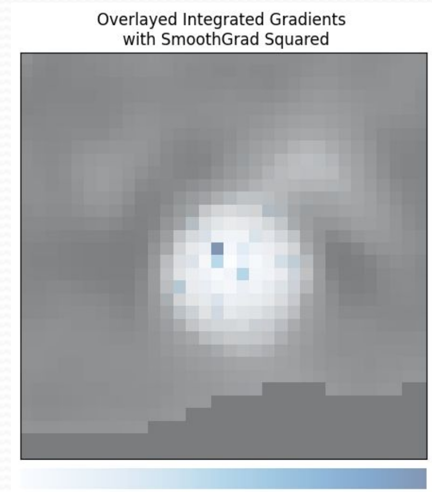
Y_true [False] & Y_pred = [True] -> Size= 30 Nodule

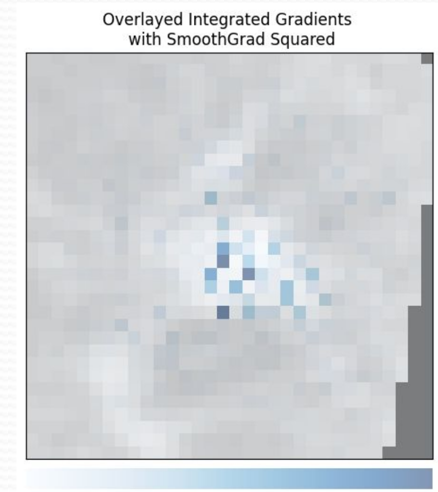


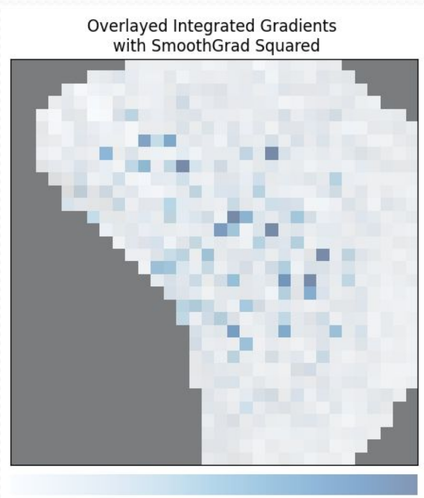
Y_true [False] & Y_pred = [True] -> Size= 30
IG_Abs

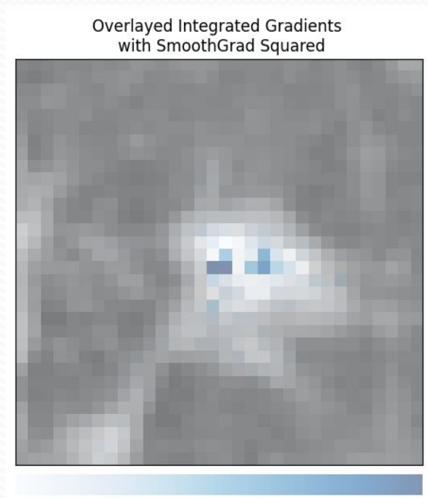


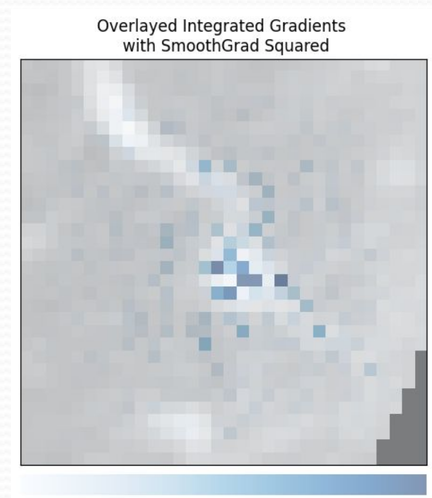


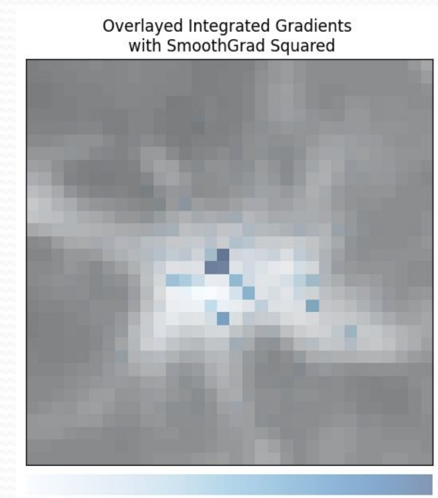




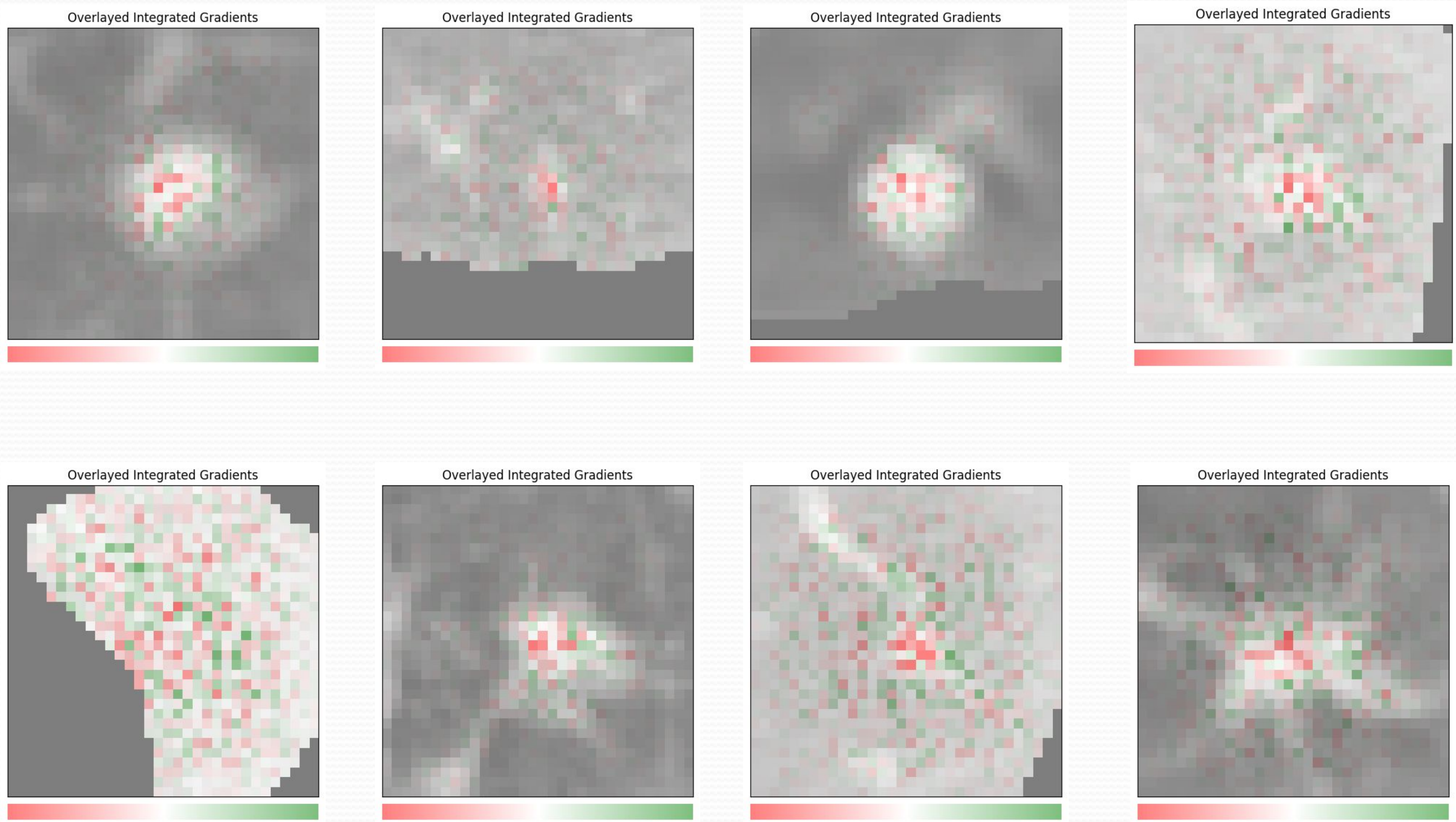




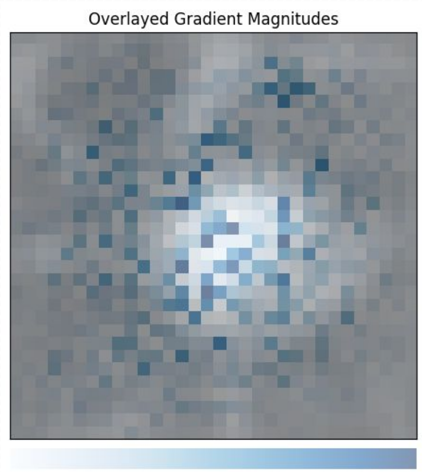


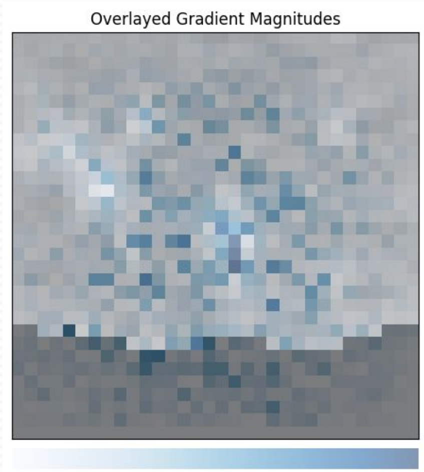


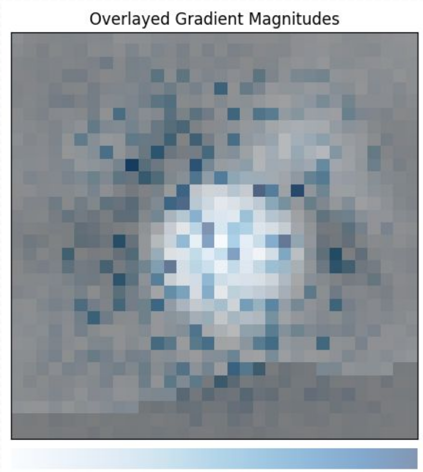
Y_true [False] & Y_pred = [True] -> Size= 30
IG_All

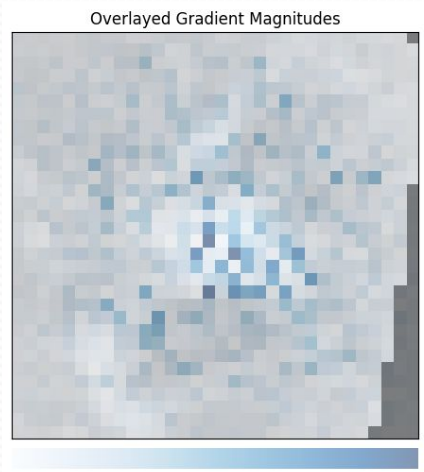


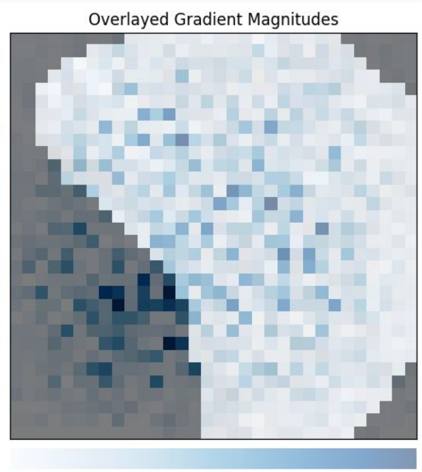
Y_true [False] & Y_pred = [True] -> Size= 30 SM

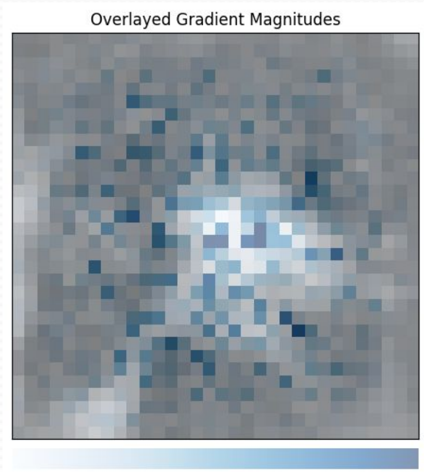


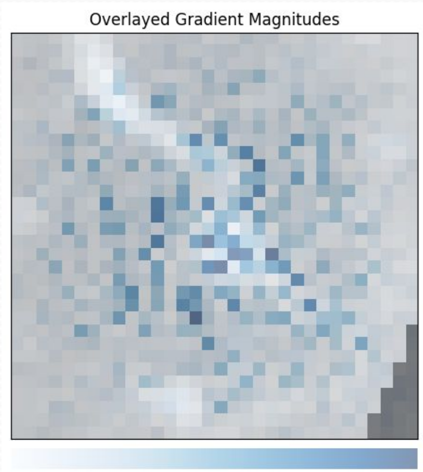


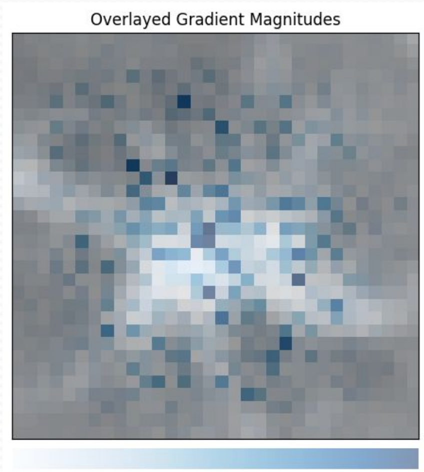




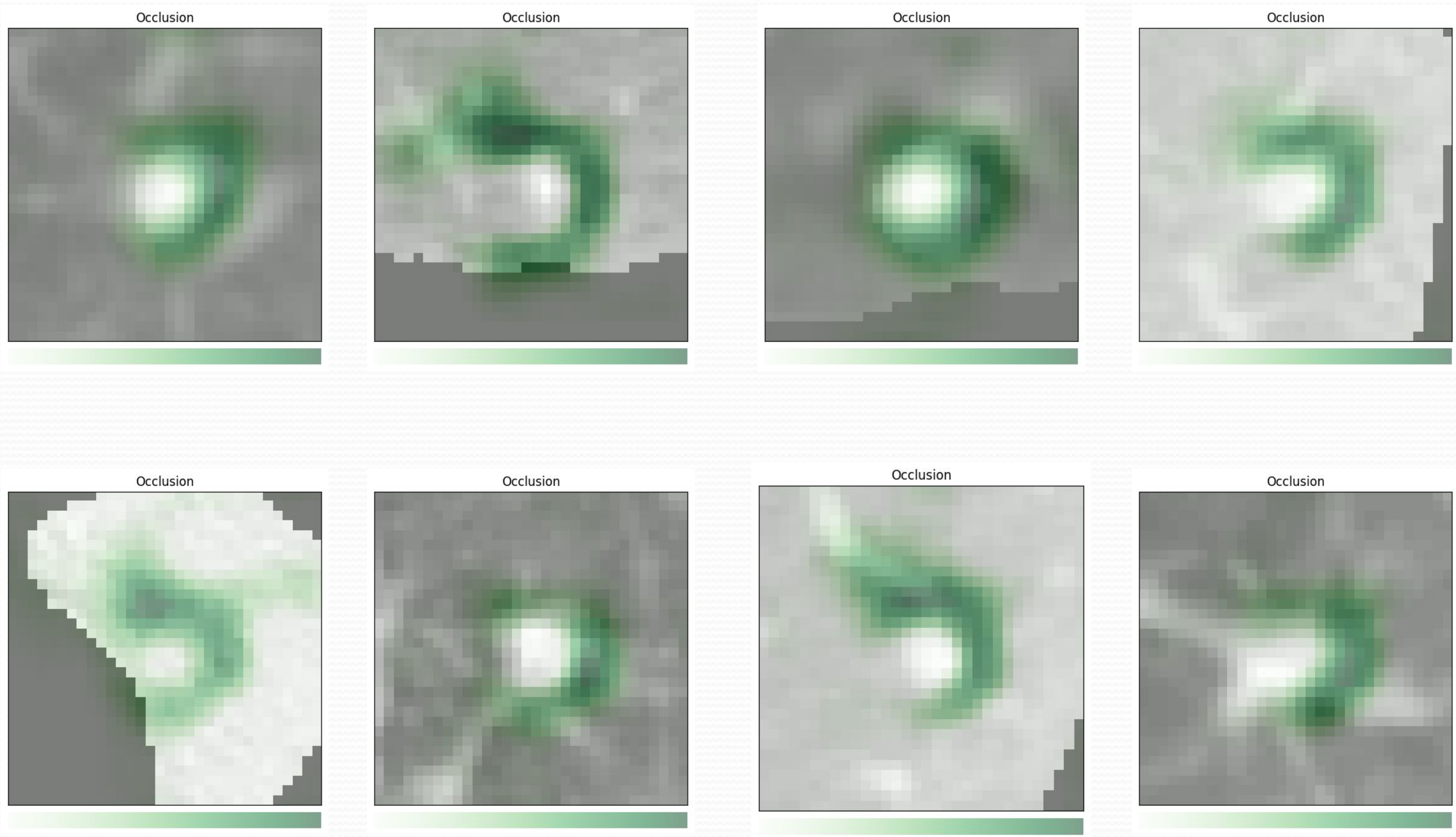




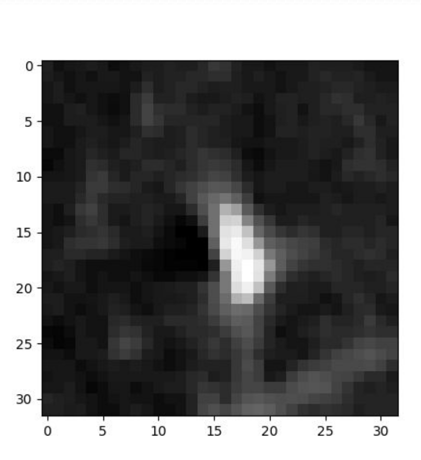
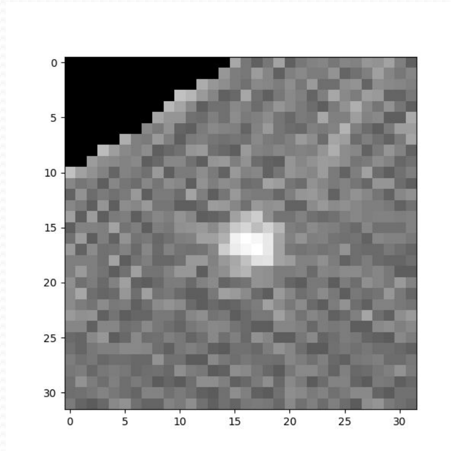
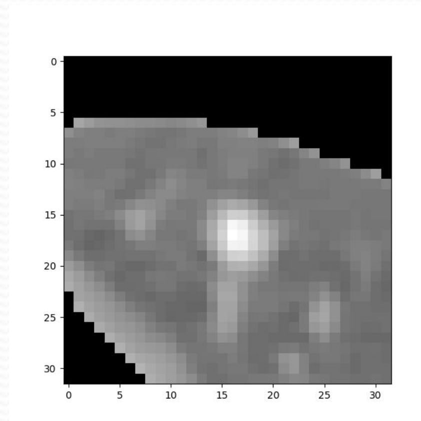
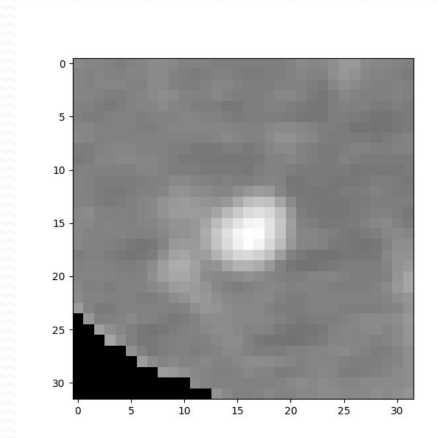
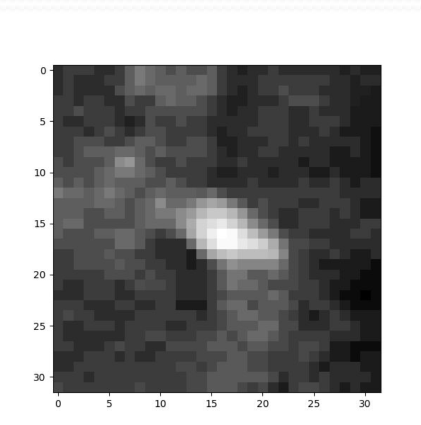
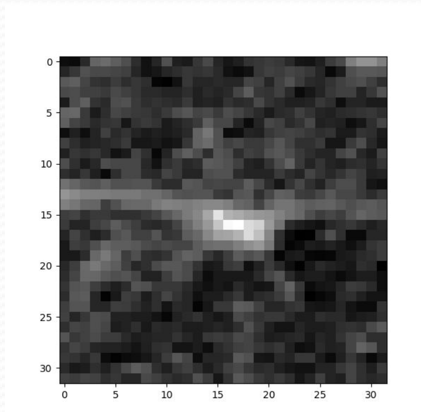
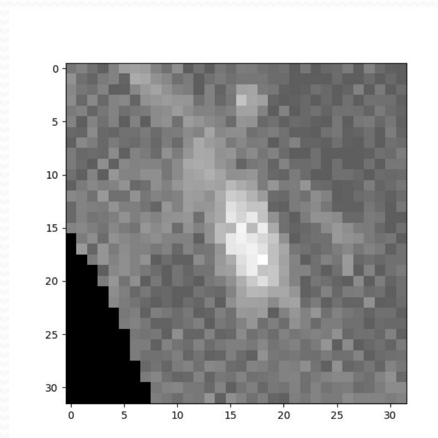
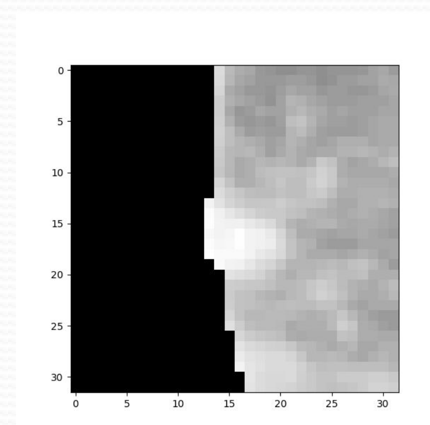




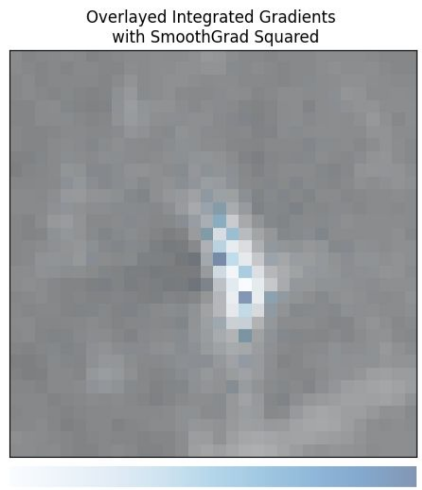
Y_true [False] & Y_pred = [True] -> Size= 30 OCC

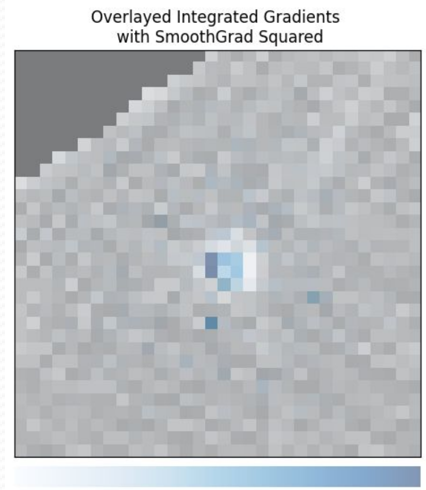


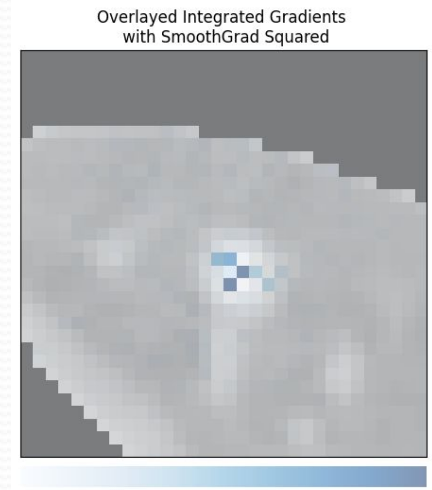
Y_true [False] & Y_pred = [False] -> Size= 109 Nodule

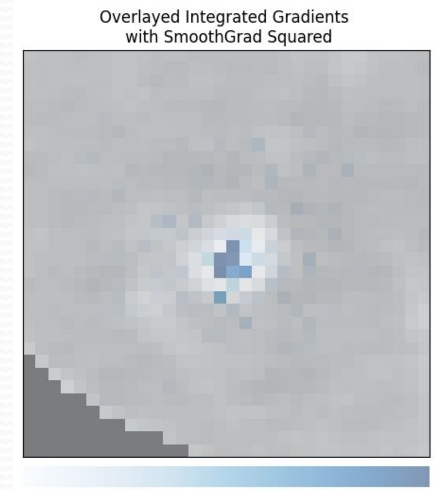









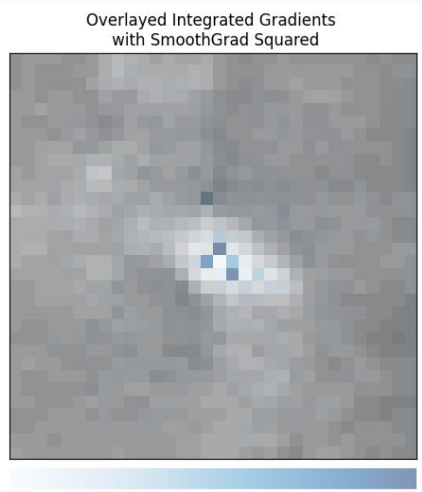
Y_true [False] & Y_pred = [False] -> Size= 109
IG_Abs

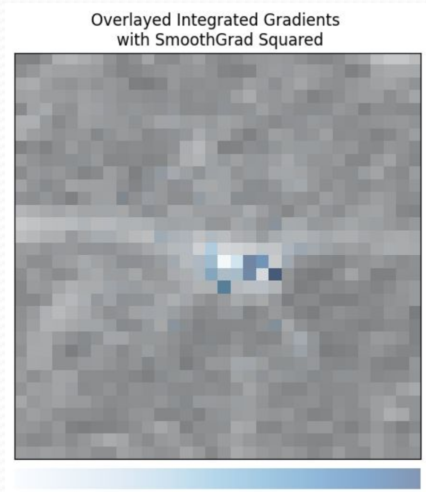


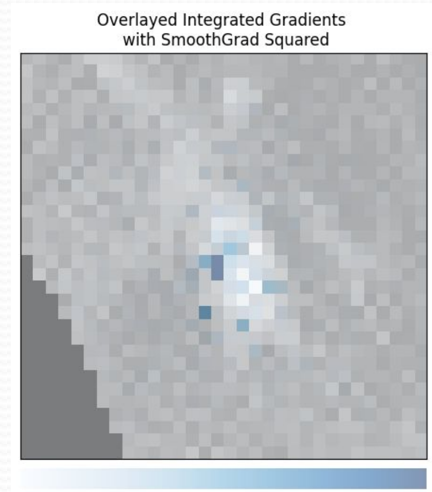


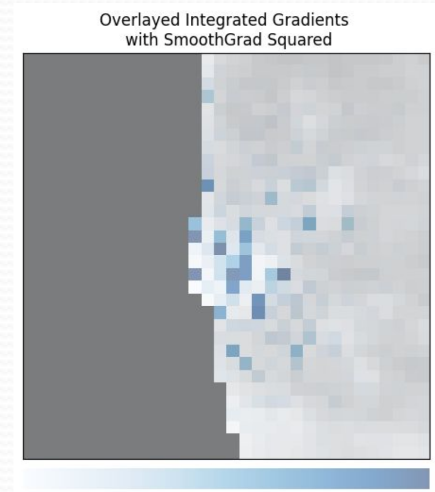




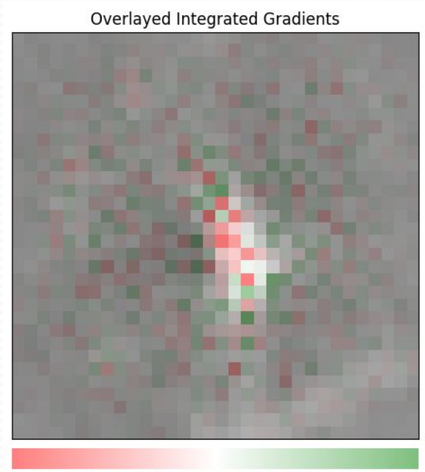


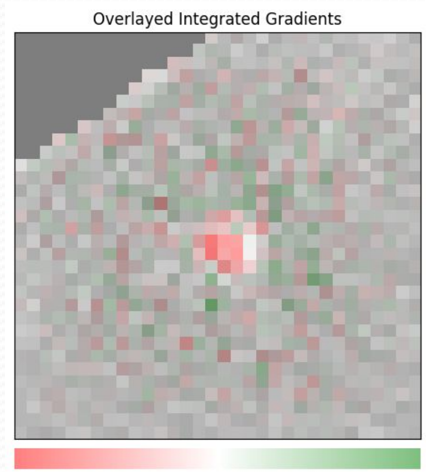


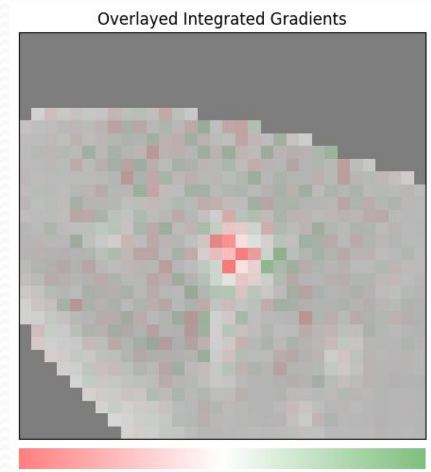


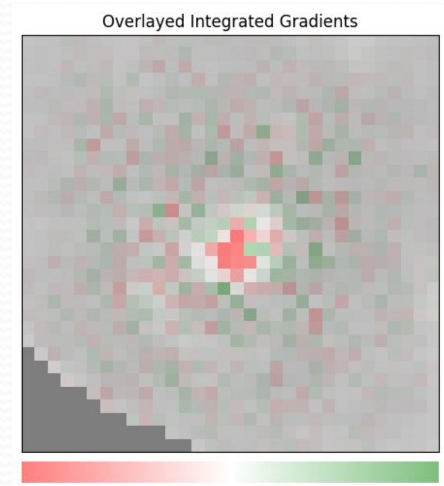


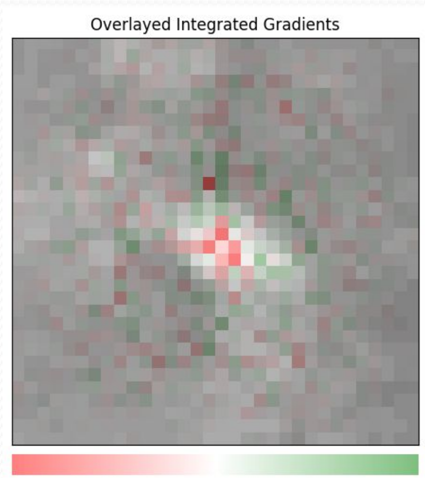
Y_true [False] & Y_pred = [False] -> Size= 109
IG_All

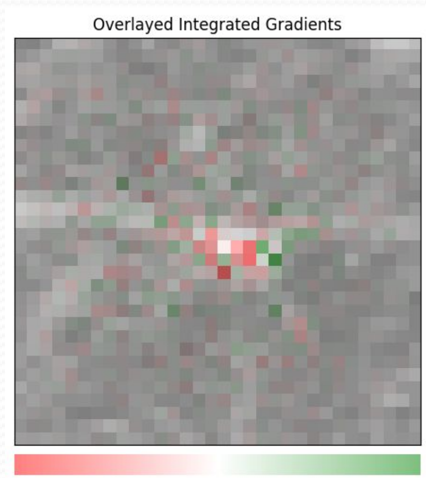


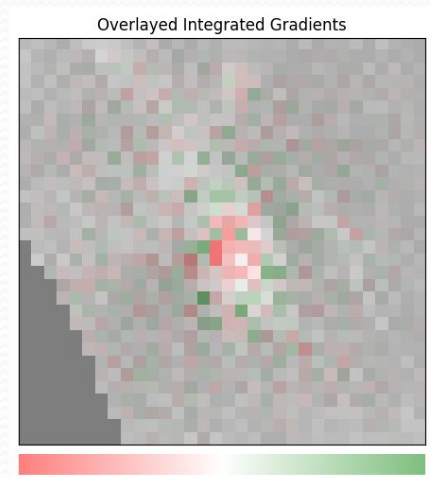


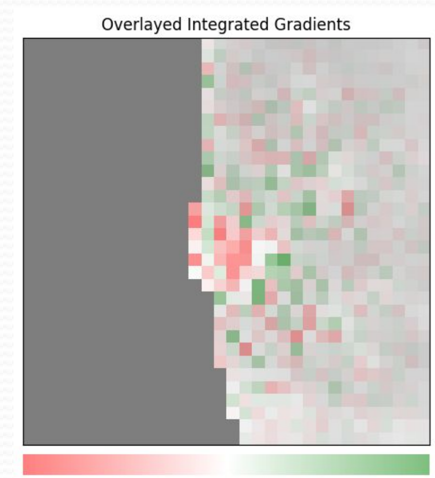




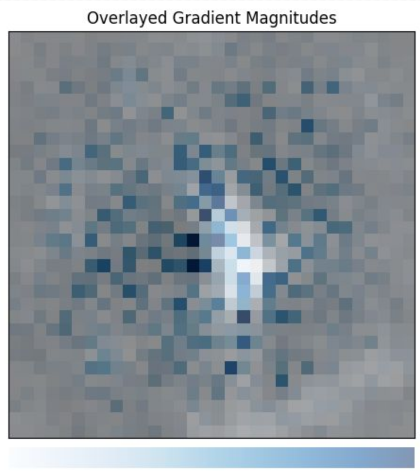


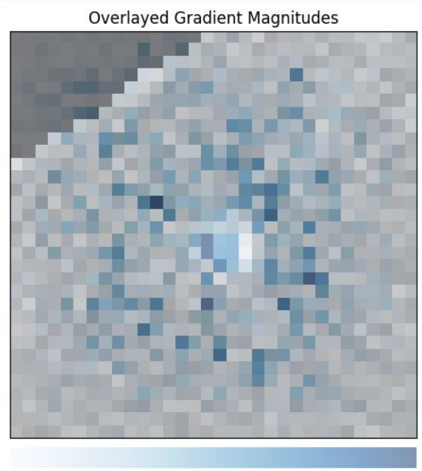


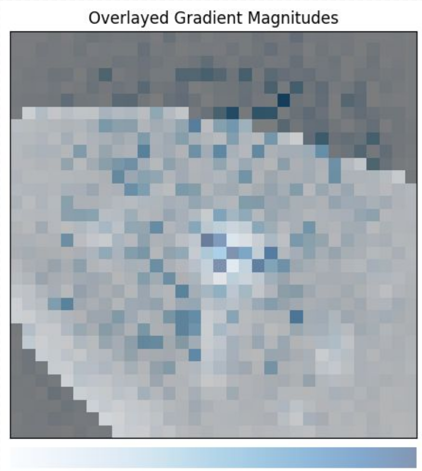


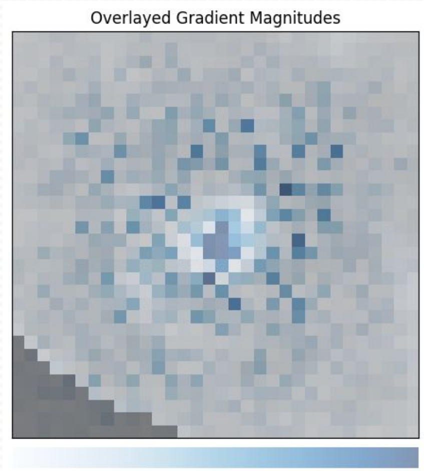


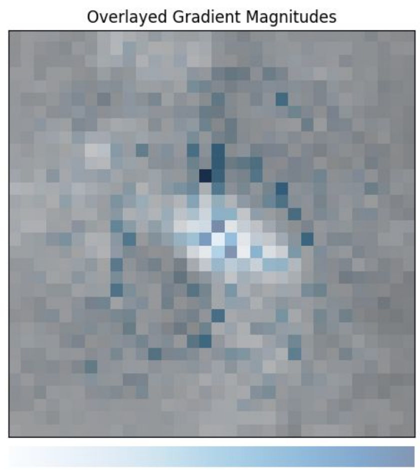
Y_true [False] & Y_pred = [False] -> Size= 109 SM

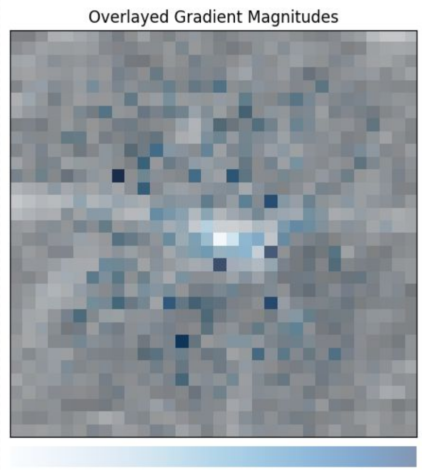


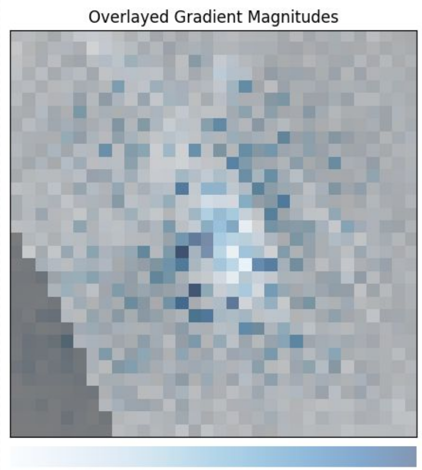


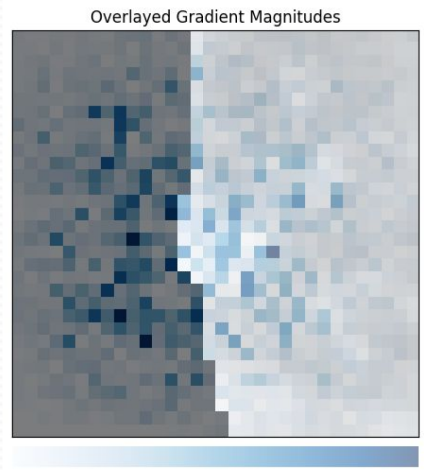




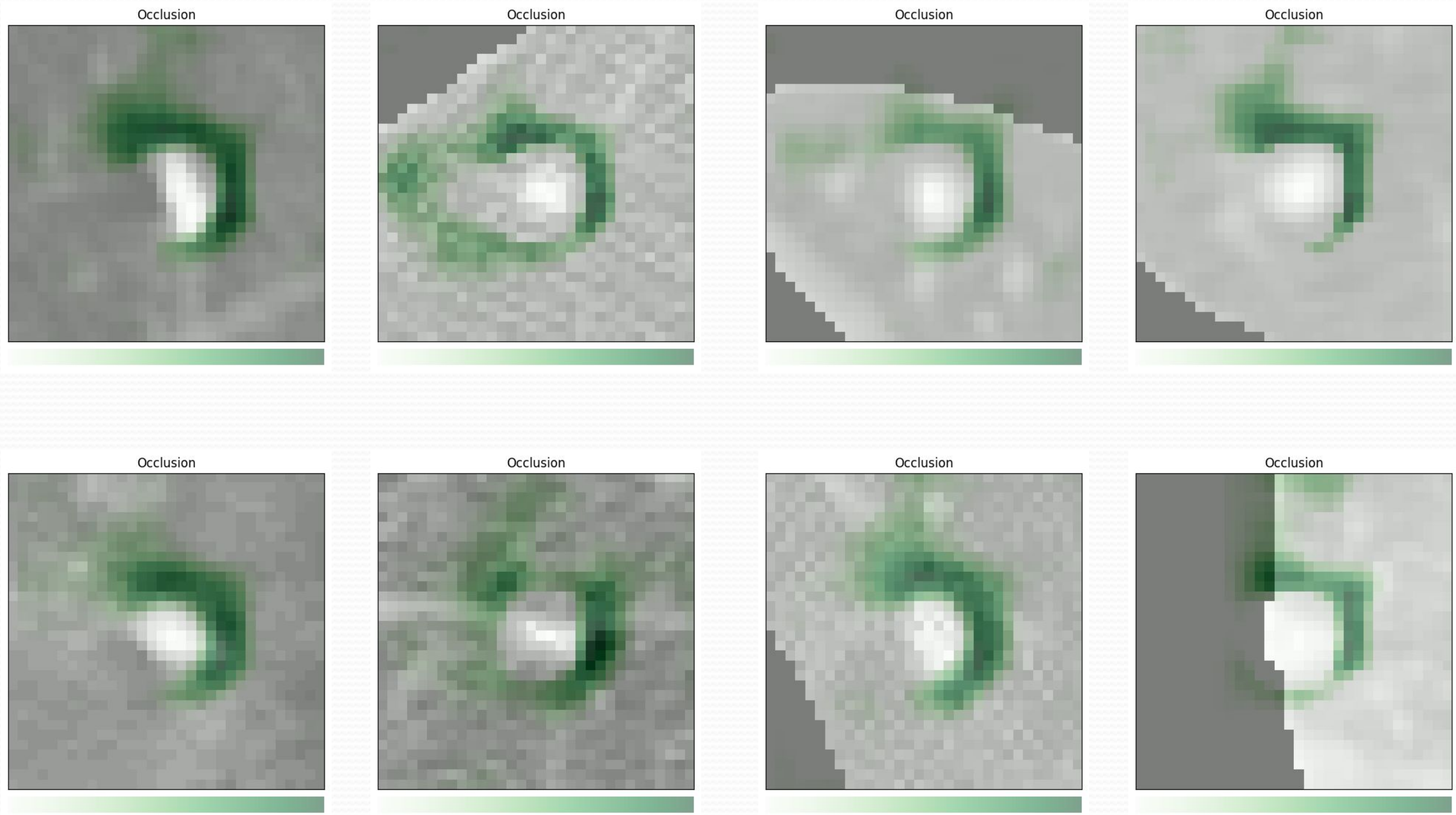






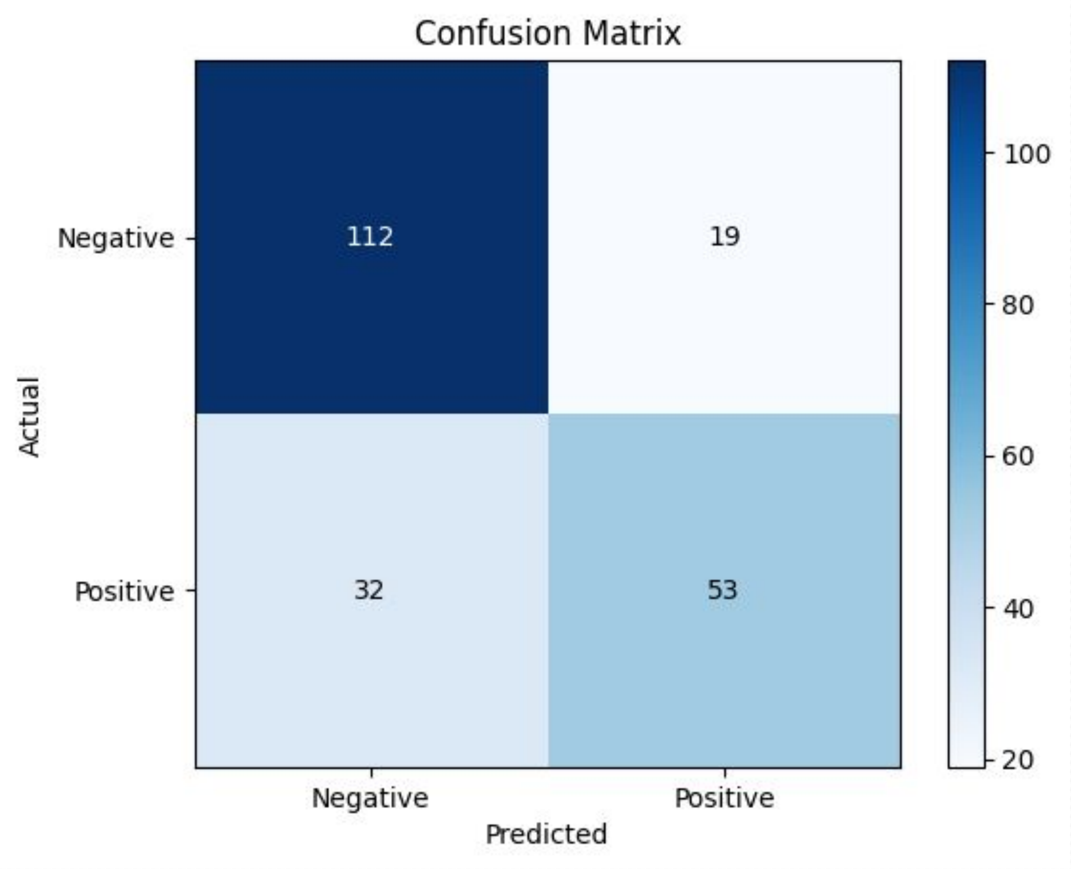


Y_true [False] & Y_pred = [False] -> Size= 109
OCC



Visual Results

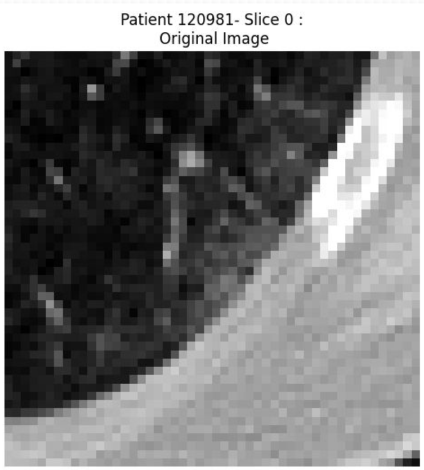
Resnet 3D

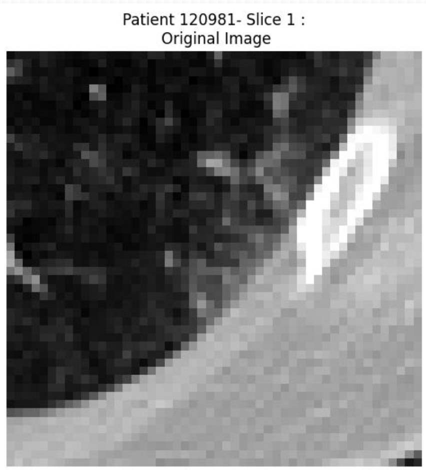


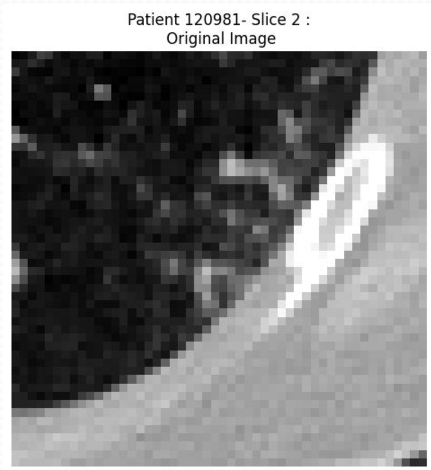
ACC: 76,39

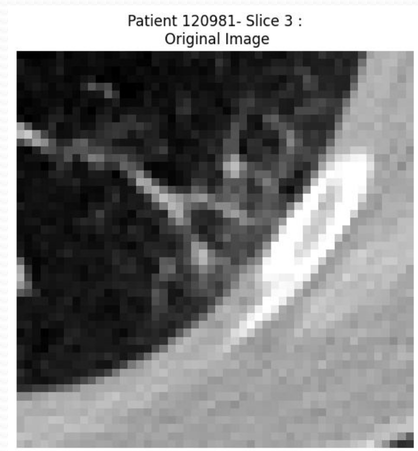
Y_true [True] & Y_pred = [True] -> Size= 53
Nodule

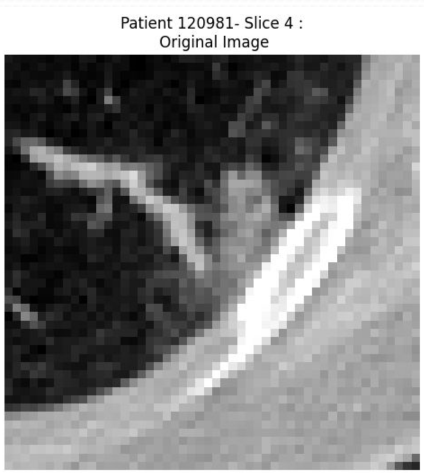
104960

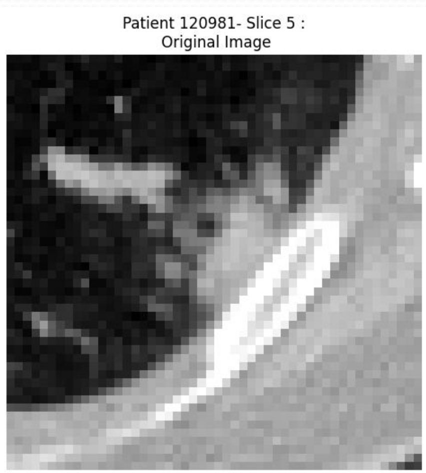


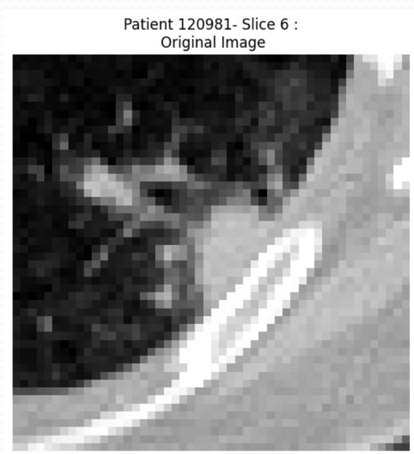






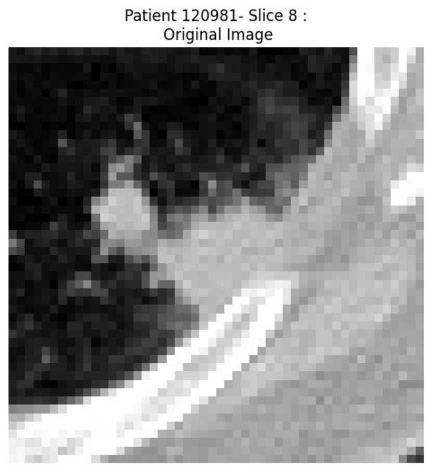


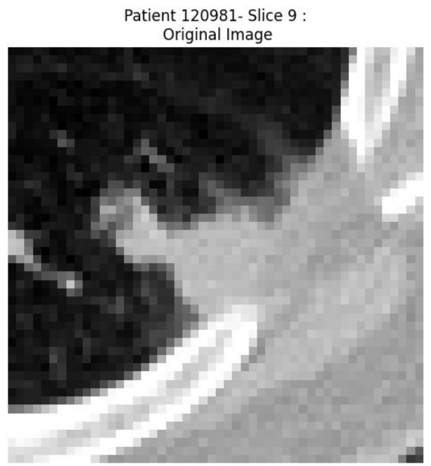


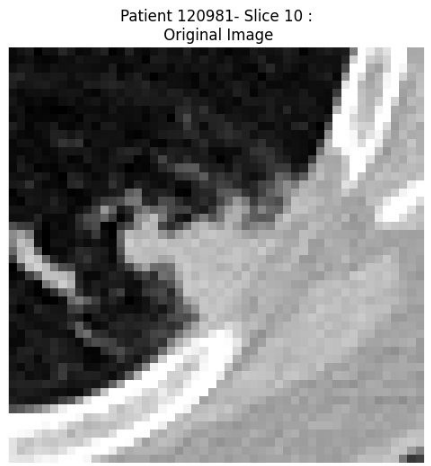


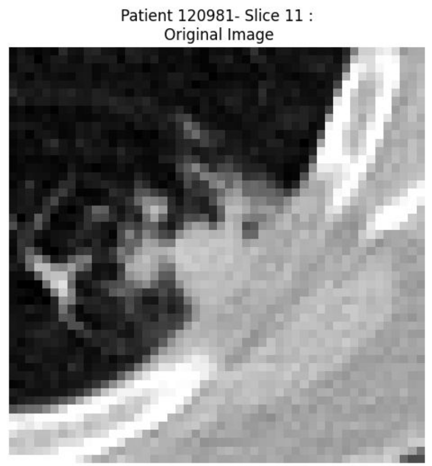
Y_true [True] & Y_pred = [True] -> Size= 53 Nodule

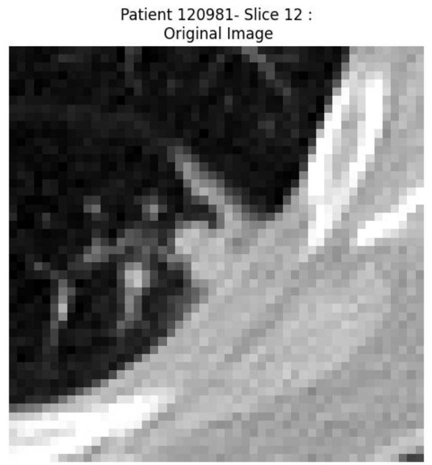
104960

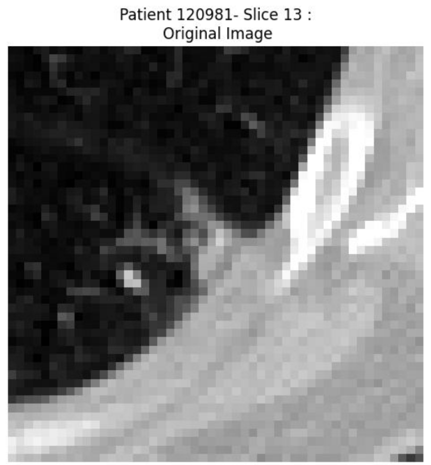


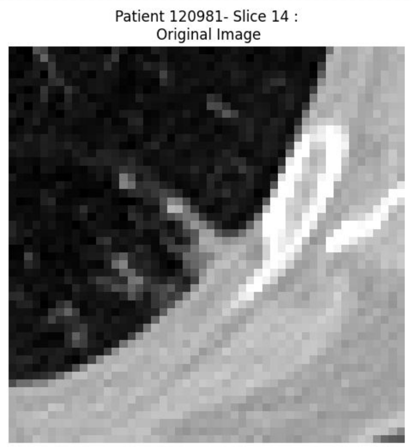


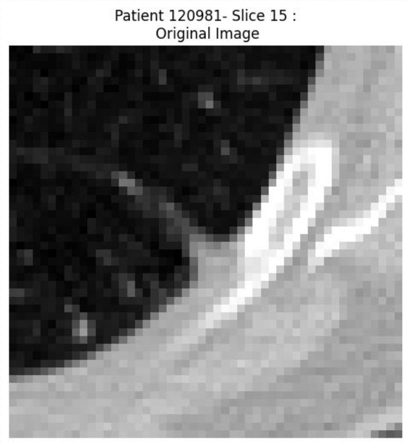






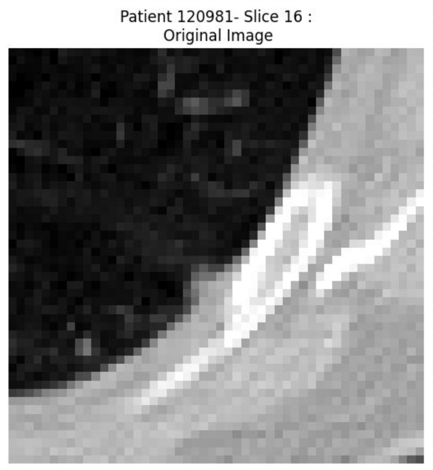


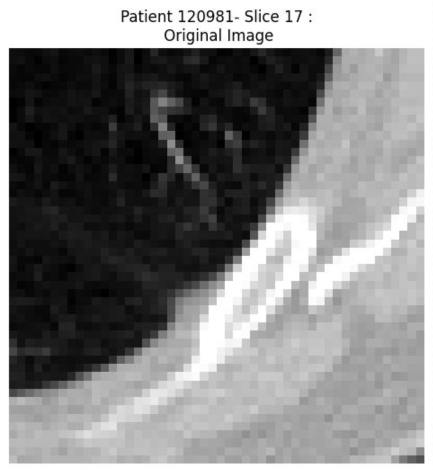


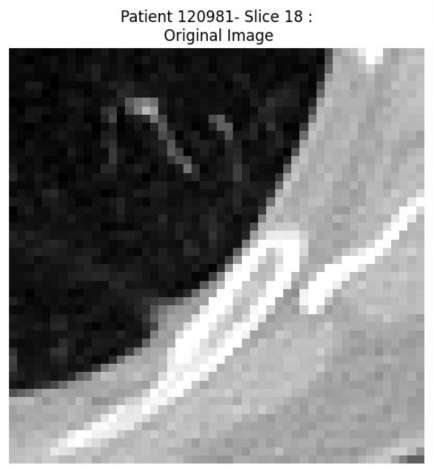


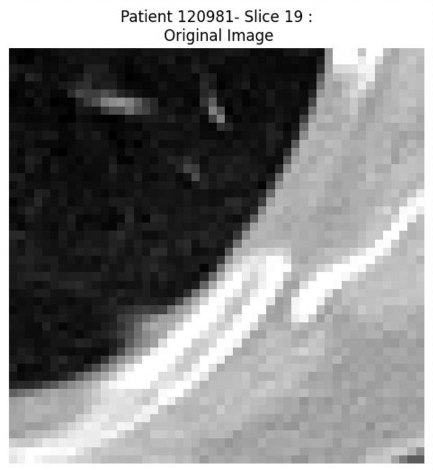
Y_true [True] & Y_pred = [True] -> Size= 53 Nodule

104960



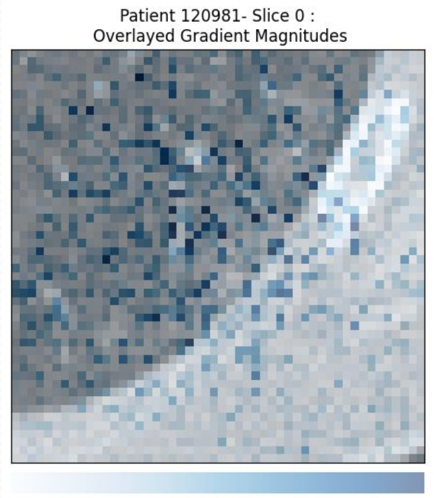


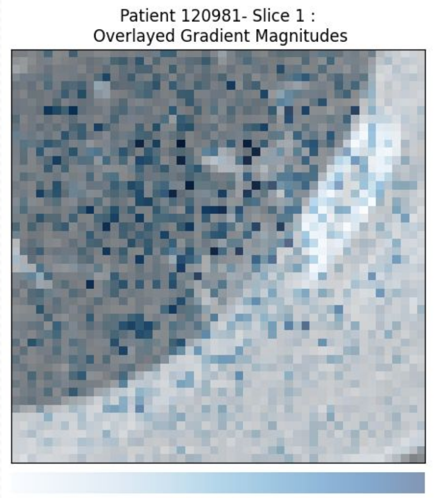


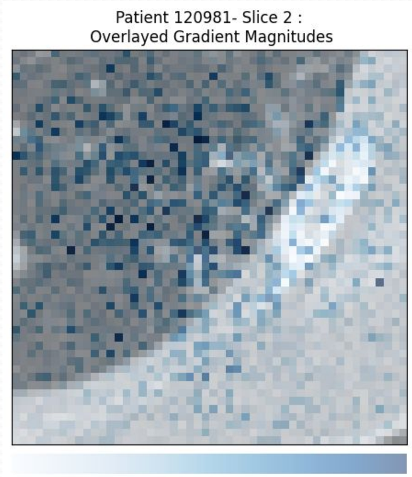


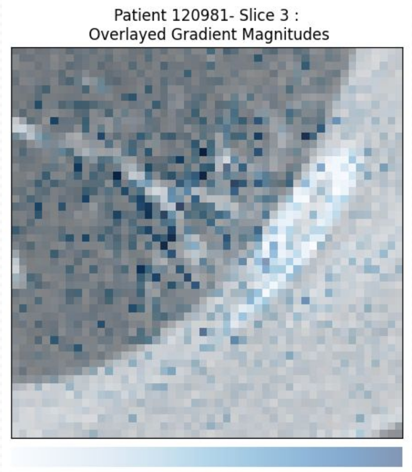
Y_true [True] & Y_pred = [True] -> Size= 53 SM

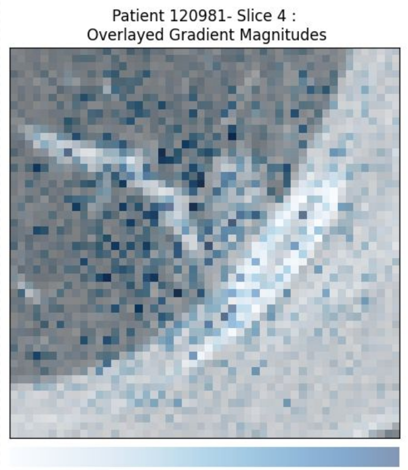
104960

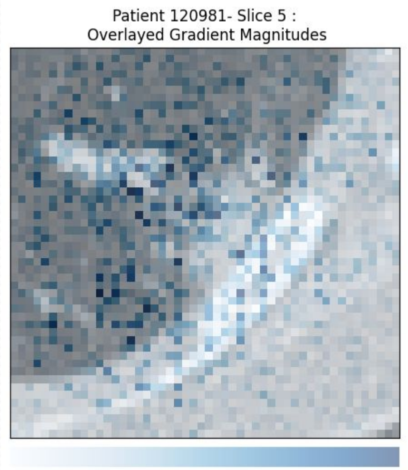


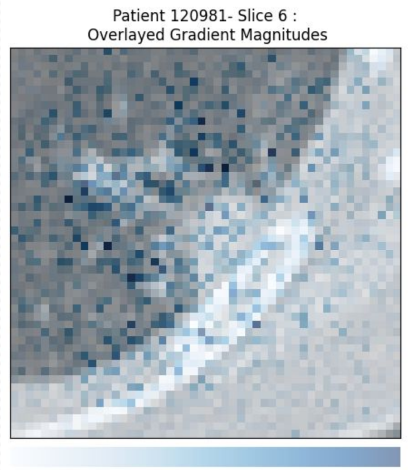


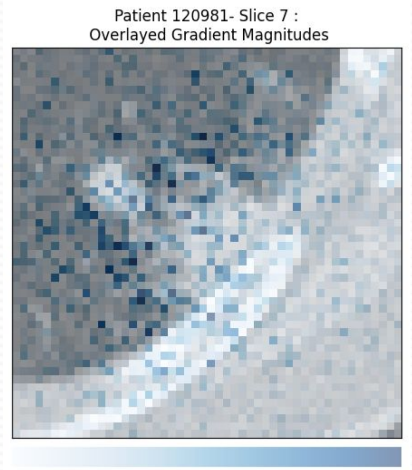








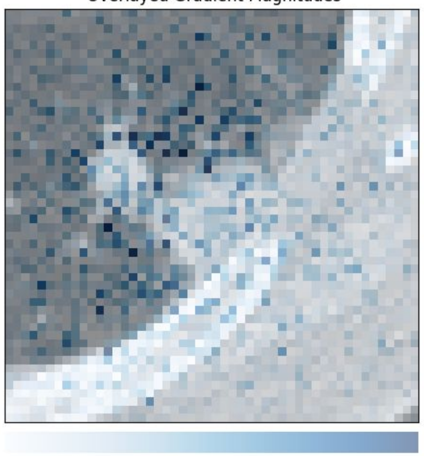




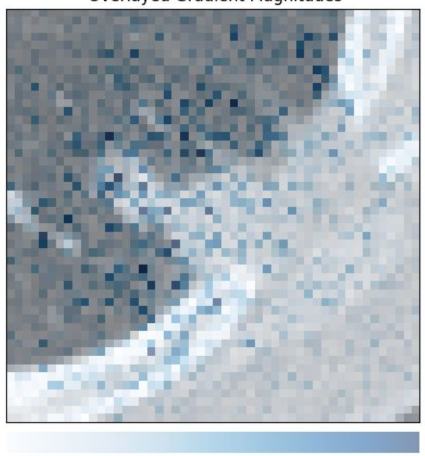
Y_true [True] & Y_pred = [True] -> Size= 53 SM

104960

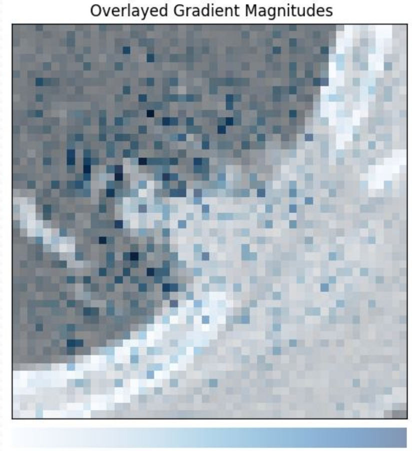
Patient 120981- Slice 8 :
Overlayed Gradient Magnitudes



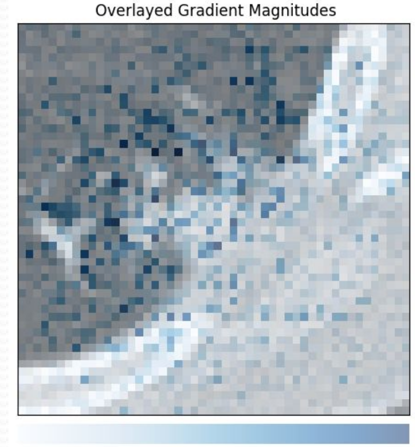
Patient 120981- Slice 9 :
Overlayed Gradient Magnitudes



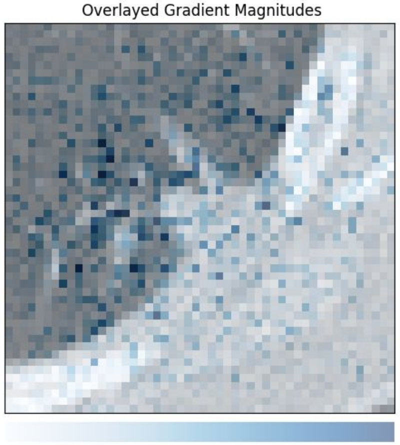
Patient 120981- Slice 10 :
Overlayed Gradient Magnitudes



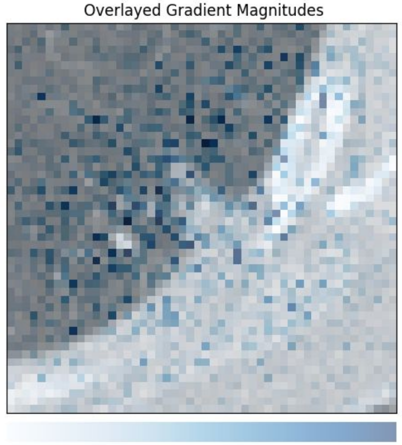
Patient 120981- Slice 11 :
Overlayed Gradient Magnitudes



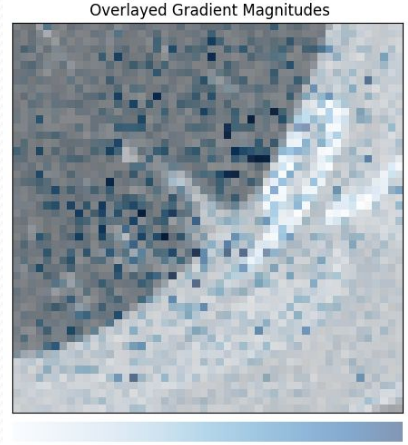
Patient 120981- Slice 12 :
Overlayed Gradient Magnitudes



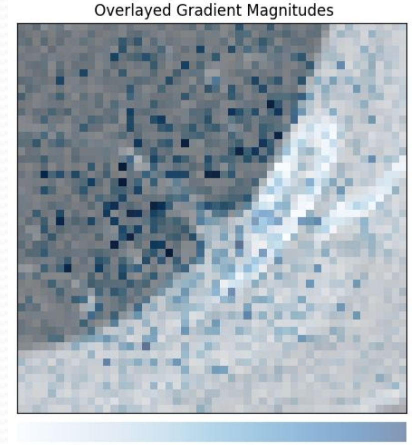
Patient 120981- Slice 13 :
Overlayed Gradient Magnitudes



Patient 120981- Slice 14 :
Overlayed Gradient Magnitudes

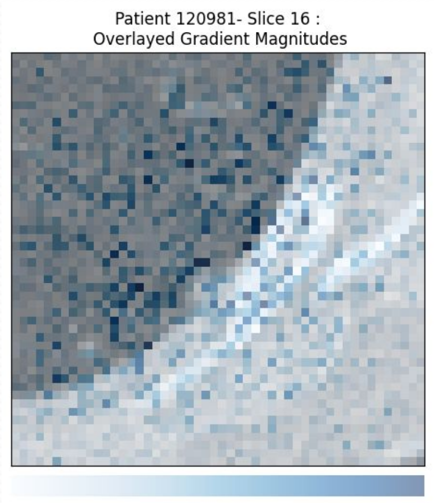


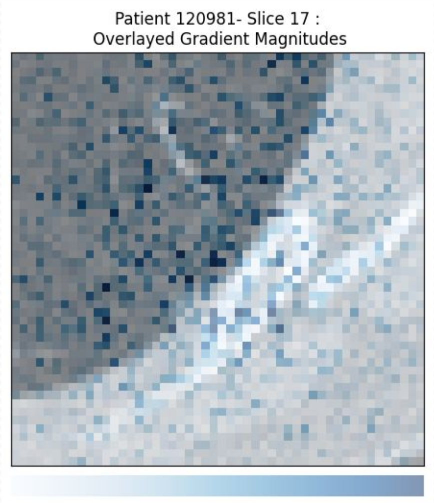
Patient 120981- Slice 15 :
Overlayed Gradient Magnitudes

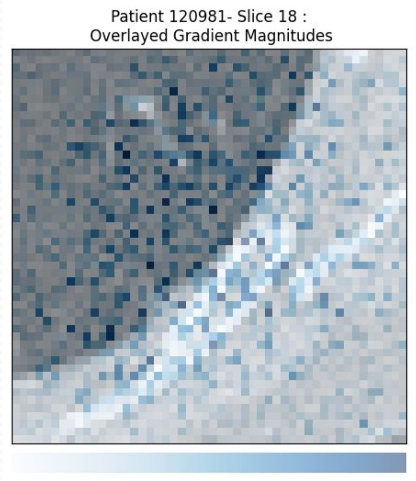


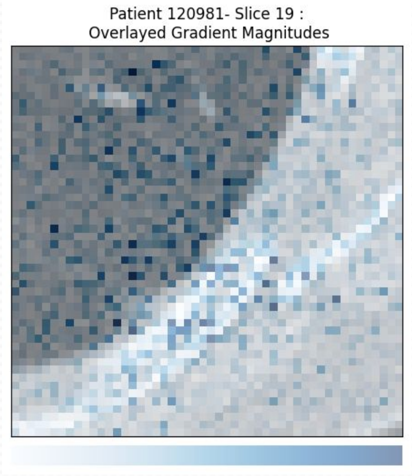
Y_true [True] & Y_pred = [True] -> Size= 53 SM

104960



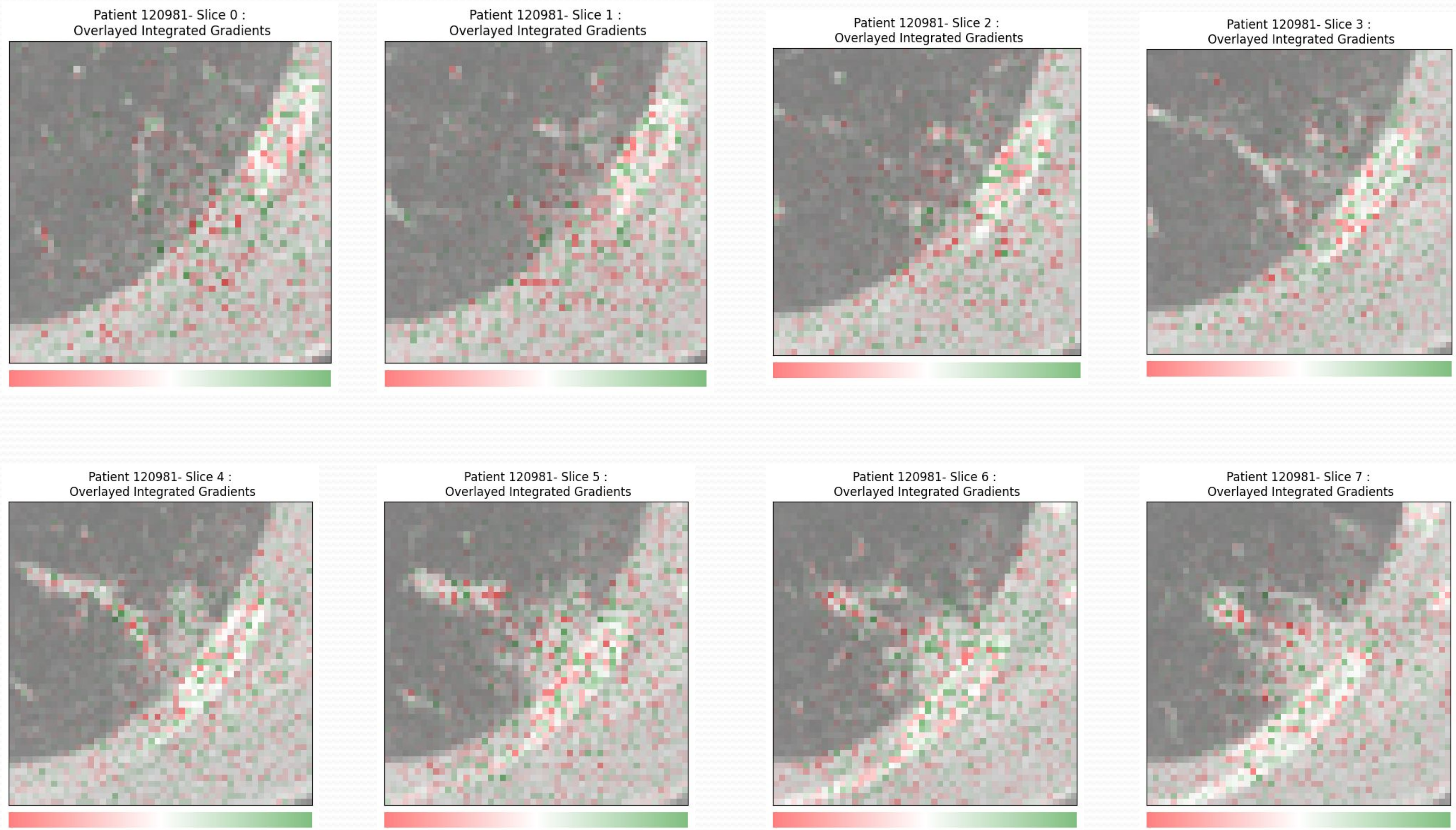






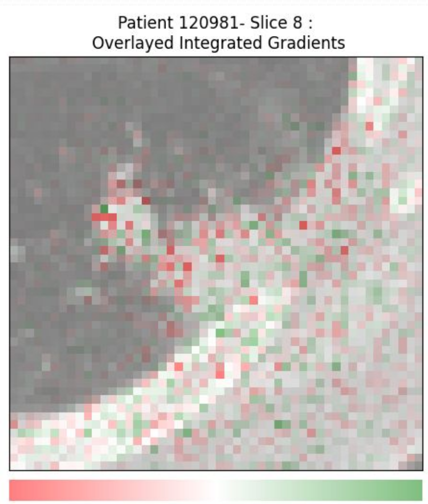
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_All

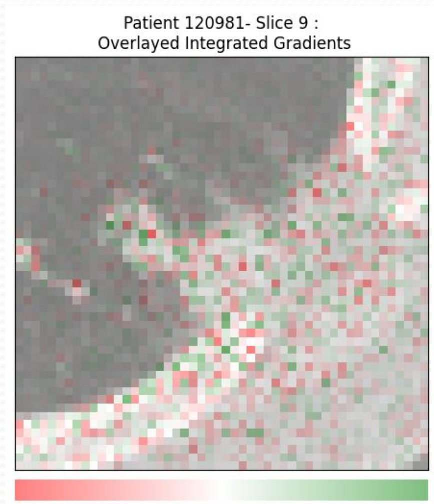
104960

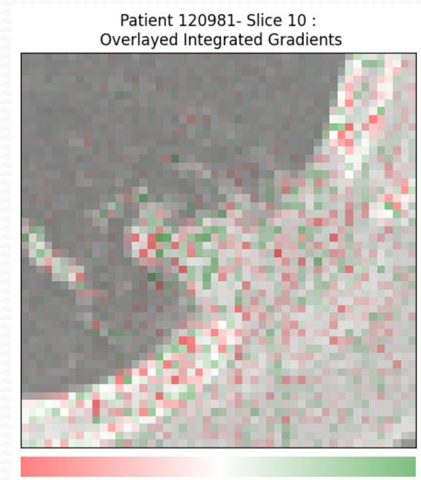


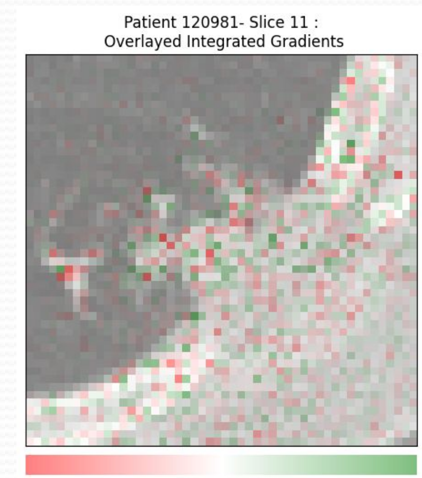
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_All

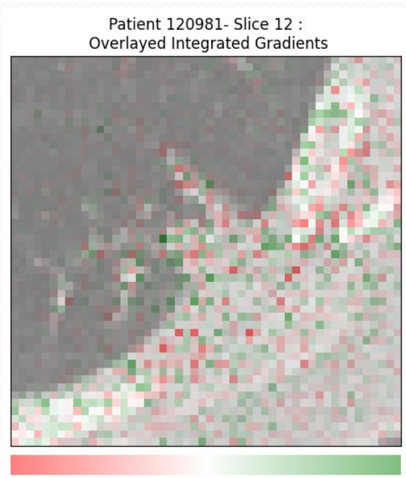
104960

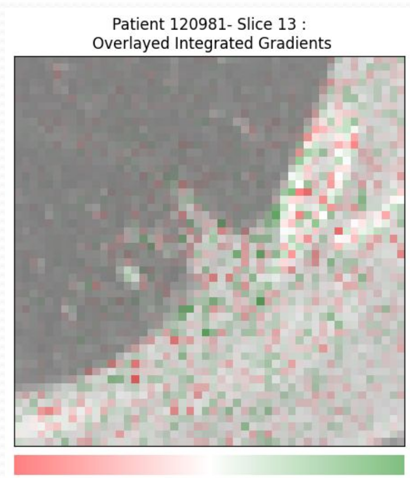


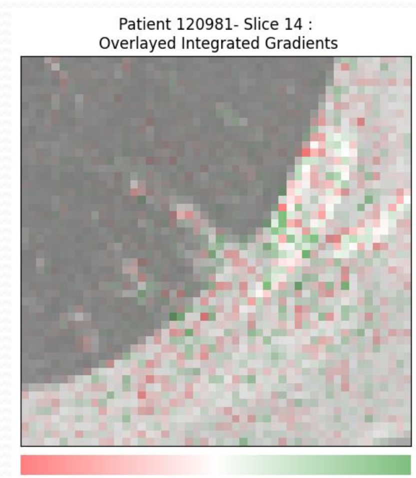


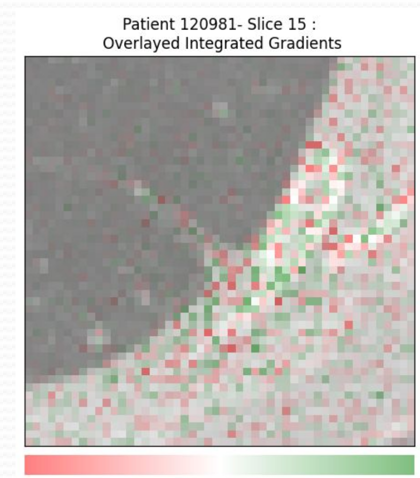






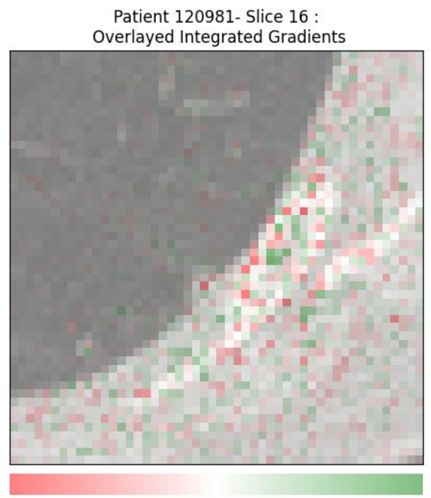


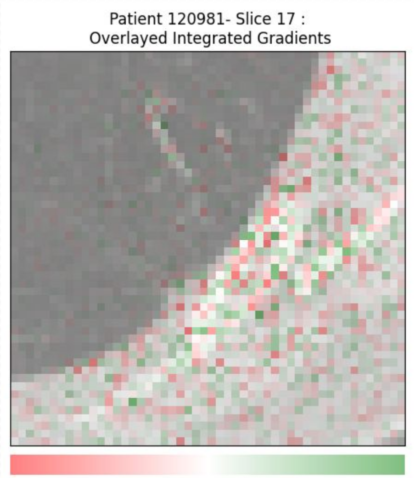


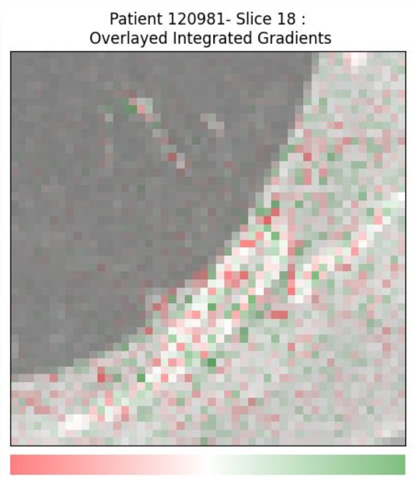


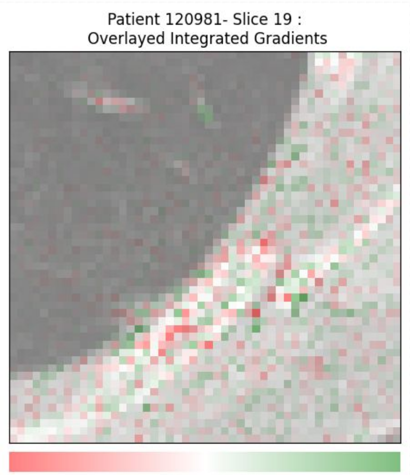
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_All

104960



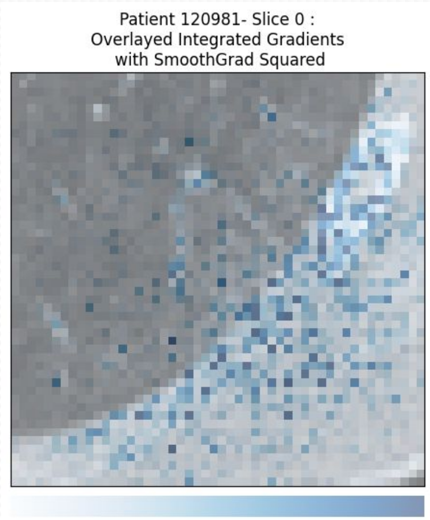


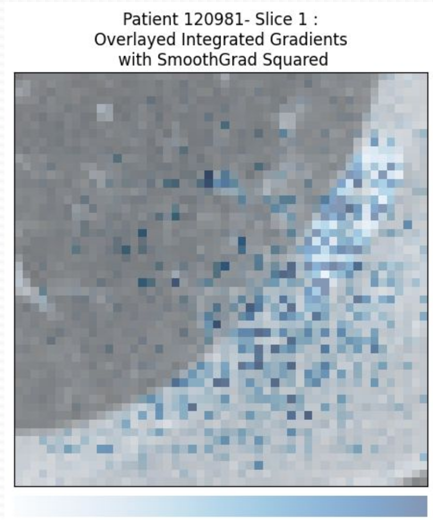


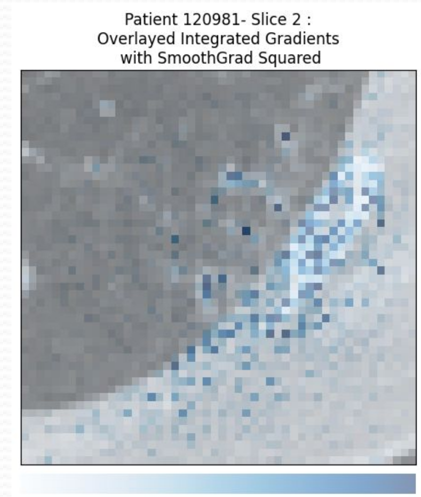


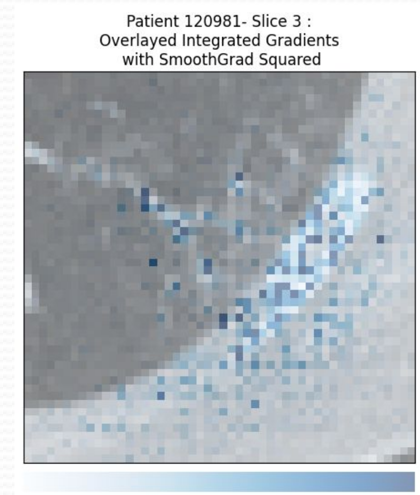
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_Abs

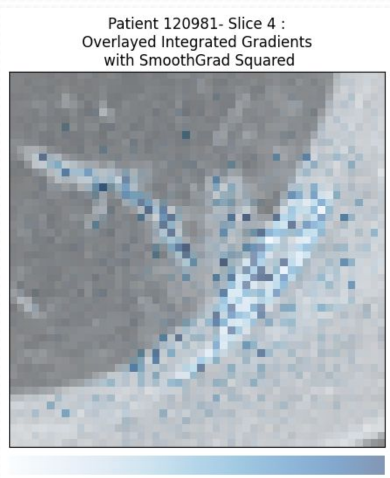
104960

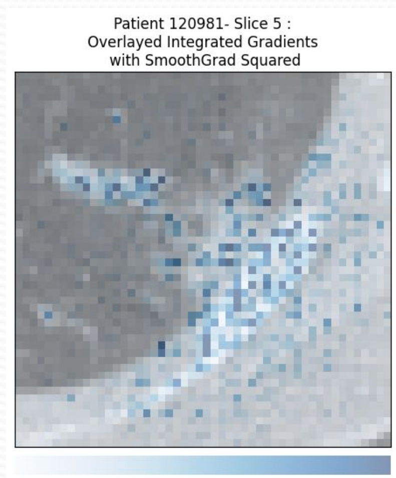


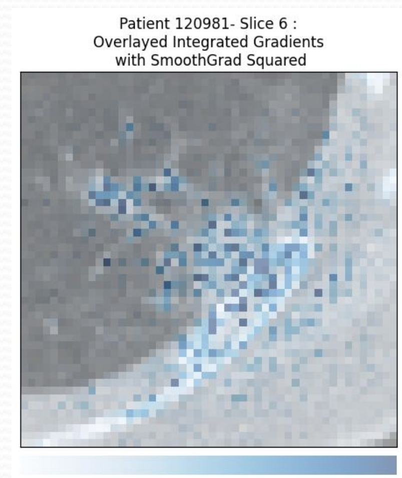


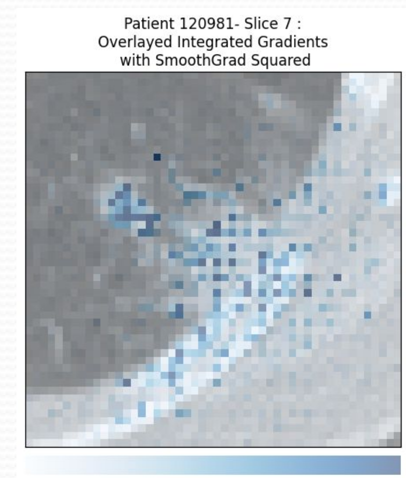








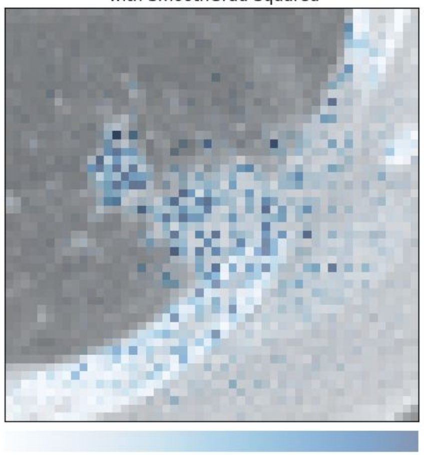




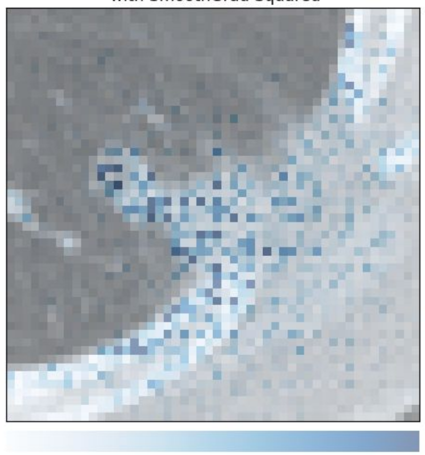
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_Abs

104960

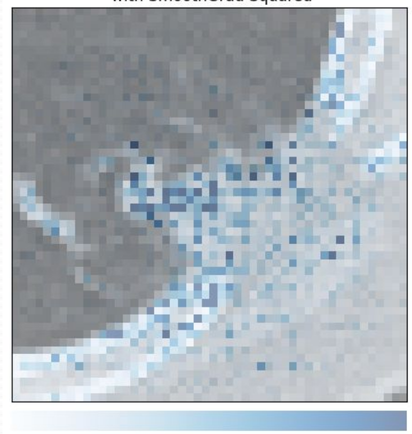
Patient 120981- Slice 8 :
Overlayed Integrated Gradients
with SmoothGrad Squared



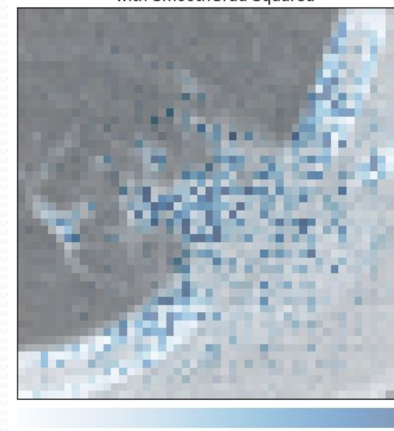
Patient 120981- Slice 9 :
Overlayed Integrated Gradients
with SmoothGrad Squared



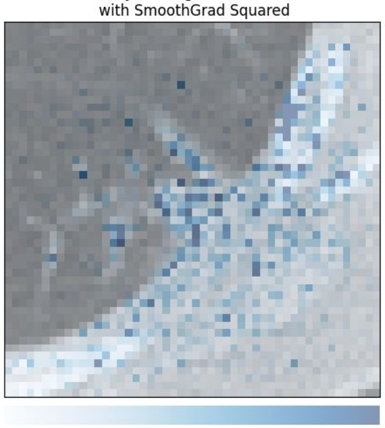
Patient 120981- Slice 10 :
Overlayed Integrated Gradients
with SmoothGrad Squared



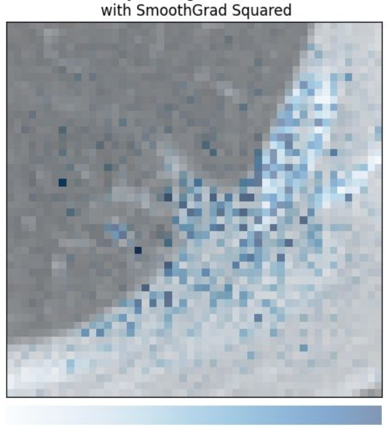
Patient 120981- Slice 11 :
Overlayed Integrated Gradients
with SmoothGrad Squared



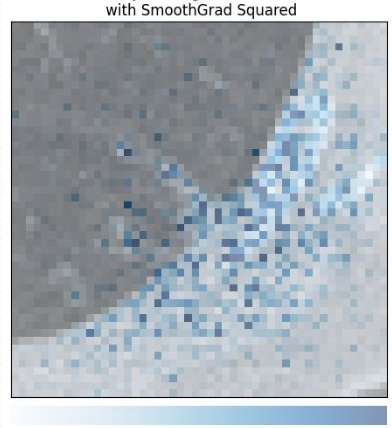
Patient 120981- Slice 12 :
Overlayed Integrated Gradients
with SmoothGrad Squared



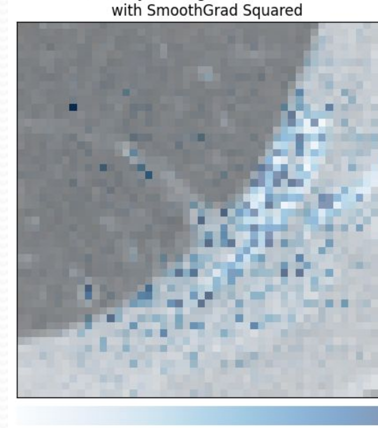
Patient 120981- Slice 13 :
Overlayed Integrated Gradients
with SmoothGrad Squared



Patient 120981- Slice 14 :
Overlayed Integrated Gradients
with SmoothGrad Squared

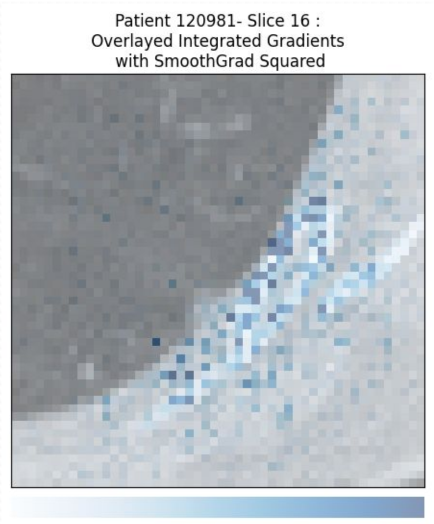


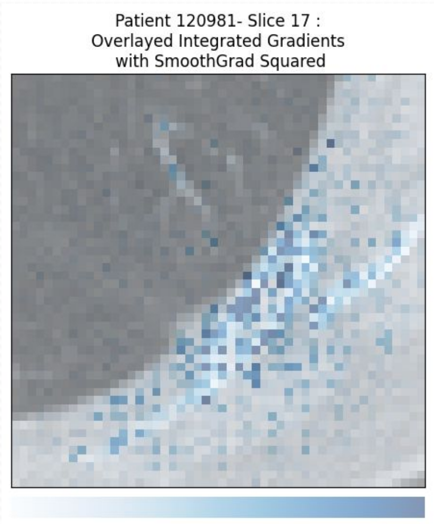
Patient 120981- Slice 15 :
Overlayed Integrated Gradients
with SmoothGrad Squared

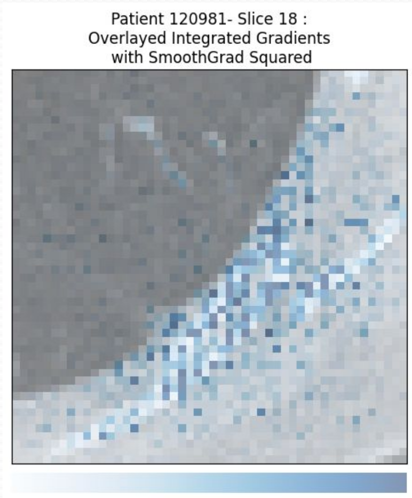


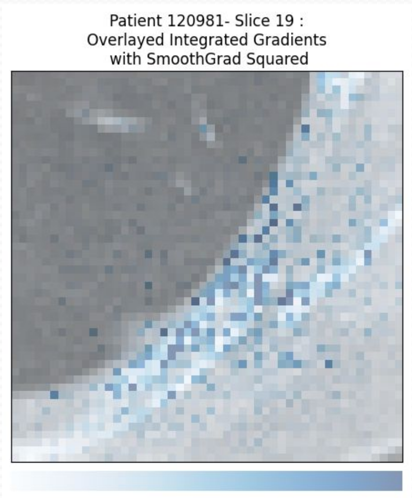
Y_true [True] & Y_pred = [True] -> Size= 53
Ig_Abs

104960



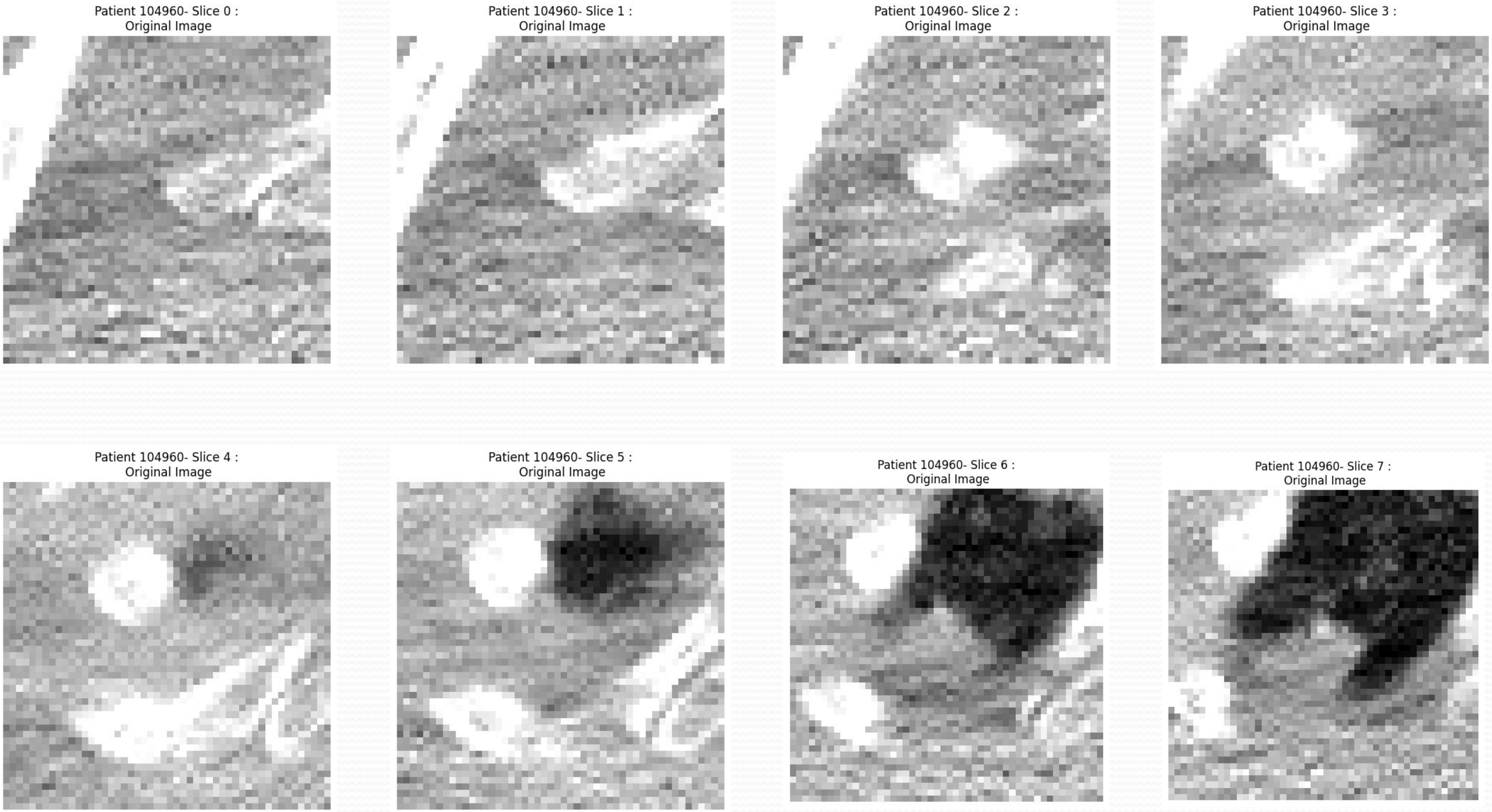






Y_true [True] & Y_pred = [False] -> Size= 32 Nodule

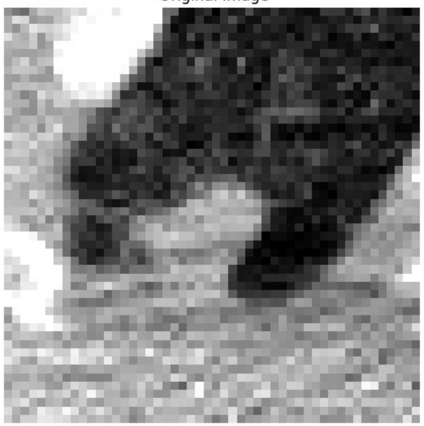
104960



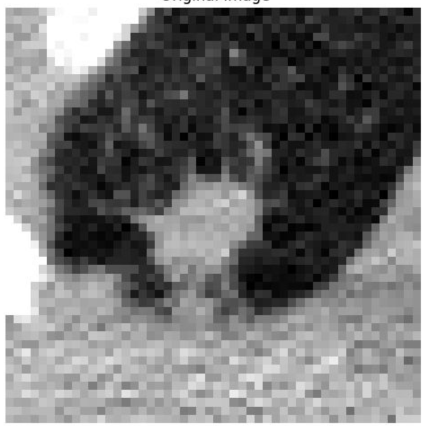
Y_true [True] & Y_pred = [False] -> Size= 32 Nodule

104960

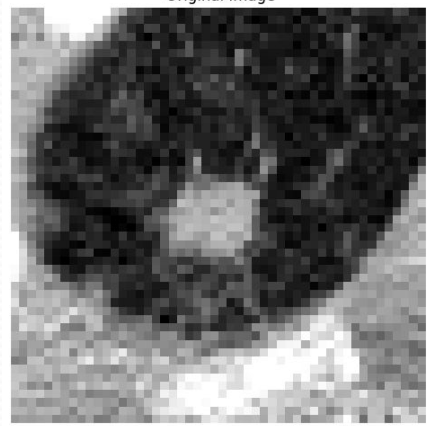
Patient 104960- Slice 8 :
Original Image



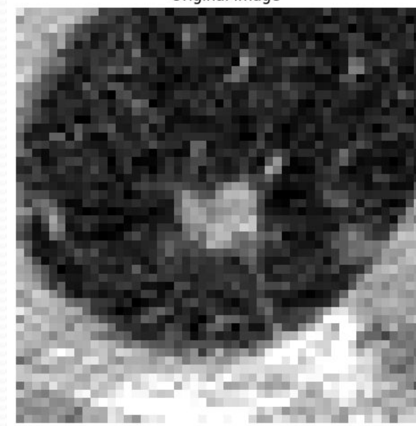
Patient 104960- Slice 9 :
Original Image



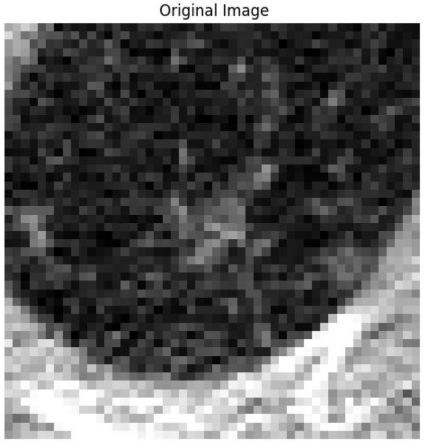
Patient 104960- Slice 10 :
Original Image



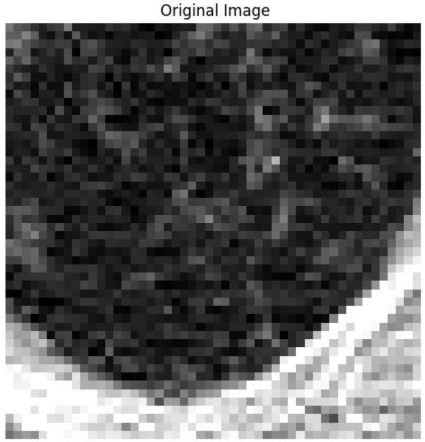
Patient 104960- Slice 11 :
Original Image



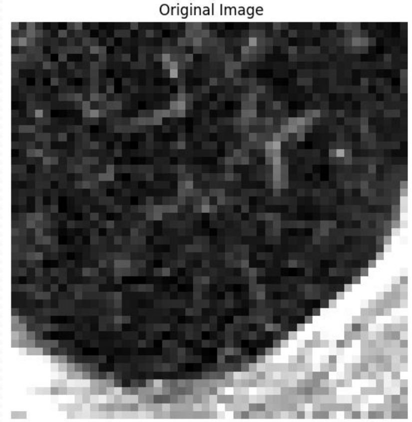
Patient 104960- Slice 12 :
Original Image



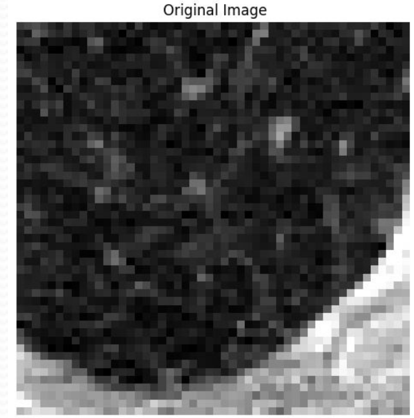
Patient 104960- Slice 13 :
Original Image



Patient 104960- Slice 14 :
Original Image

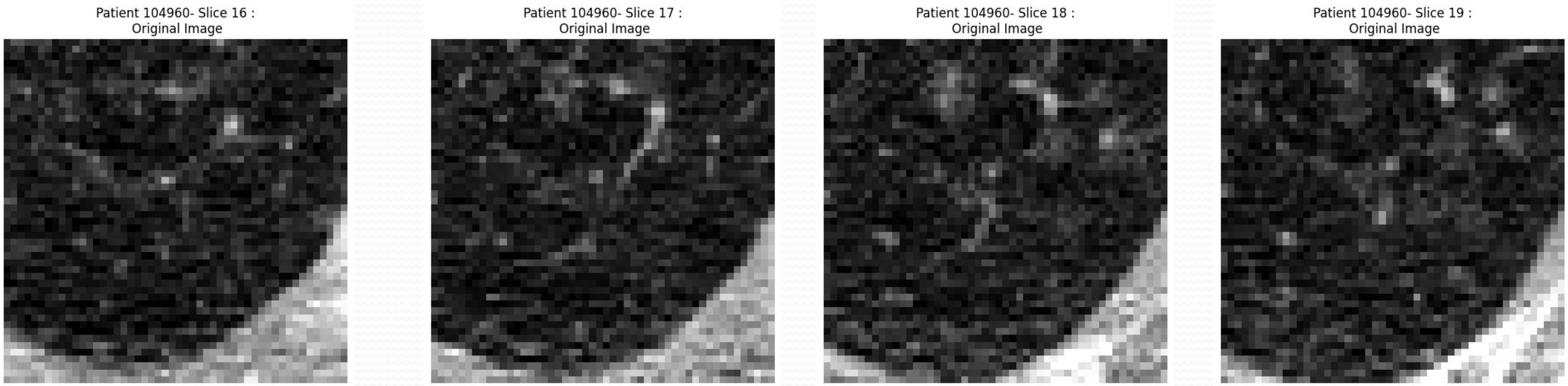


Patient 104960- Slice 15 :
Original Image



Y_true [True] & Y_pred = [False] -> Size= 32 Nodule

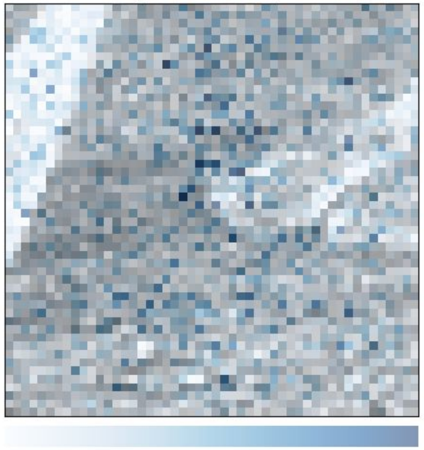
104960



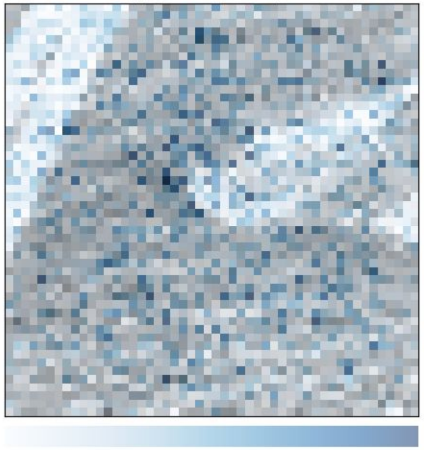
Y_true [True] & Y_pred = [False] -> Size= 32 SM

104960

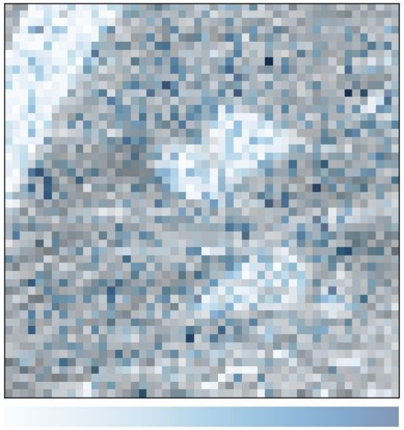
Patient 104960- Slice 0 : Overlayed Gradient Magnitudes



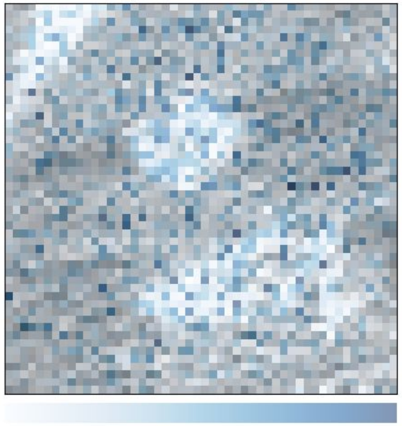
Patient 104960- Slice 1 : Overlayed Gradient Magnitudes



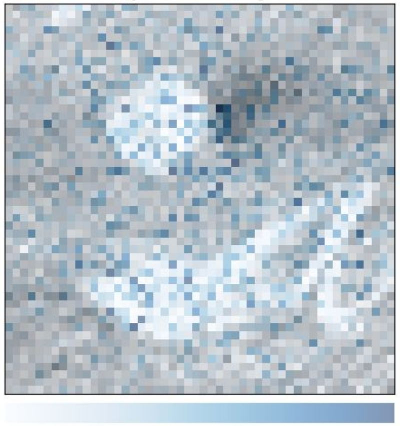
Patient 104960- Slice 2 : Overlayed Gradient Magnitudes



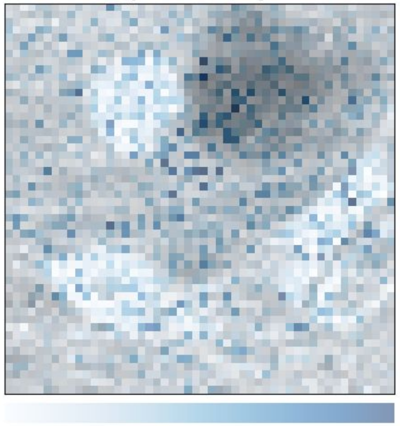
Patient 104960- Slice 3 : Overlayed Gradient Magnitudes



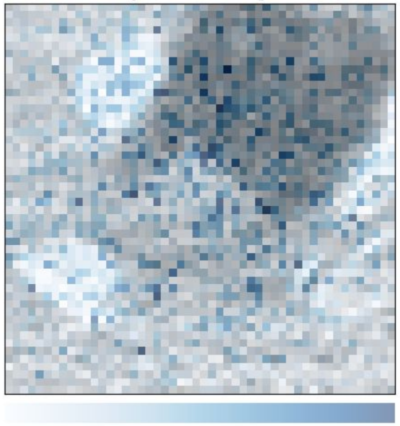
Patient 104960- Slice 4 : Overlayed Gradient Magnitudes



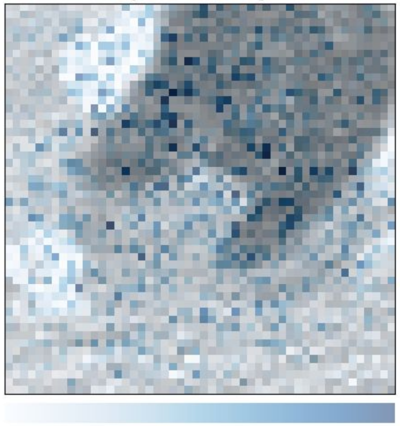
Patient 104960- Slice 5 : Overlayed Gradient Magnitudes



Patient 104960- Slice 6 : Overlayed Gradient Magnitudes

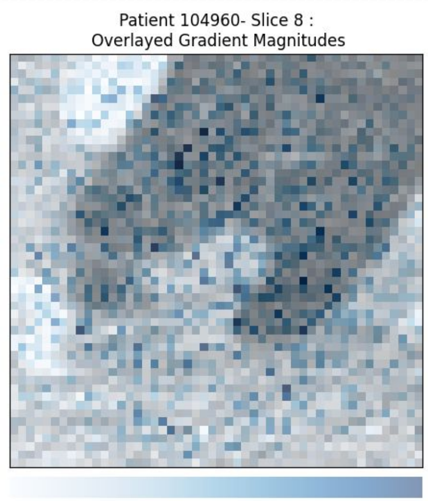


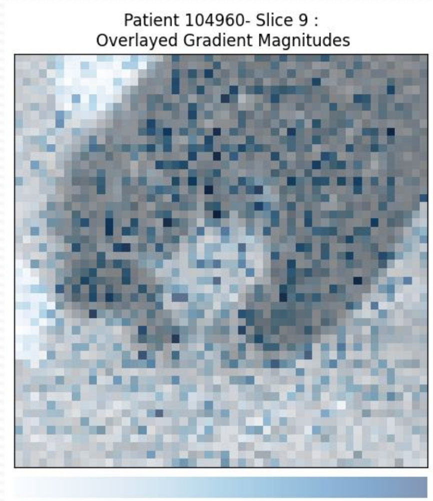
Patient 104960- Slice 7 : Overlayed Gradient Magnitudes

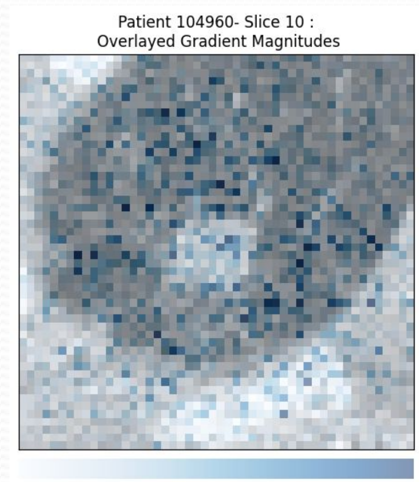


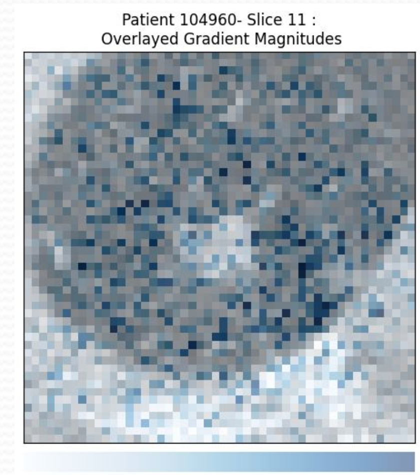
Y_true [True] & Y_pred = [False] -> Size= 32 SM

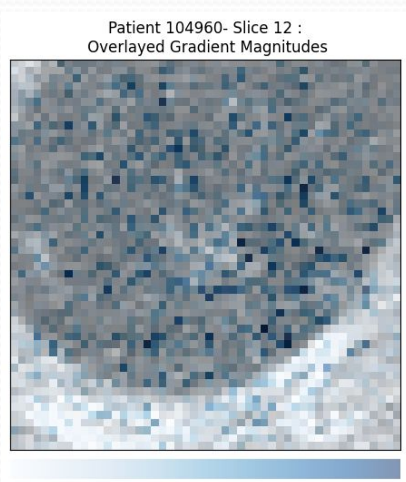
104960

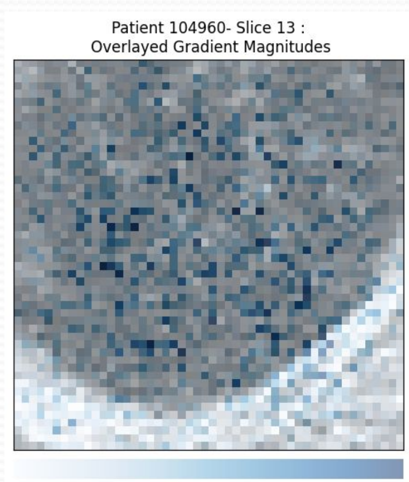


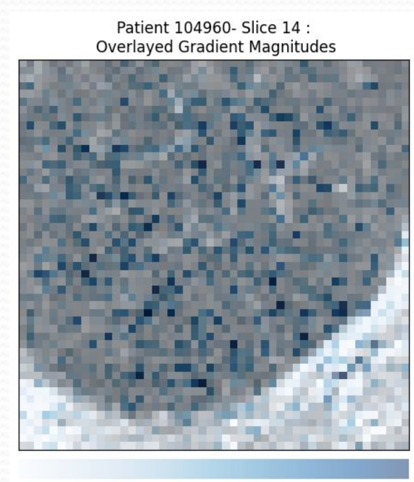


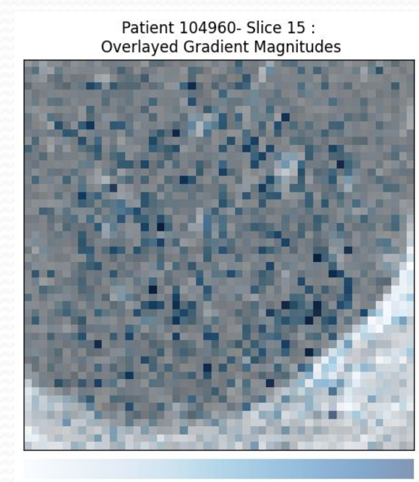






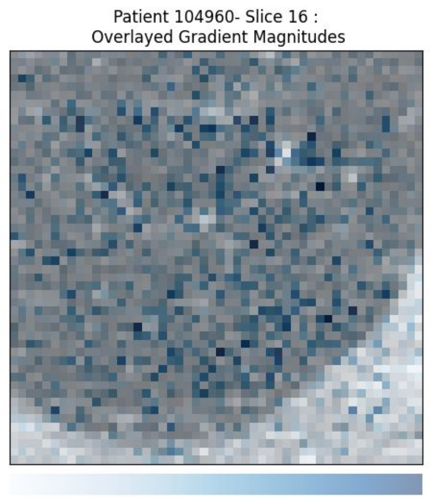


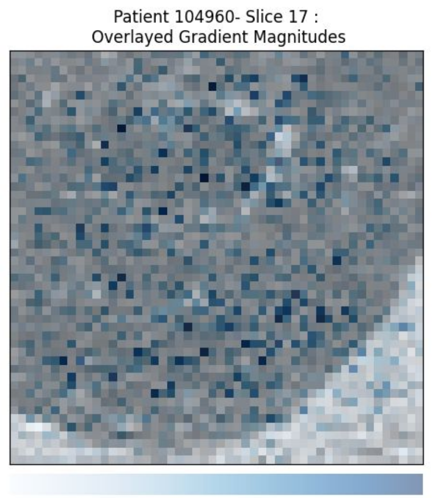


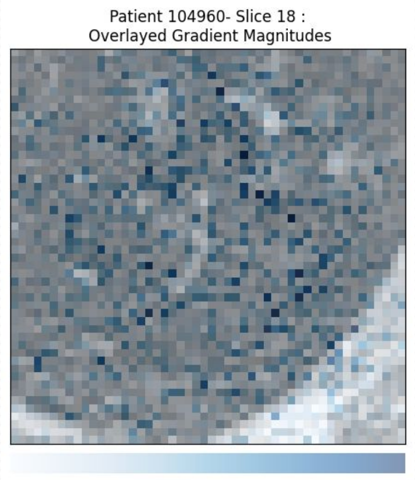


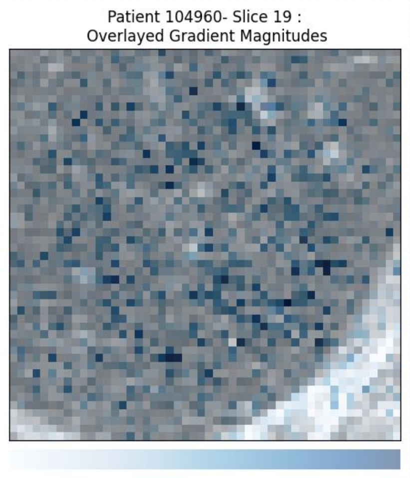
Y_true [True] & Y_pred = [False] -> Size= 32 SM

104960



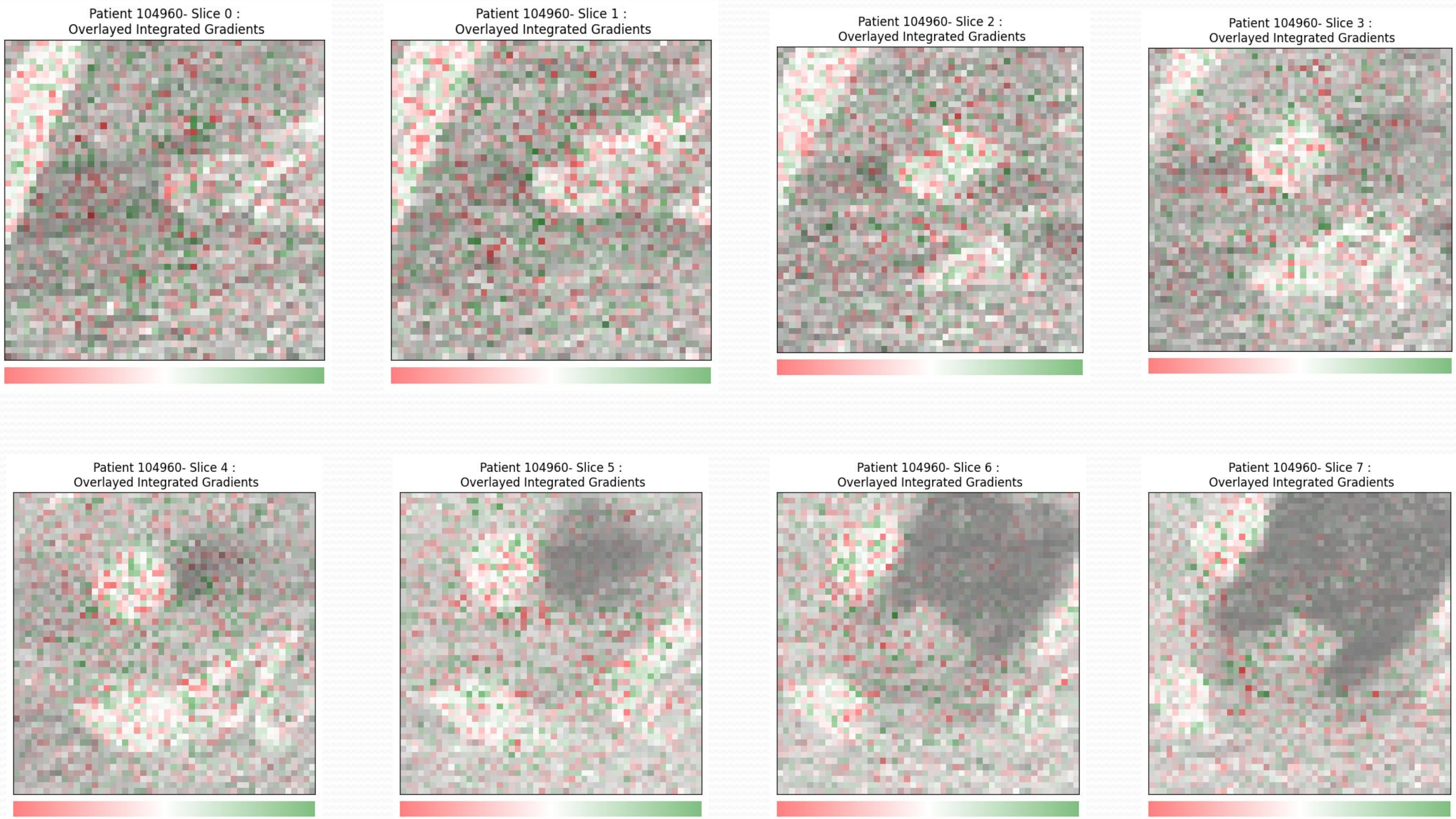






Y_true [True] & Y_pred = [False] -> Size= 32 IG_All

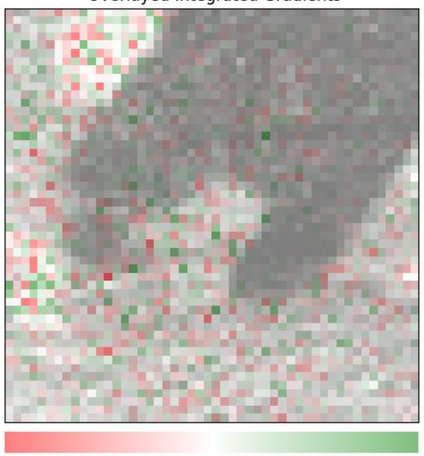
104960



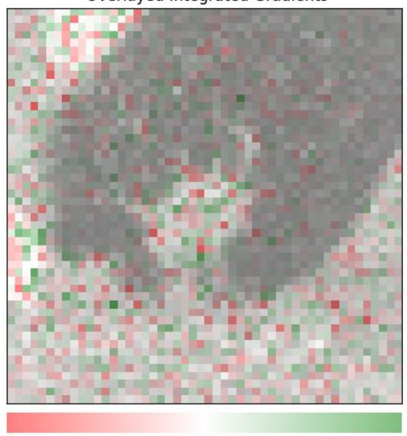
Y_true [True] & Y_pred = [False] -> Size= 32 IG_All

104960

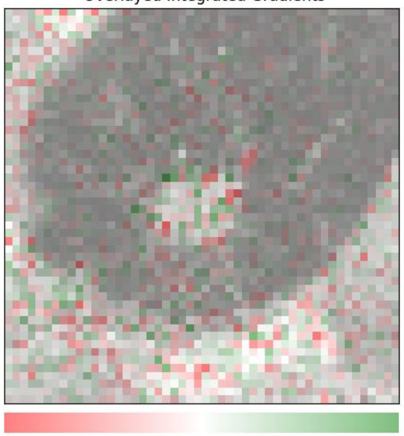
Patient 104960- Slice 8 :
Overlayed Integrated Gradients



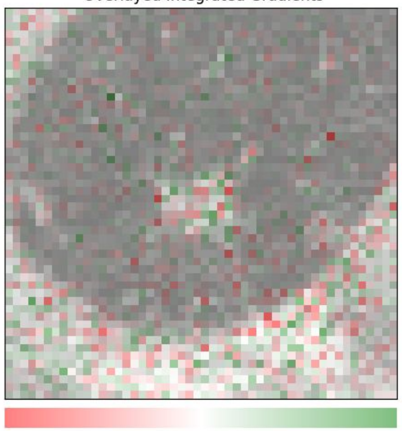
Patient 104960- Slice 9 :
Overlayed Integrated Gradients



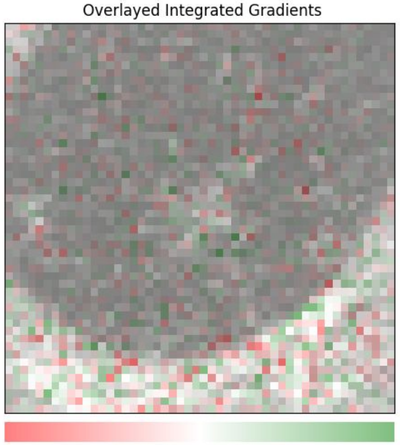
Patient 104960- Slice 10 :
Overlayed Integrated Gradients



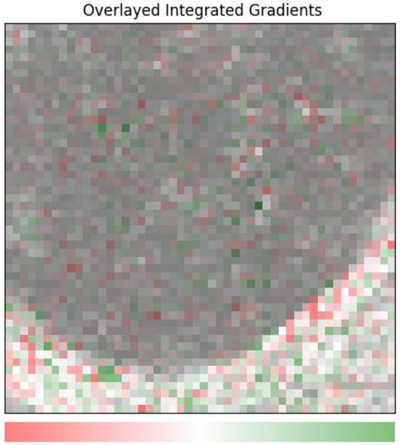
Patient 104960- Slice 11 :
Overlayed Integrated Gradients



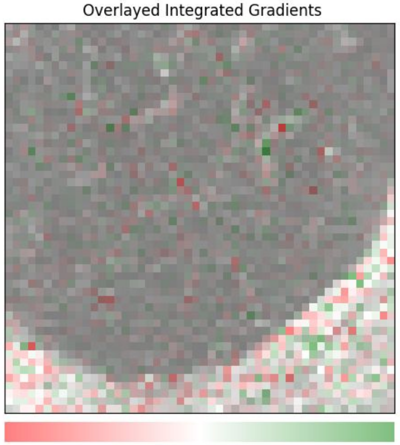
Patient 104960- Slice 12 :
Overlayed Integrated Gradients



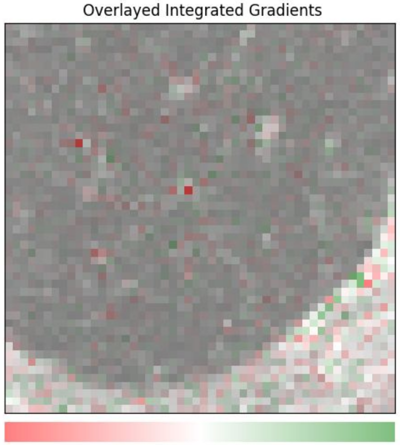
Patient 104960- Slice 13 :
Overlayed Integrated Gradients



Patient 104960- Slice 14 :
Overlayed Integrated Gradients



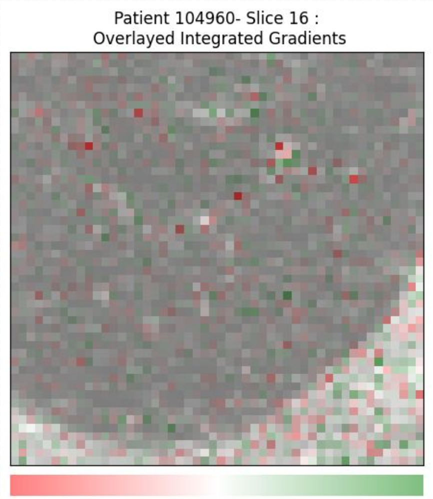
Patient 104960- Slice 15 :
Overlayed Integrated Gradients

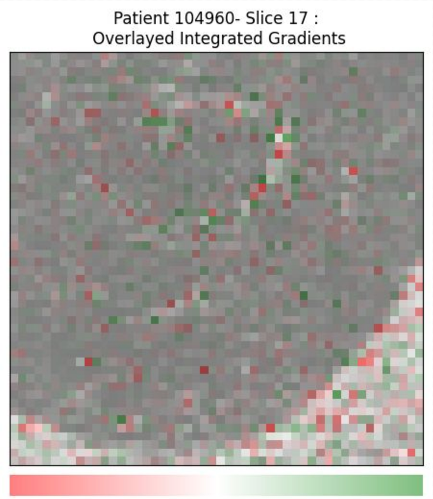


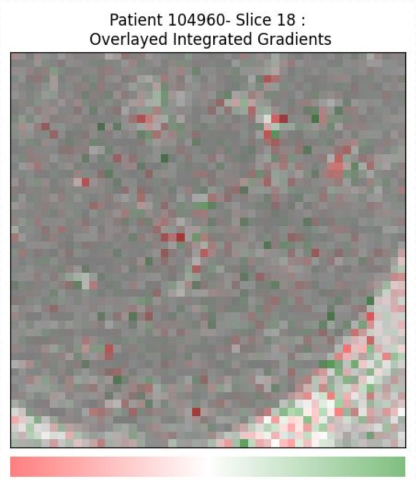
Y_true [True] & Y_pred = [False] -> Size= 32

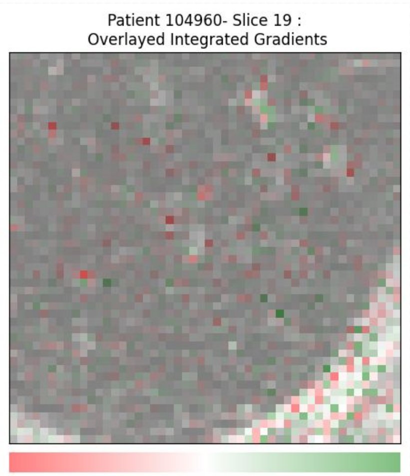
104960

IG_All



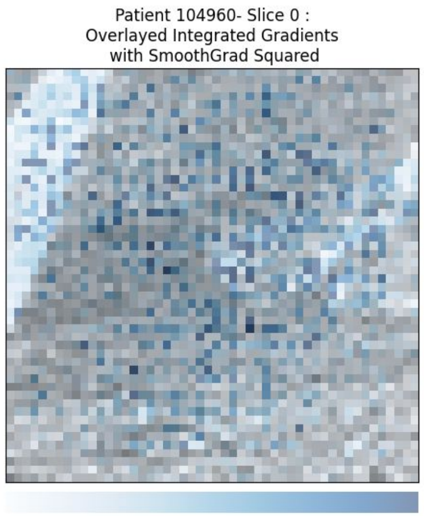


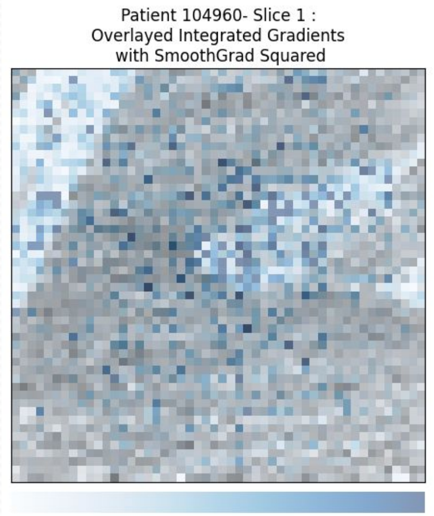


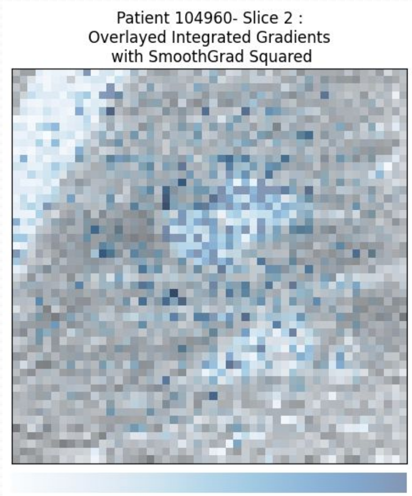


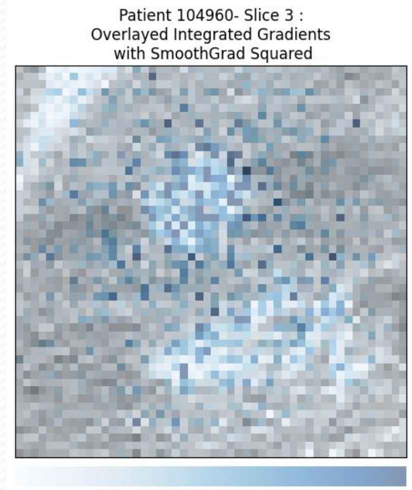
Y_true [True] & Y_pred = [False] -> Size= 32 IG_Abs

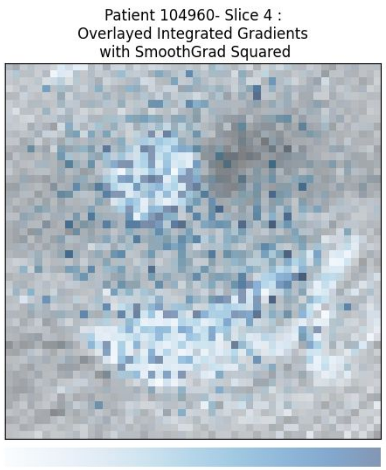
104960

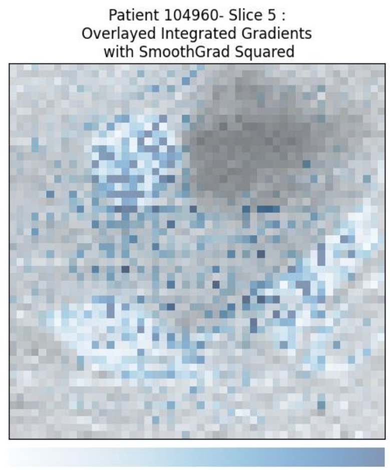


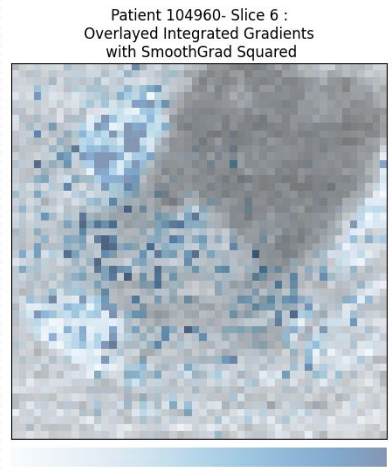


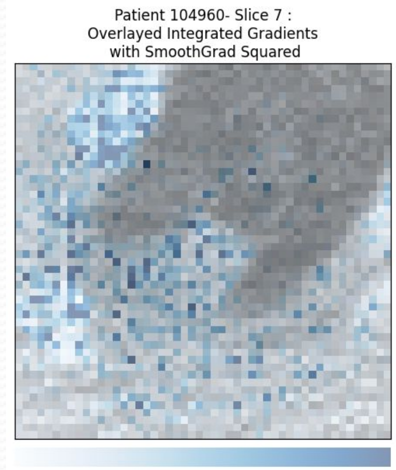








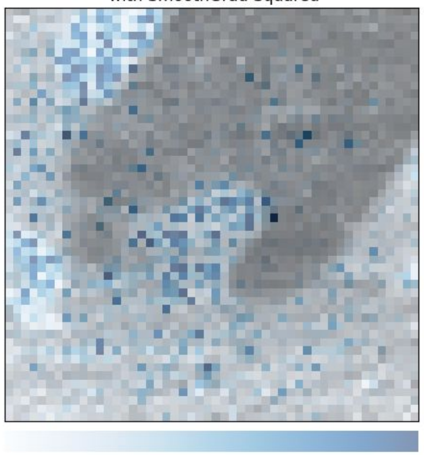




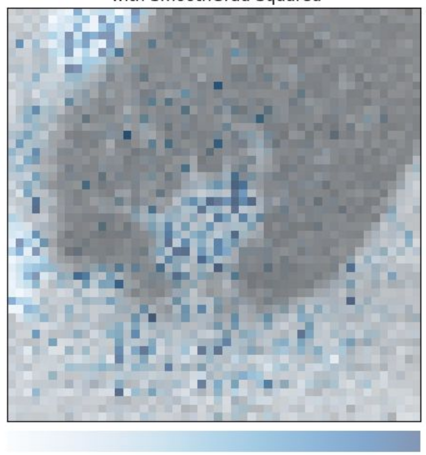
Y_true [True] & Y_pred = [False] -> Size= 32 104960

IG_Abs

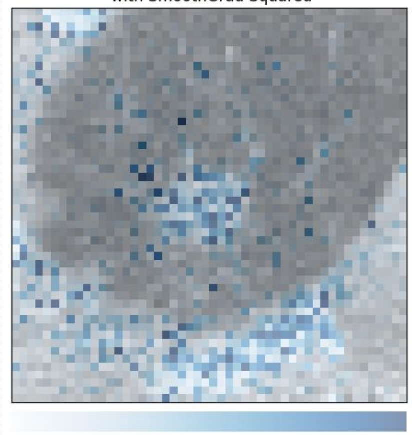
Patient 104960- Slice 8 :
Overlayed Integrated Gradients
with SmoothGrad Squared



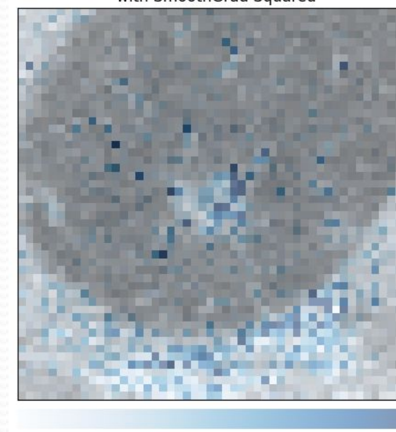
Patient 104960- Slice 9 :
Overlayed Integrated Gradients
with SmoothGrad Squared



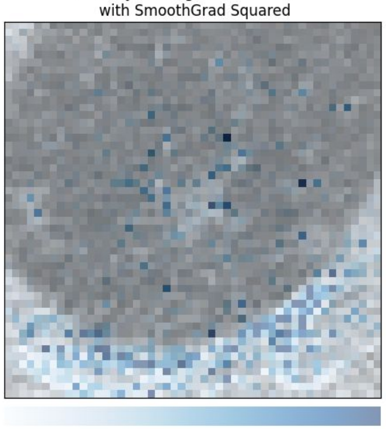
Patient 104960- Slice 10 :
Overlayed Integrated Gradients
with SmoothGrad Squared



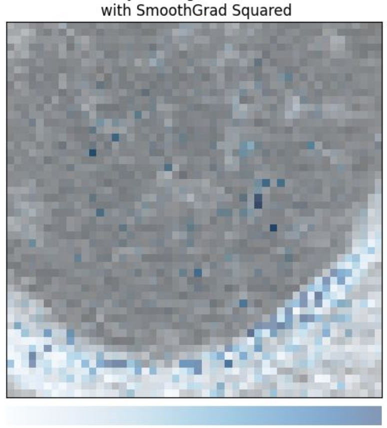
Patient 104960- Slice 11 :
Overlayed Integrated Gradients
with SmoothGrad Squared



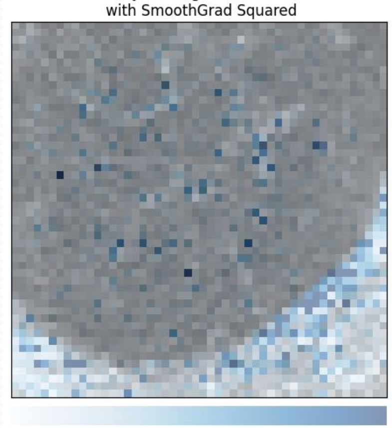
Patient 104960- Slice 12 :
Overlayed Integrated Gradients
with SmoothGrad Squared



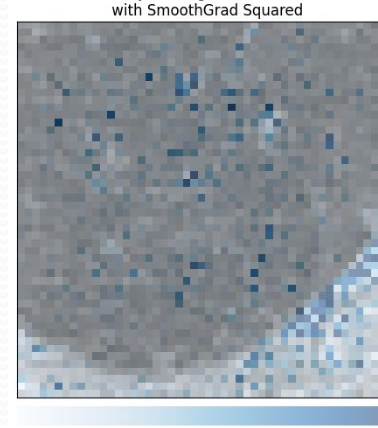
Patient 104960- Slice 13 :
Overlayed Integrated Gradients
with SmoothGrad Squared



Patient 104960- Slice 14 :
Overlayed Integrated Gradients
with SmoothGrad Squared

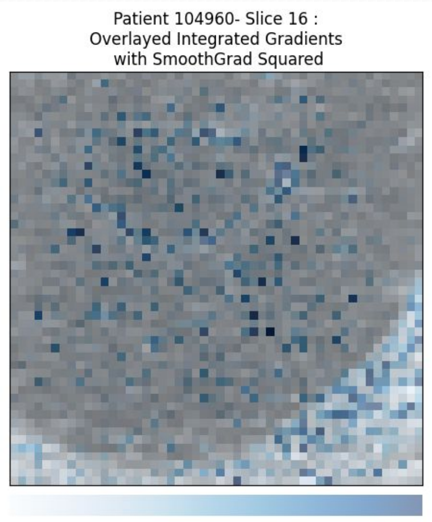


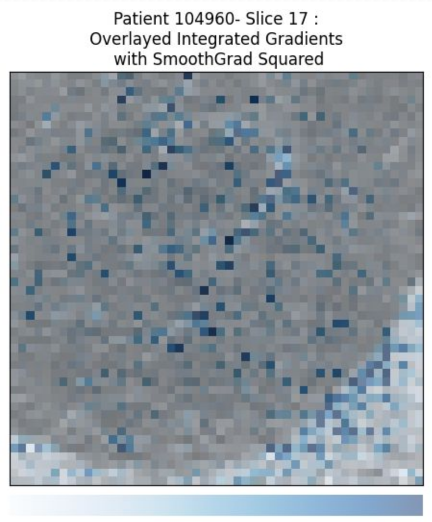
Patient 104960- Slice 15 :
Overlayed Integrated Gradients
with SmoothGrad Squared

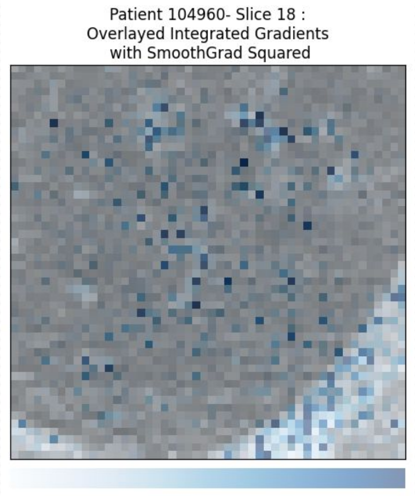


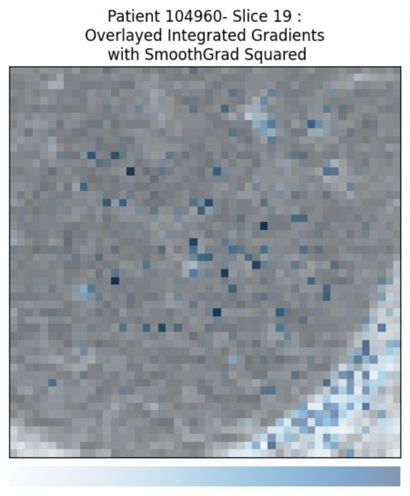
Y_true [True] & Y_pred = [False] -> Size= 32 IG_Abs

104960



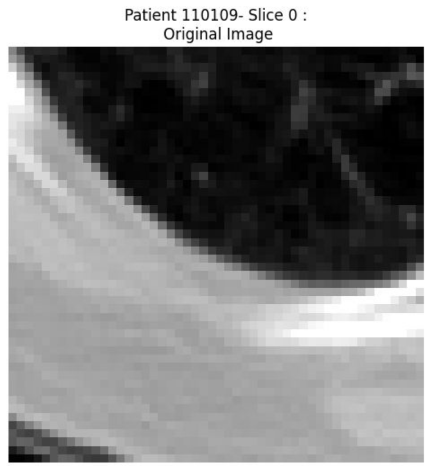


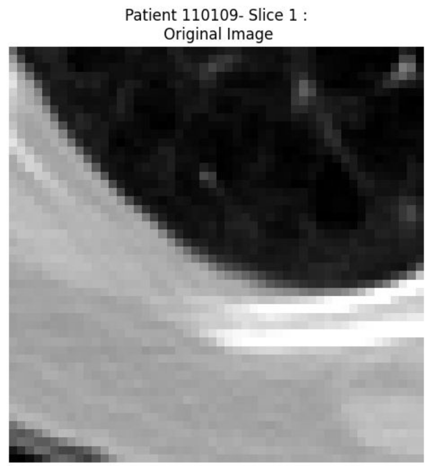


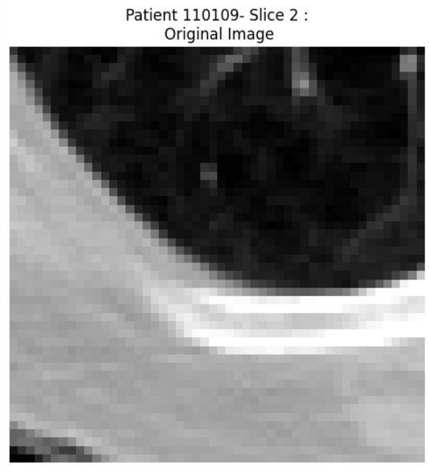


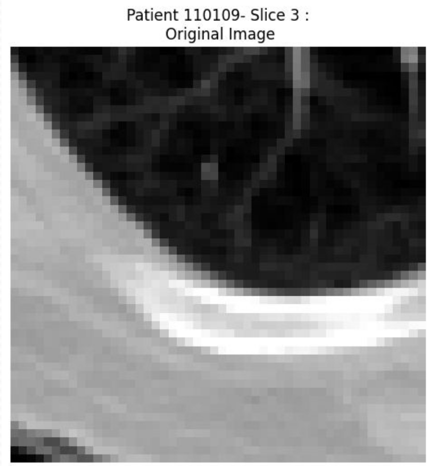
Y_true [False] & Y_pred = [True] -> Size= 19
Nodule

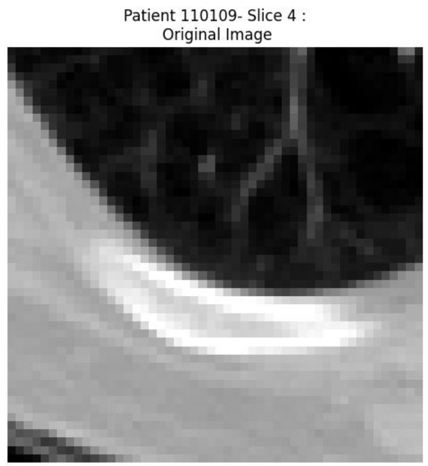
110109

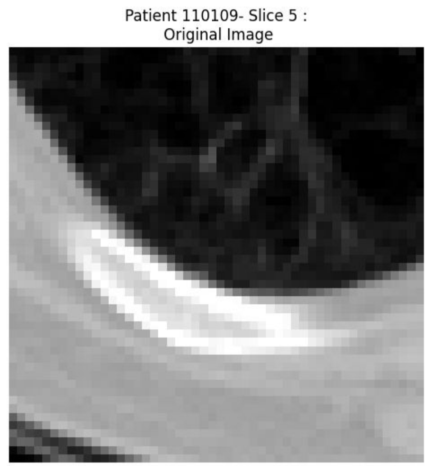


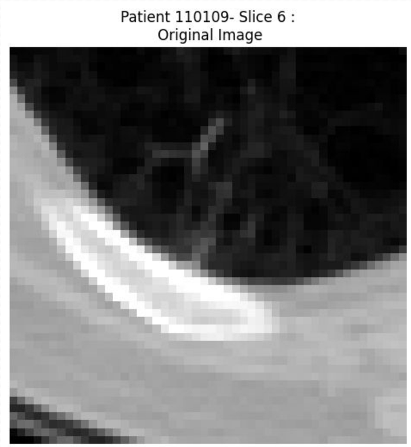


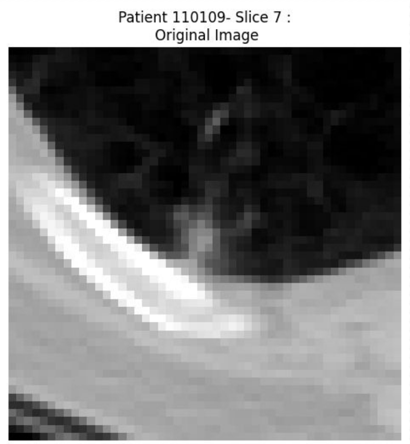






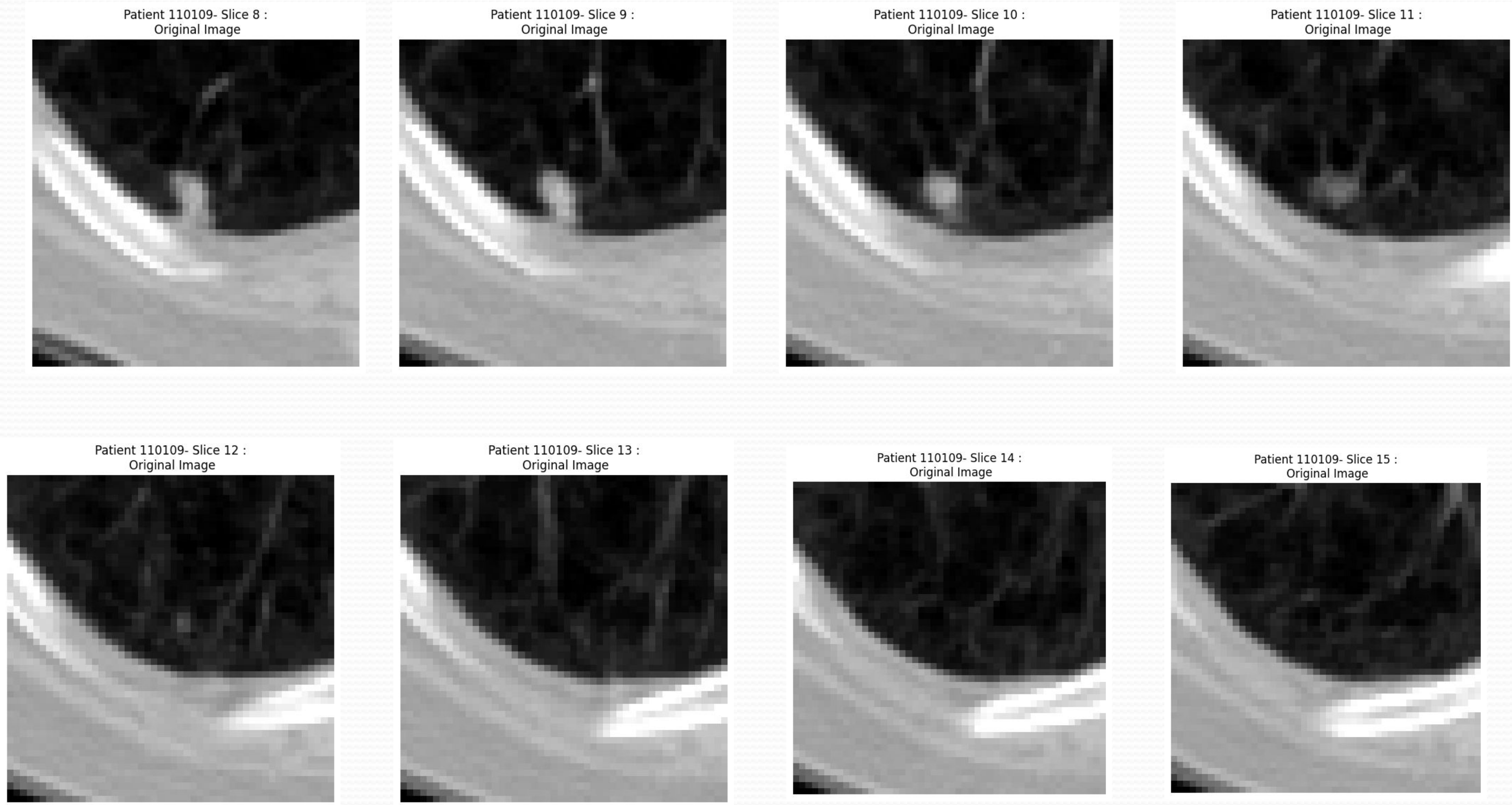






Y_true [False] & Y_pred = [True] -> Size= 19 Nodule

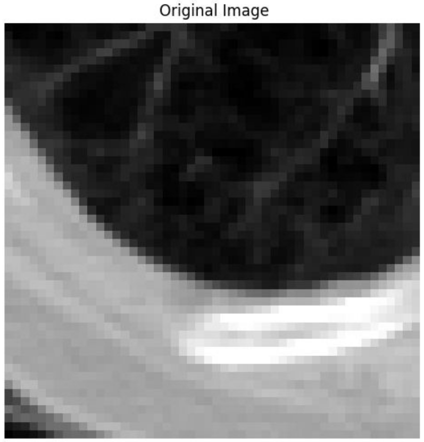
110109



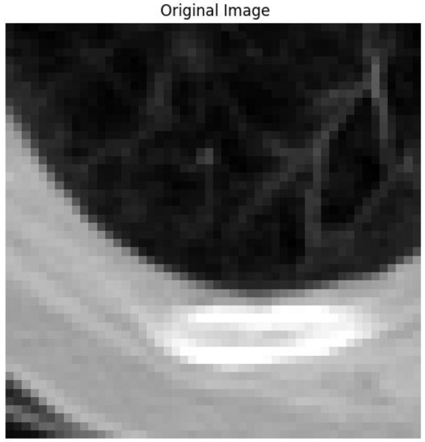
Y_true [False] & Y_pred = [True] -> Size= 19 Nodule

110109

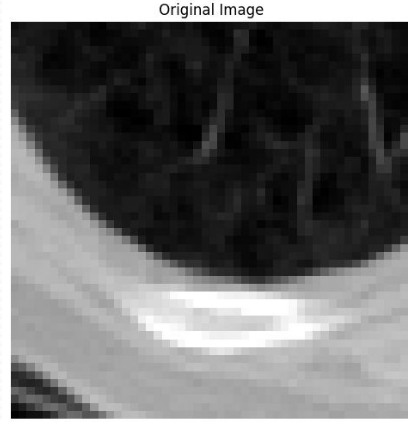
Patient 110109- Slice 16 :
Original Image



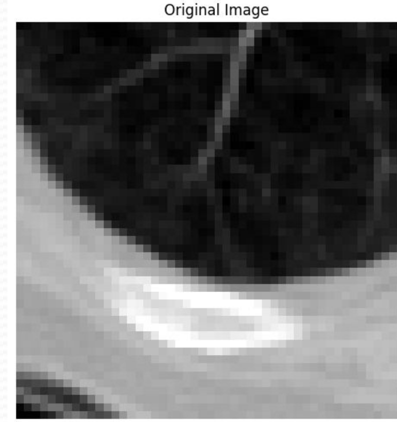
Patient 110109- Slice 17 :
Original Image



Patient 110109- Slice 18 :
Original Image

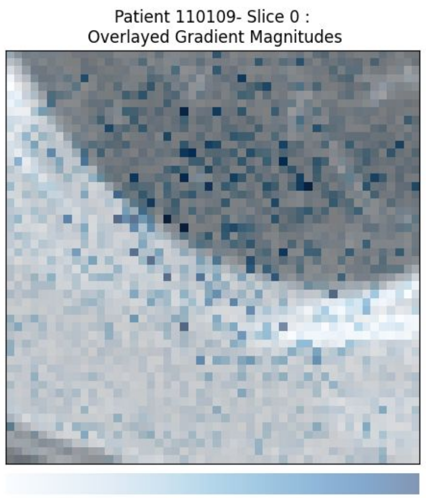


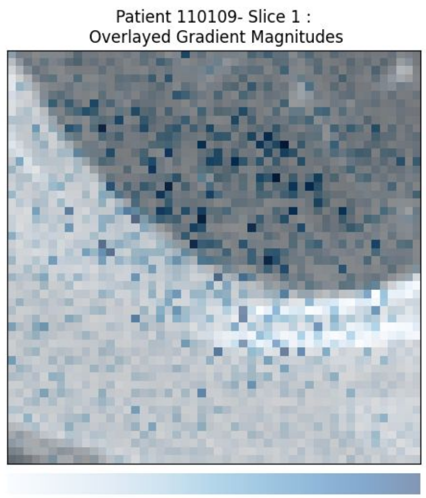
Patient 110109- Slice 19 :
Original Image

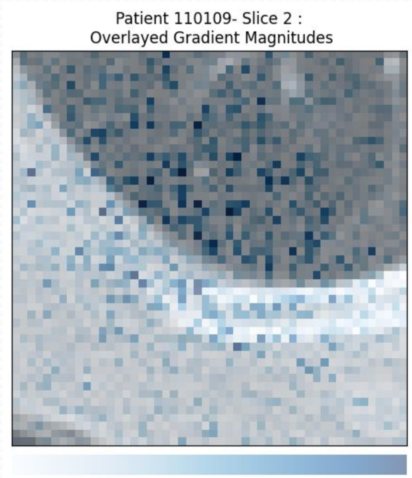


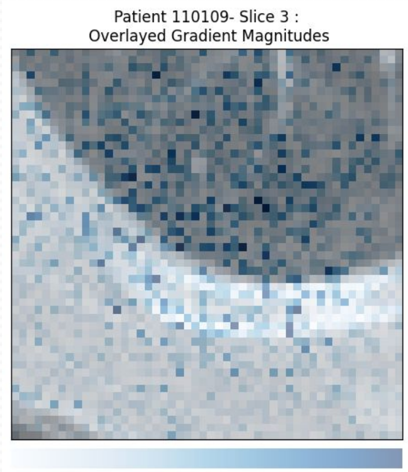
Y_true [False] & Y_pred = [True] -> Size= 19 SM

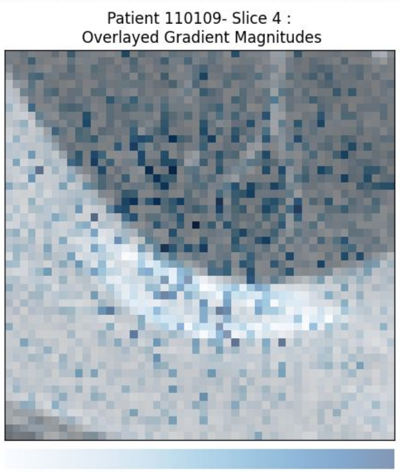
110109

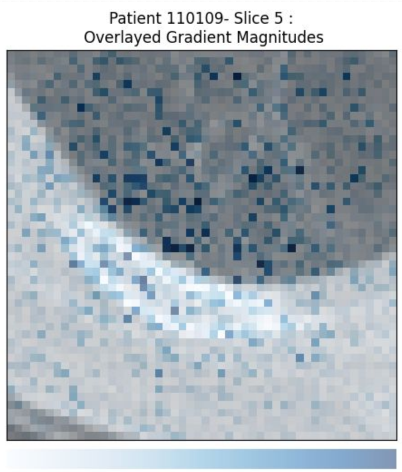


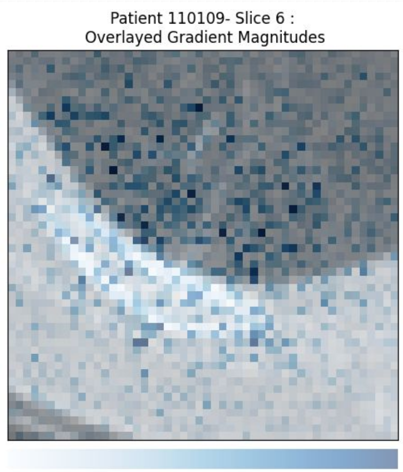


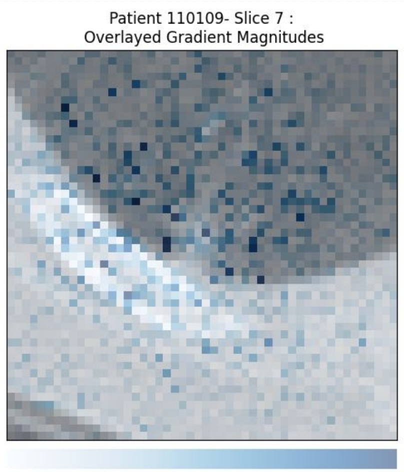






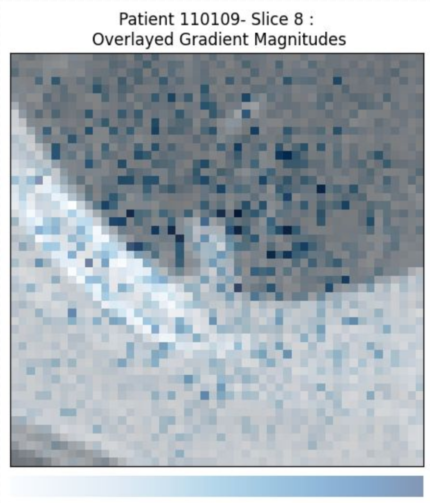


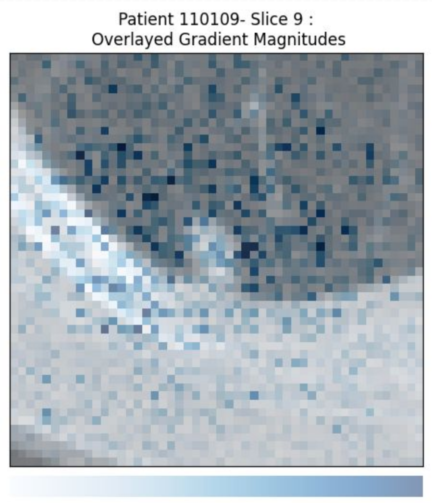


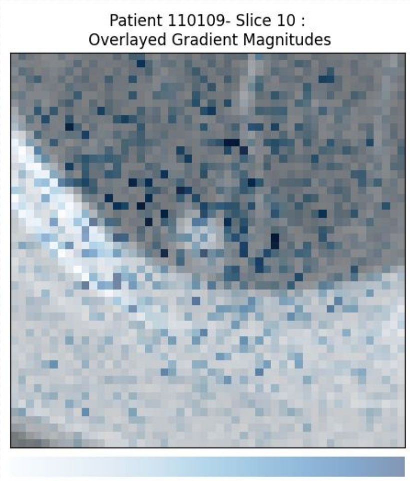


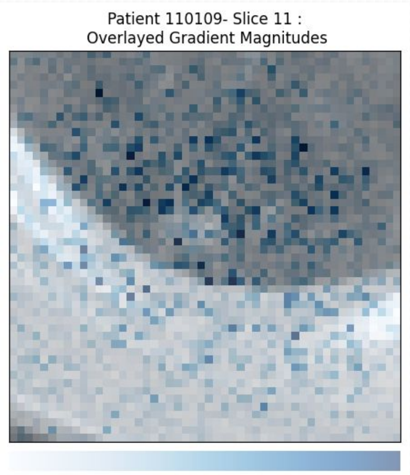
Y_true [False] & Y_pred = [True] -> Size= 19 SM

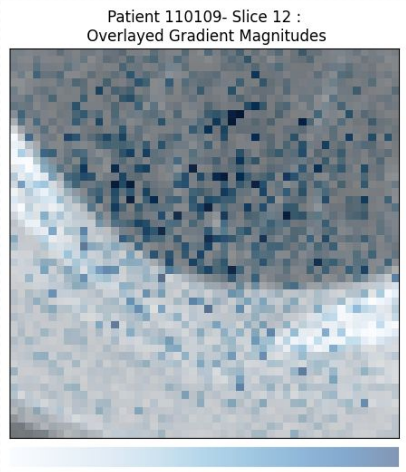
110109

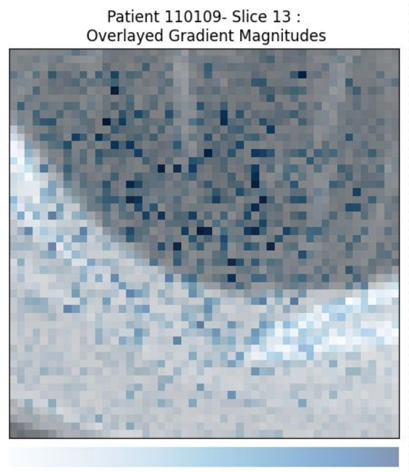


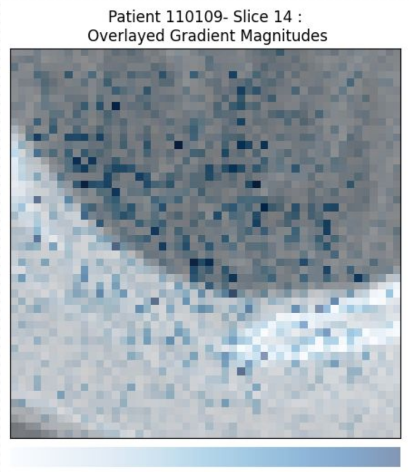


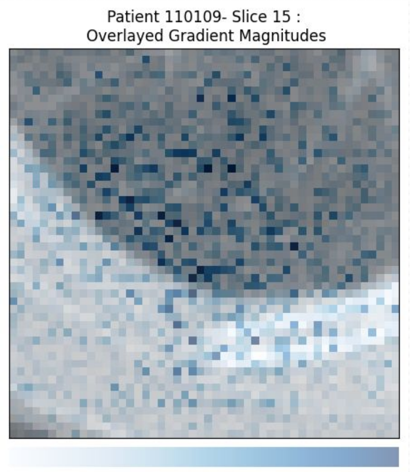






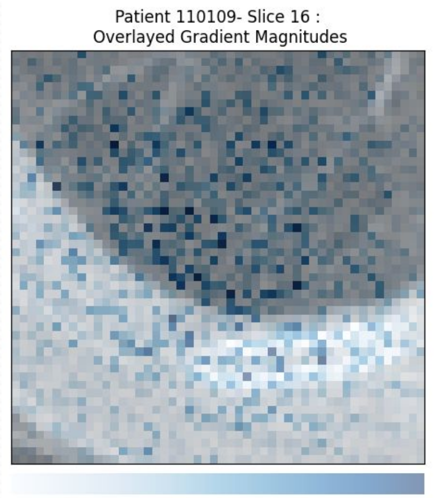


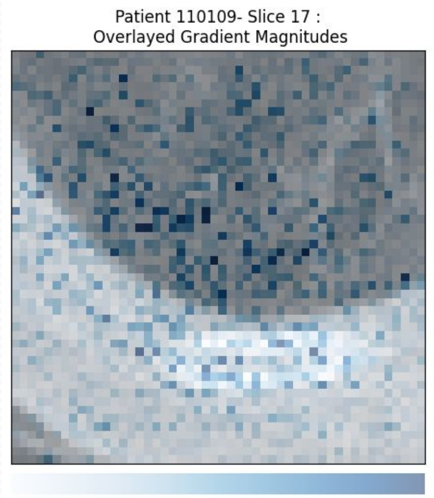


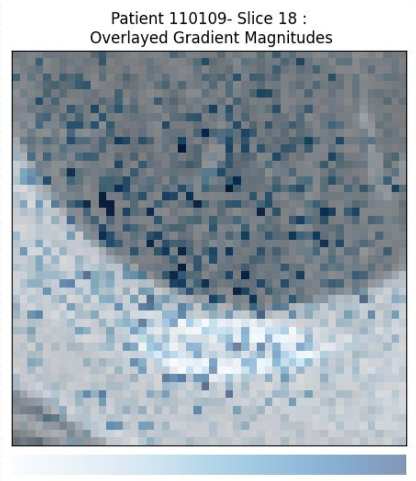


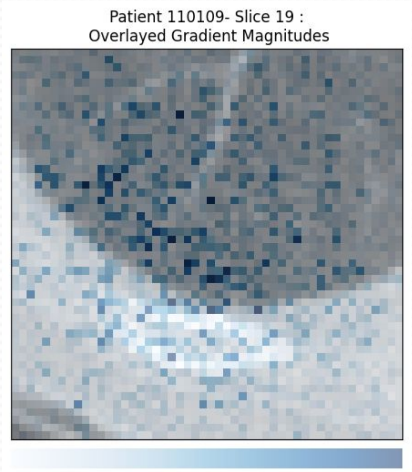
Y_true [False] & Y_pred = [True] -> Size= 19 SM

110109





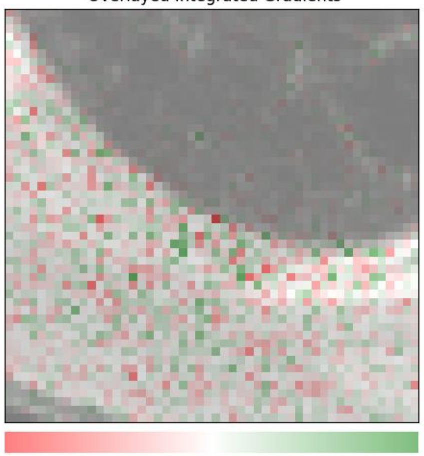




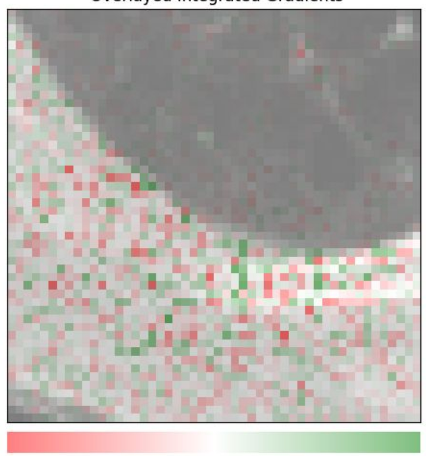
Y_true [False] & Y_pred = [True] -> Size= 19
IG_ALL

110109

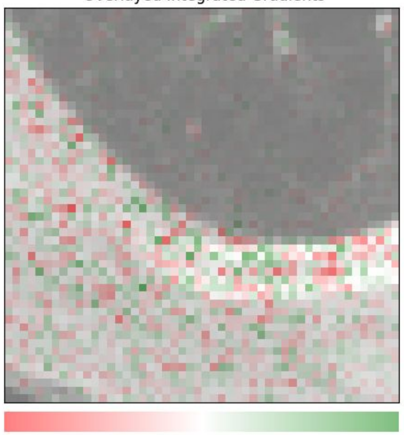
Patient 110109- Slice 0 :
Overlayed Integrated Gradients



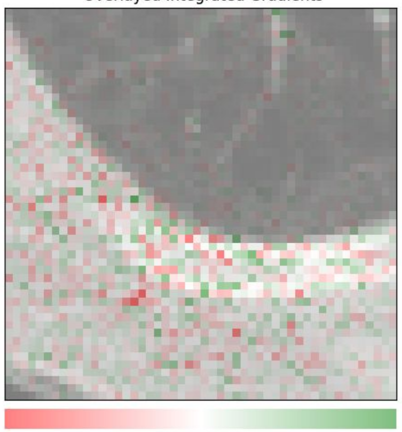
Patient 110109- Slice 1 :
Overlayed Integrated Gradients



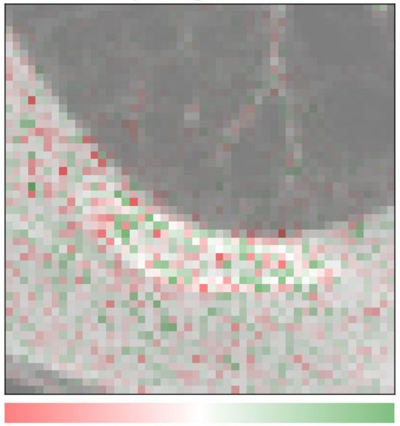
Patient 110109- Slice 2 :
Overlayed Integrated Gradients



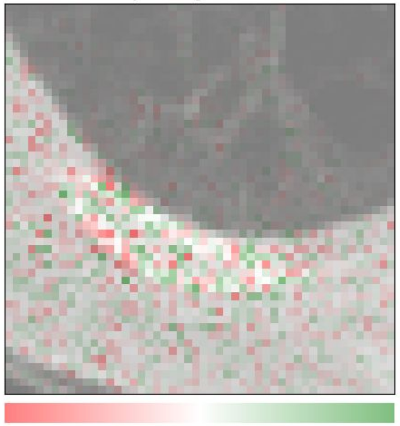
Patient 110109- Slice 3 :
Overlayed Integrated Gradients



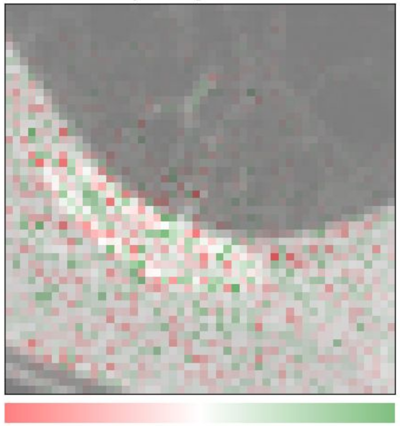
Patient 110109- Slice 4 :
Overlayed Integrated Gradients



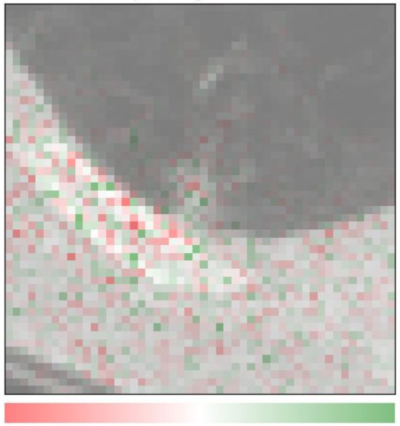
Patient 110109- Slice 5 :
Overlayed Integrated Gradients



Patient 110109- Slice 6 :
Overlayed Integrated Gradients



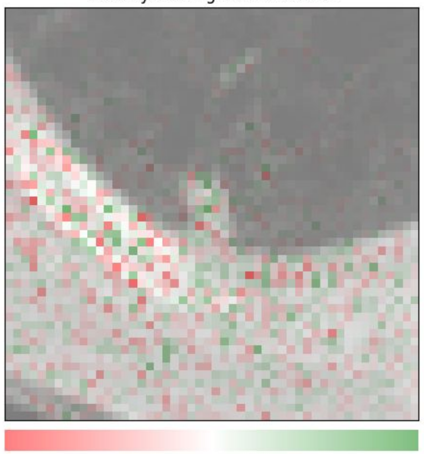
Patient 110109- Slice 7 :
Overlayed Integrated Gradients



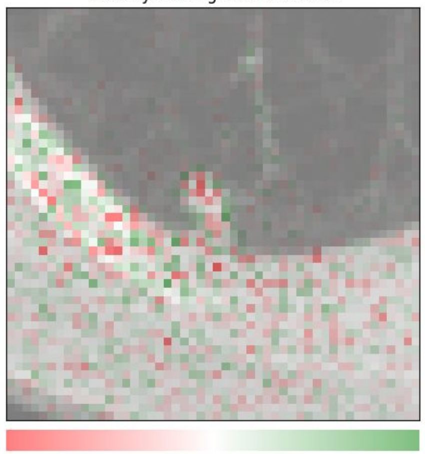
Y_true [False] & Y_pred = [True] -> Size= 19
IG_ALL

110109

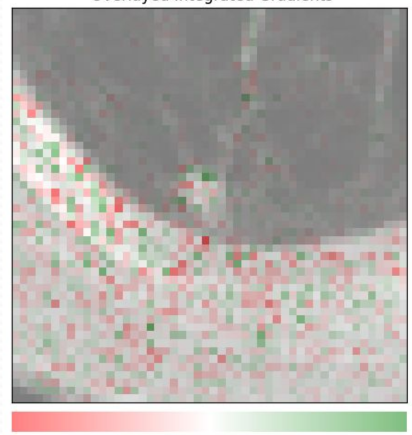
Patient 110109- Slice 8 :
Overlayed Integrated Gradients



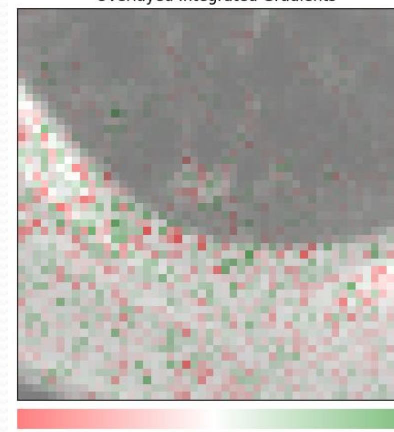
Patient 110109- Slice 9 :
Overlayed Integrated Gradients



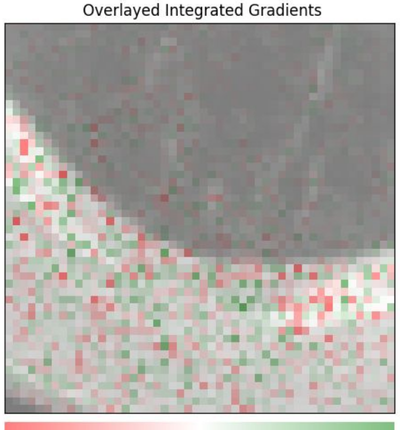
Patient 110109- Slice 10 :
Overlayed Integrated Gradients



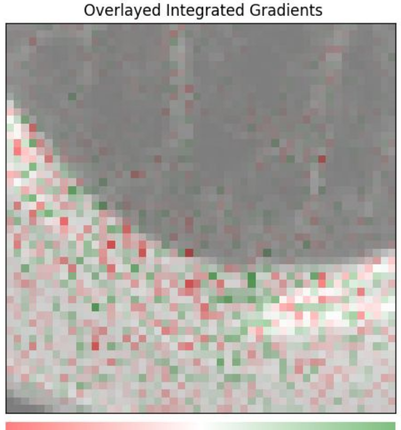
Patient 110109- Slice 11 :
Overlayed Integrated Gradients



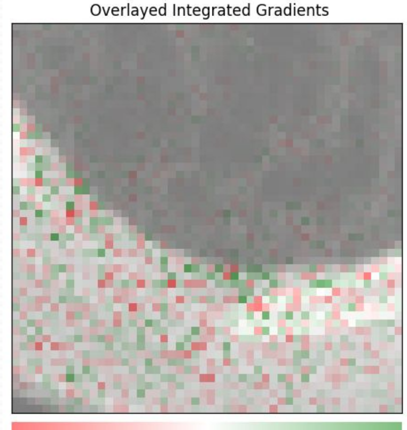
Patient 110109- Slice 12 :
Overlayed Integrated Gradients



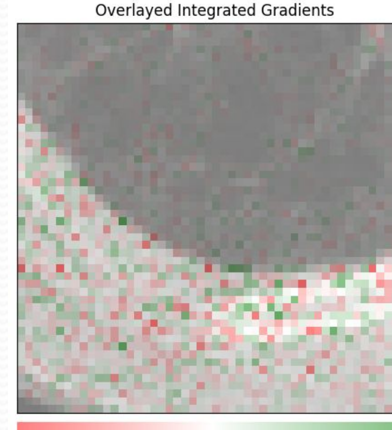
Patient 110109- Slice 13 :
Overlayed Integrated Gradients



Patient 110109- Slice 14 :
Overlayed Integrated Gradients

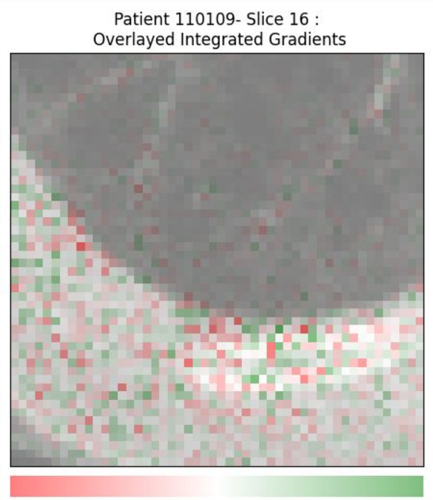


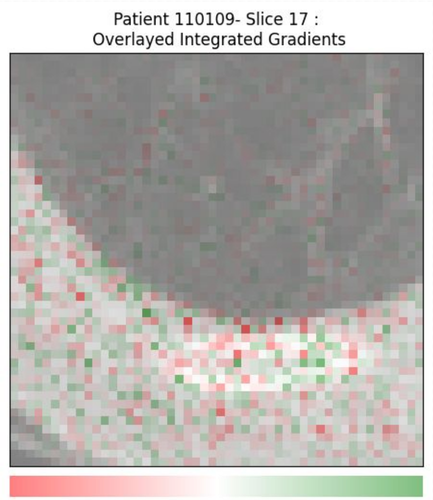
Patient 110109- Slice 15 :
Overlayed Integrated Gradients

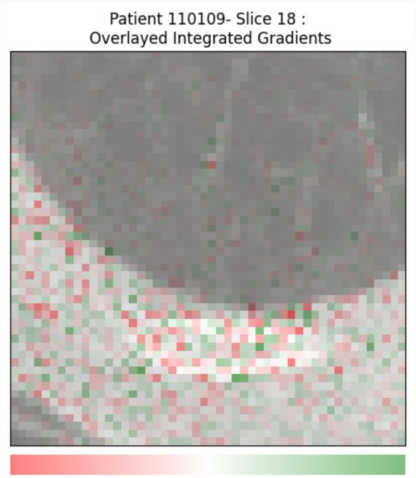


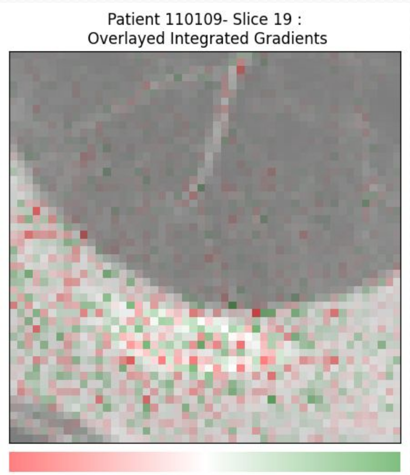
Y_true [False] & Y_pred = [True] -> Size= 19
IG_ALL

110109



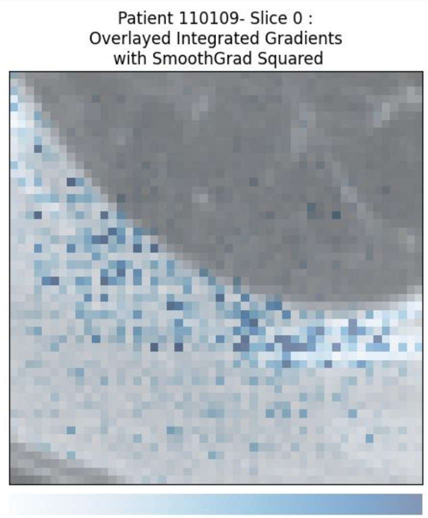


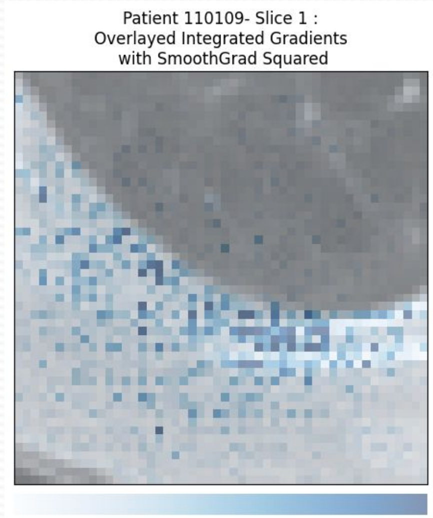


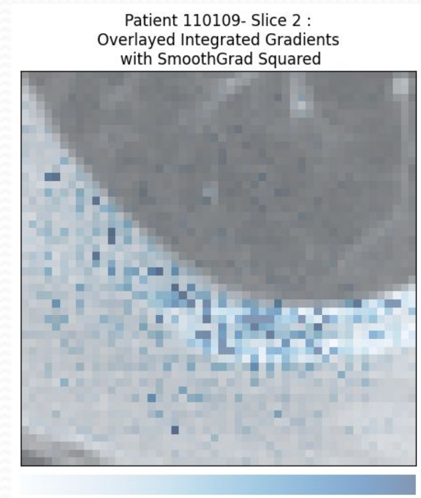


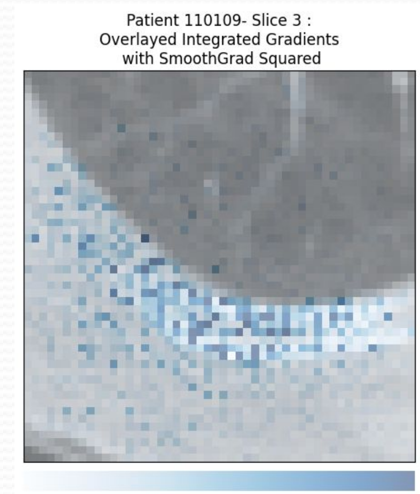
Y_true [False] & Y_pred = [True] -> Size= 19
IG_Abs

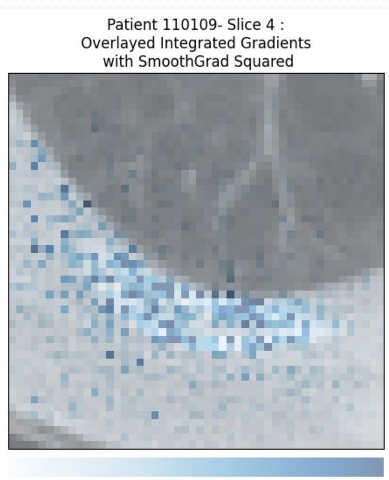
110109

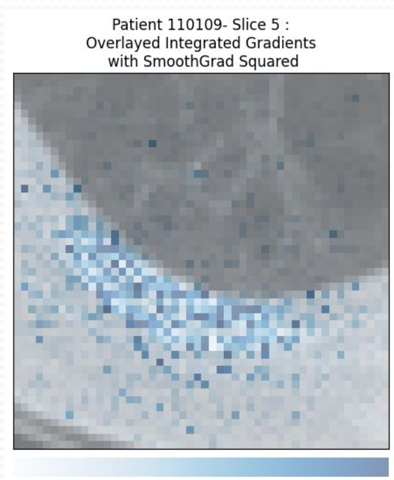


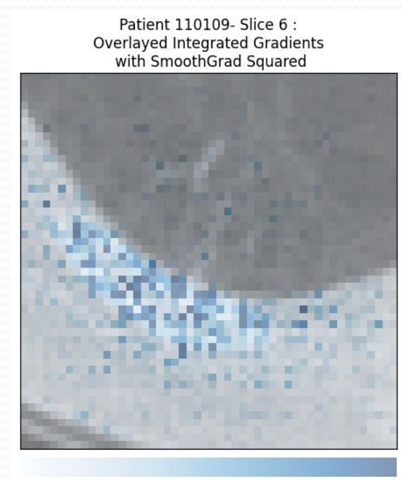


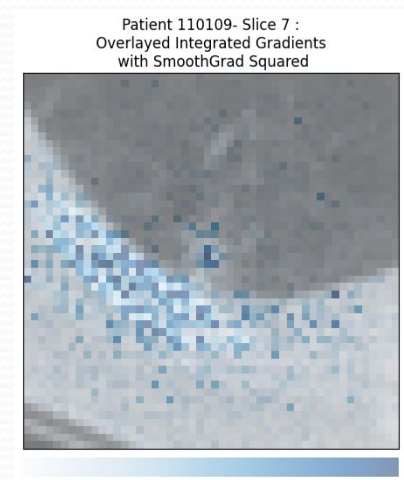






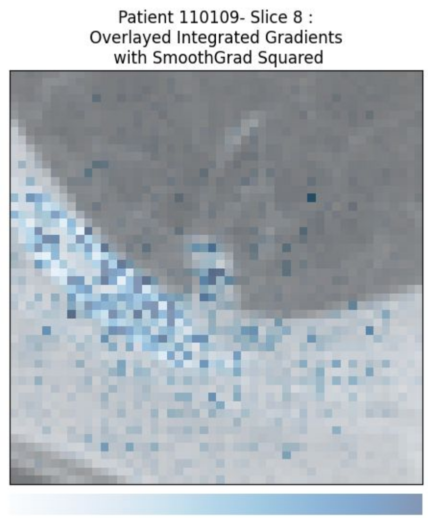


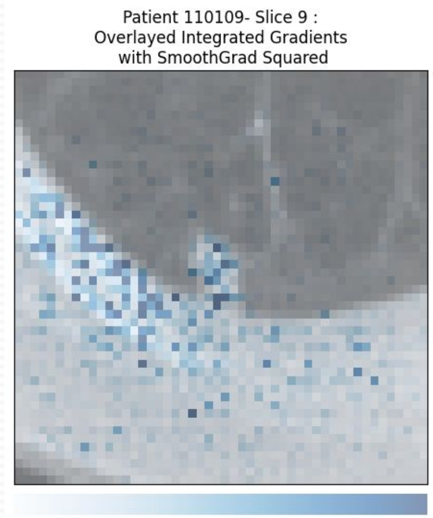


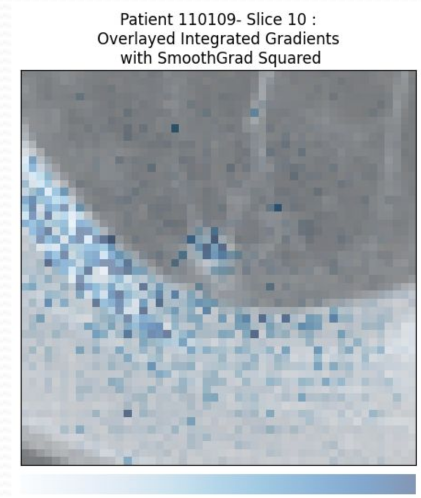


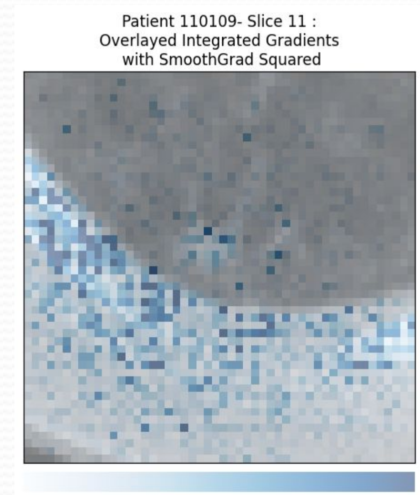
Y_true [False] & Y_pred = [True] -> Size= 19
IG_Abs

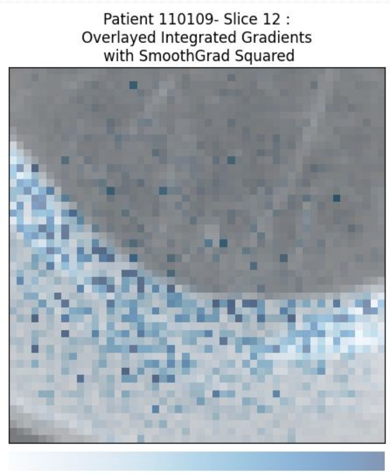
110109

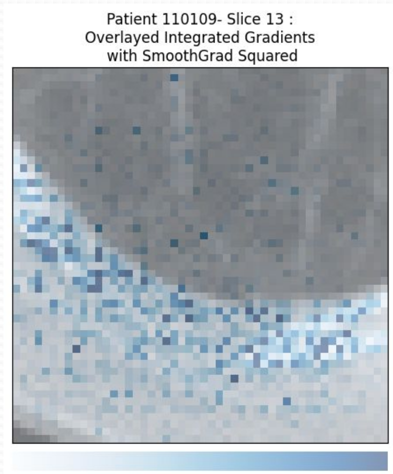


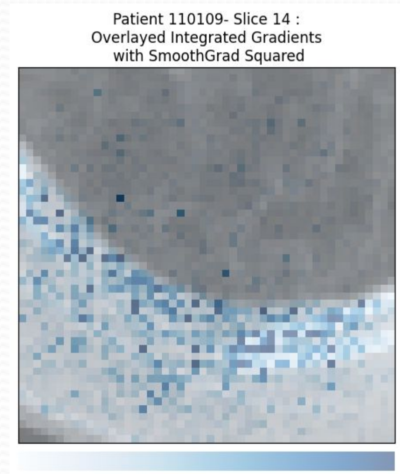


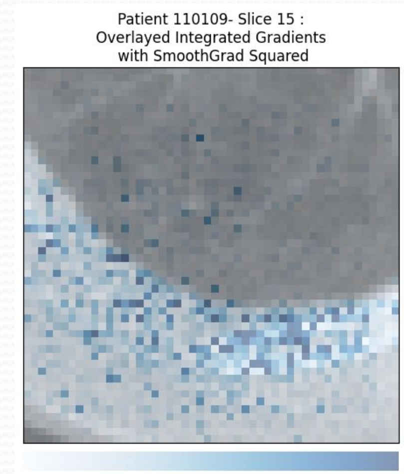






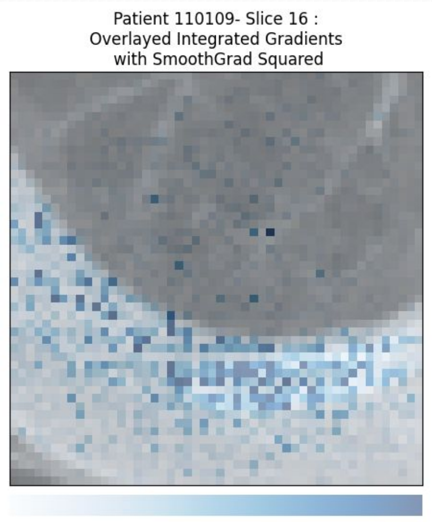


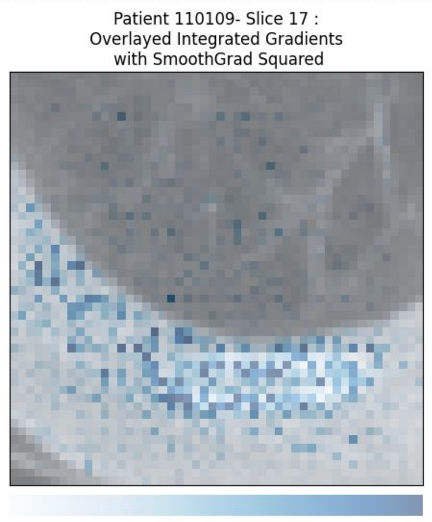


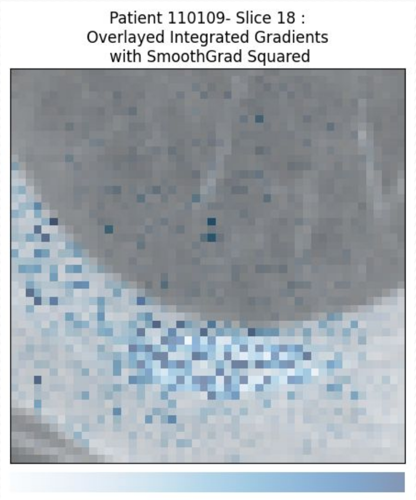


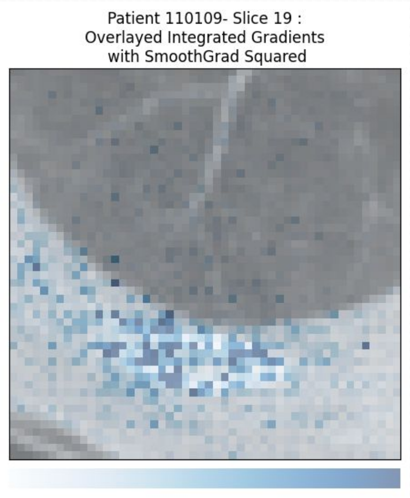
Y_true [False] & Y_pred = [True] -> Size= 19
IG_Abs

110109



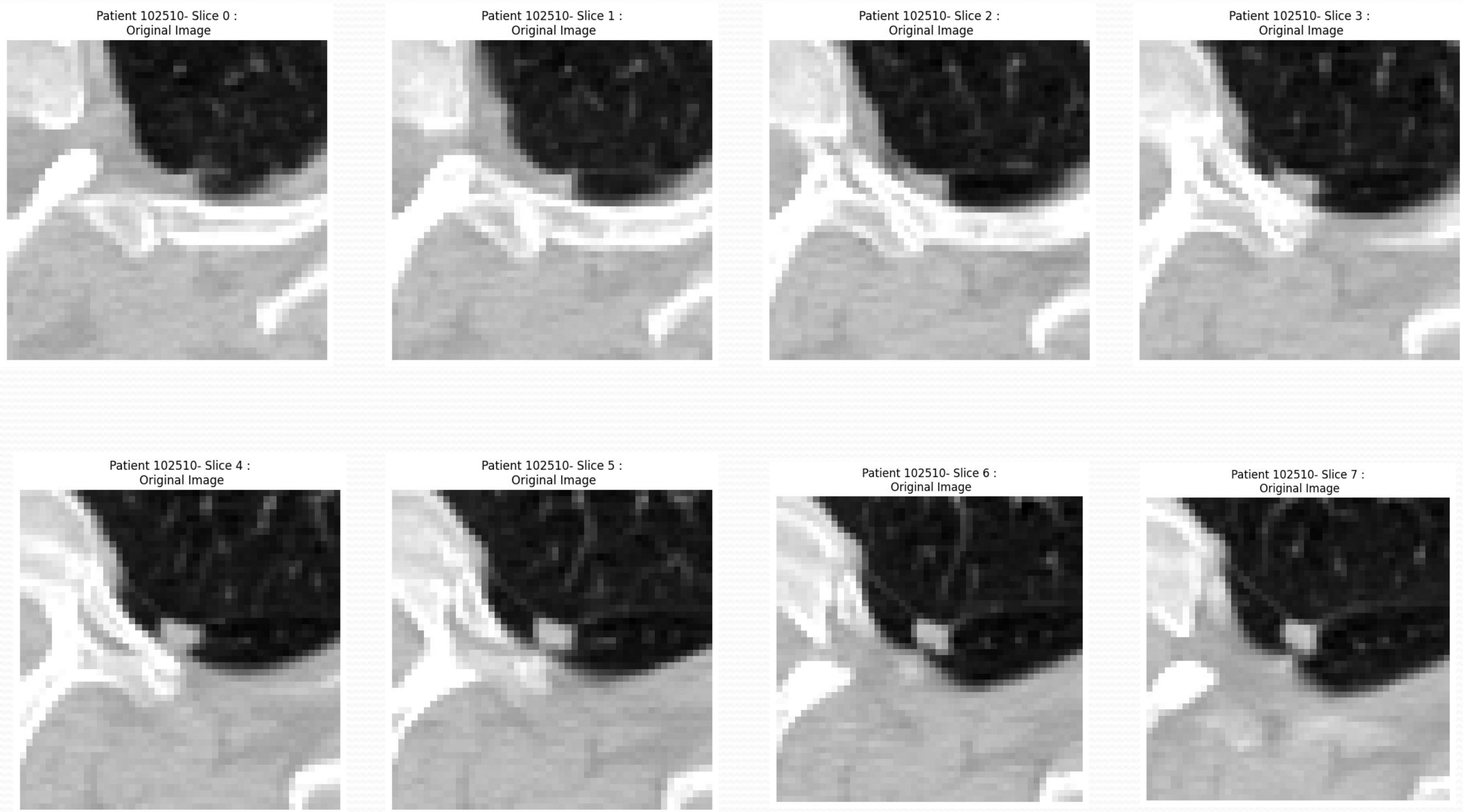






Y_true [False] & Y_pred = [False] -> Size= 112 Nodule

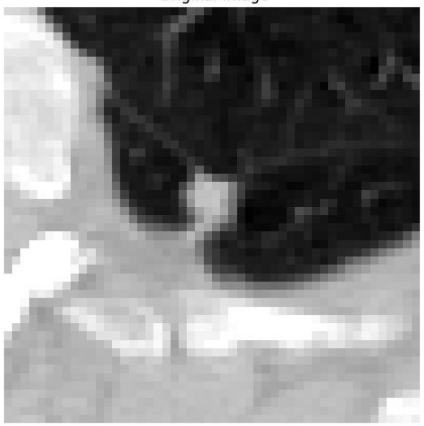
102510



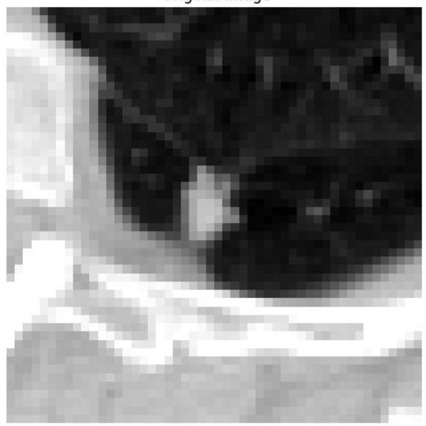
Y_true [False] & Y_pred = [False] -> Size= 112 Nodule

102510

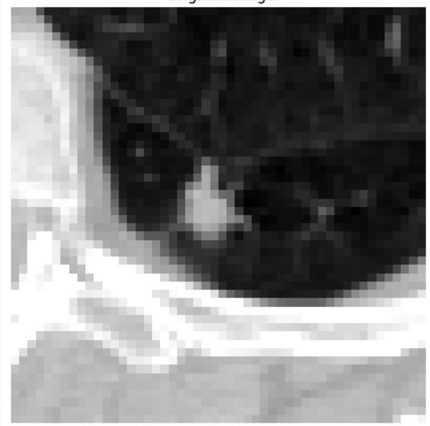
Patient 102510- Slice 8 :
Original Image



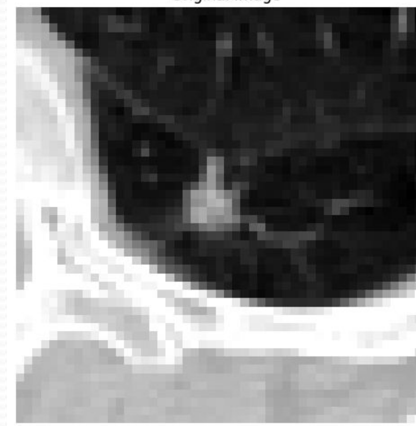
Patient 102510- Slice 9 :
Original Image



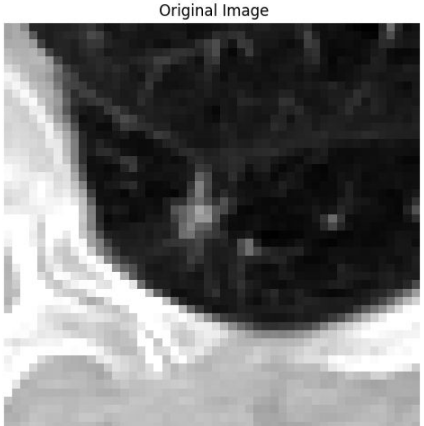
Patient 102510- Slice 10 :
Original Image



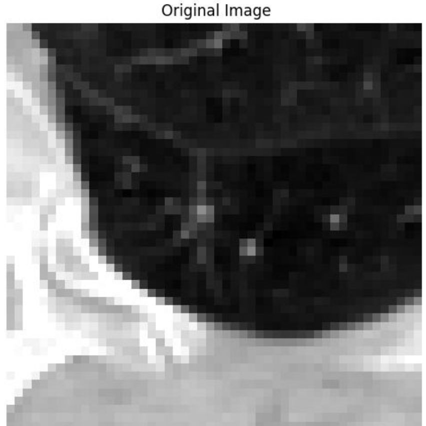
Patient 102510- Slice 11 :
Original Image



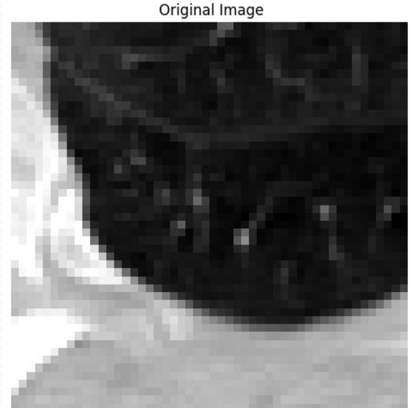
Patient 102510- Slice 12 :
Original Image



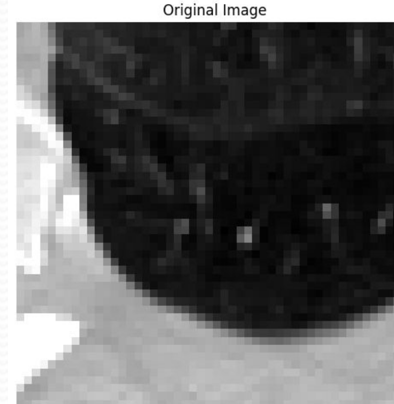
Patient 102510- Slice 13 :
Original Image



Patient 102510- Slice 14 :
Original Image

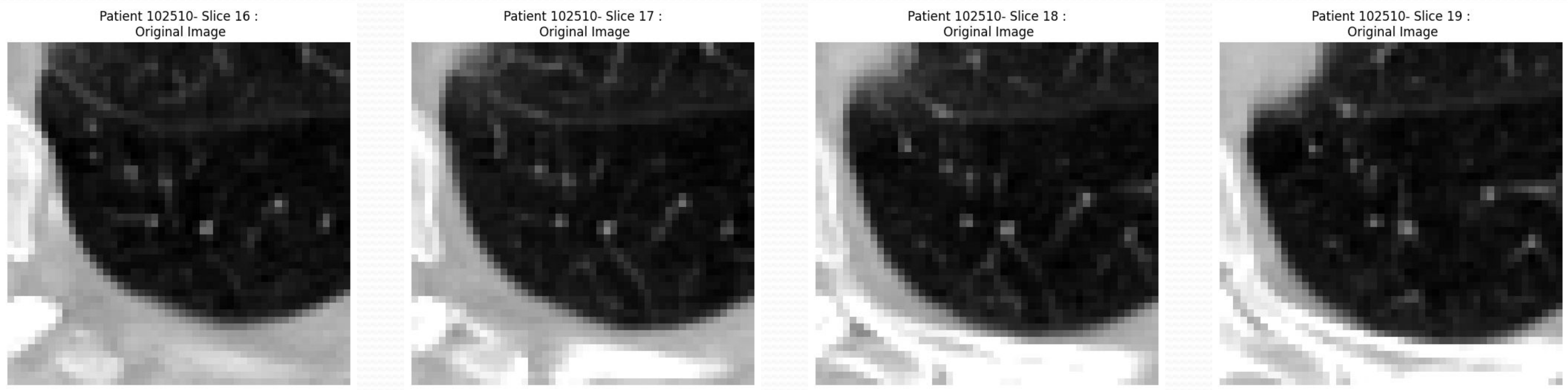


Patient 102510- Slice 15 :
Original Image



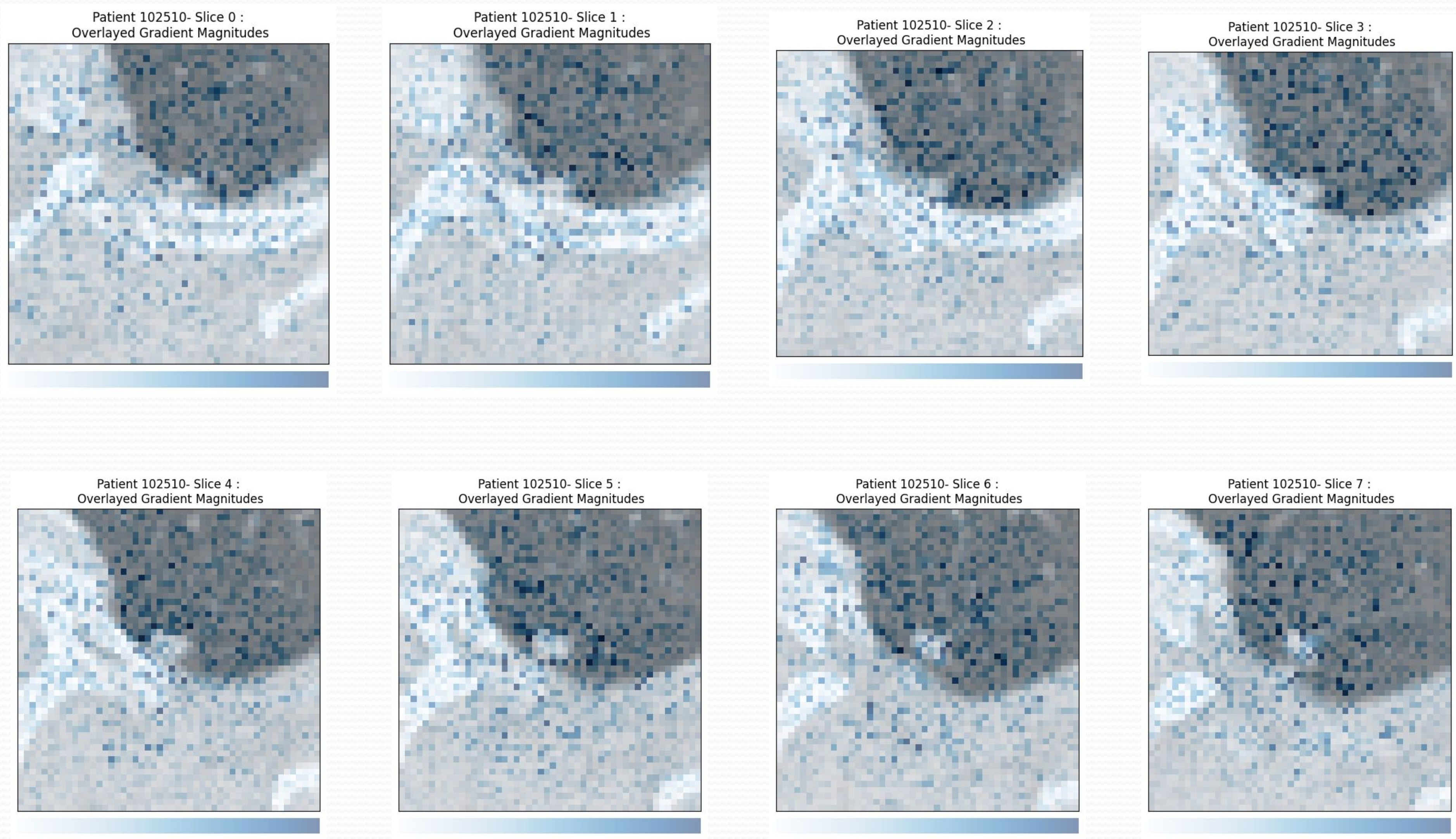
Y_true [False] & Y_pred = [False] -> Size= 112 Nodule

102510



Y_true [False] & Y_pred = [False] -> Size= 112 SM

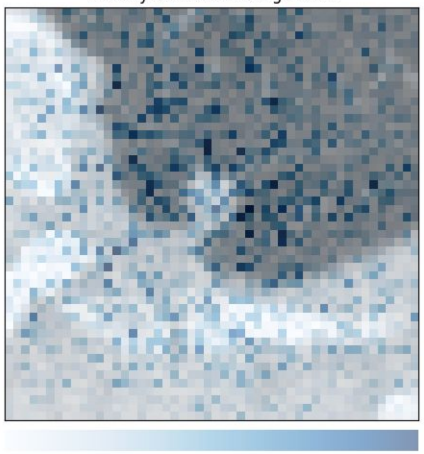
102510



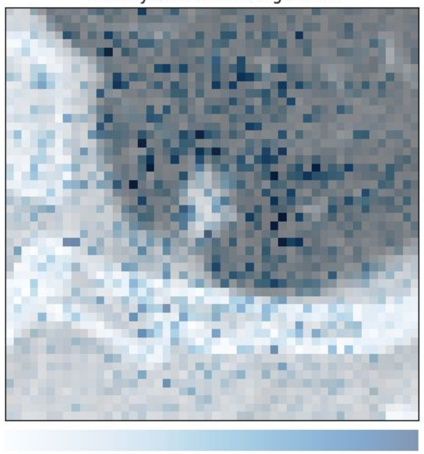
Y_true [False] & Y_pred = [False] -> Size= 112 SM

102510

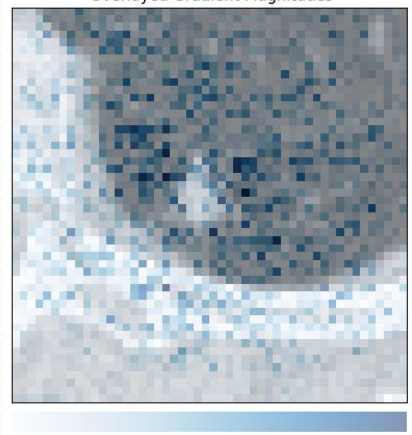
Patient 102510- Slice 8 :
Overlayed Gradient Magnitudes



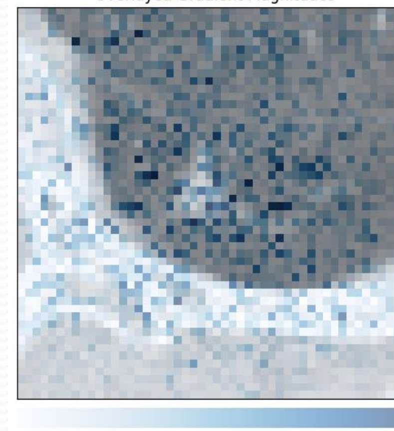
Patient 102510- Slice 9 :
Overlayed Gradient Magnitudes



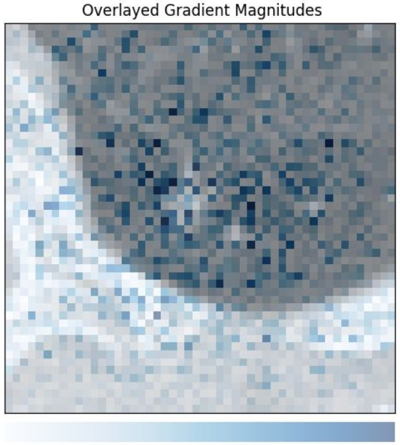
Patient 102510- Slice 10 :
Overlayed Gradient Magnitudes



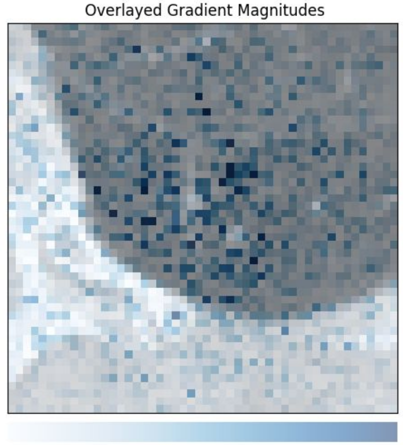
Patient 102510- Slice 11 :
Overlayed Gradient Magnitudes



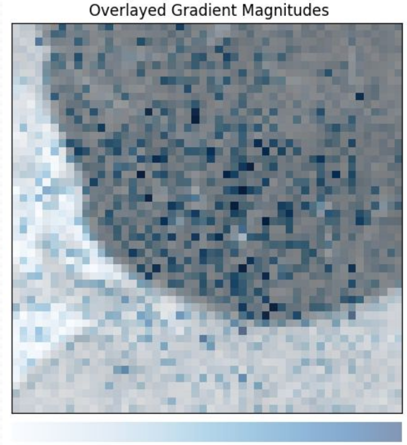
Patient 102510- Slice 12 :
Overlayed Gradient Magnitudes



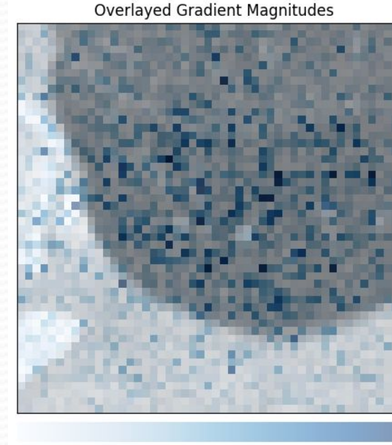
Patient 102510- Slice 13 :
Overlayed Gradient Magnitudes



Patient 102510- Slice 14 :
Overlayed Gradient Magnitudes

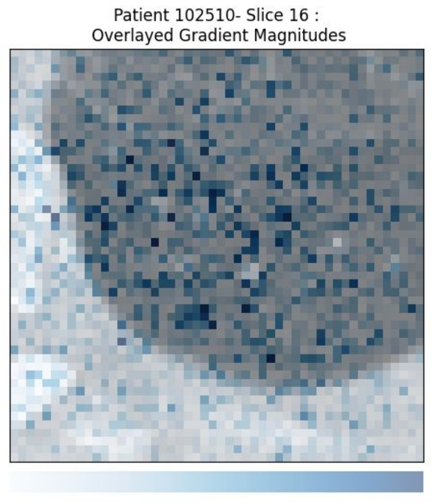


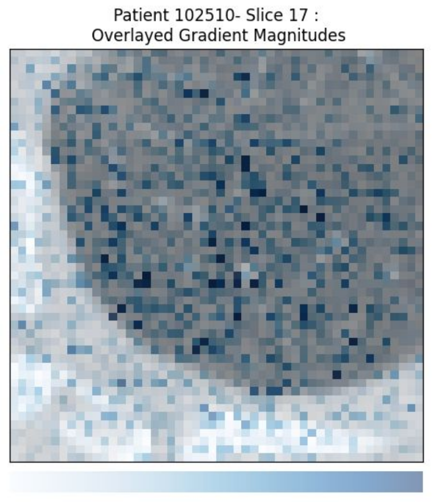
Patient 102510- Slice 15 :
Overlayed Gradient Magnitudes

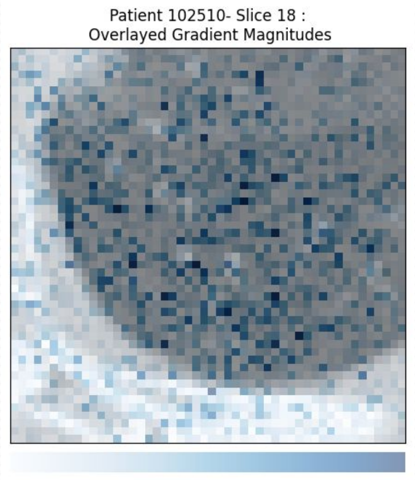


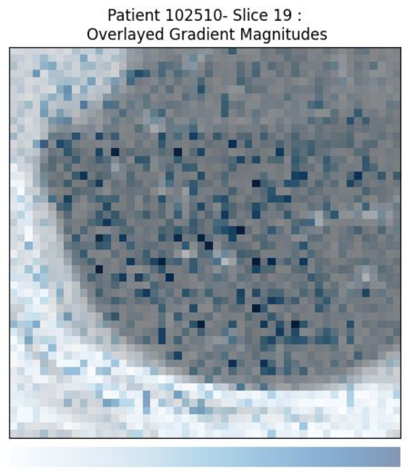
Y_true [False] & Y_pred = [False] -> Size= 112 SM

102510



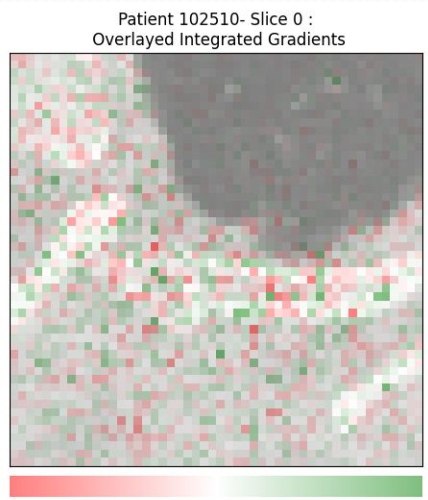


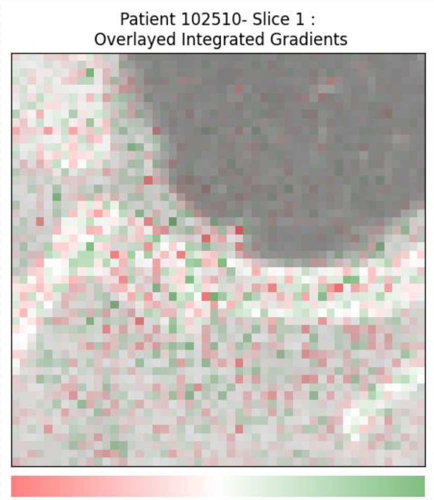


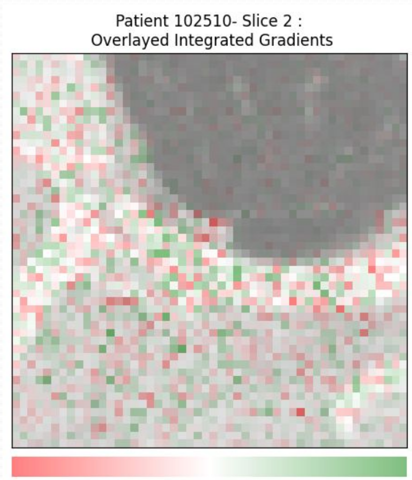


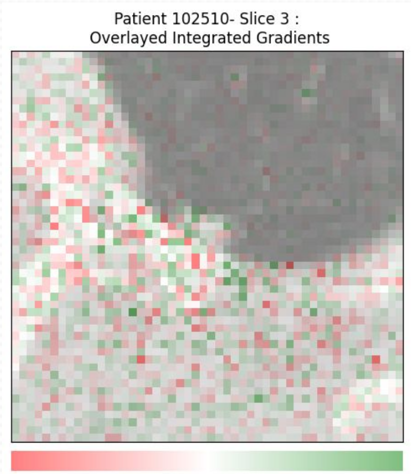
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_all

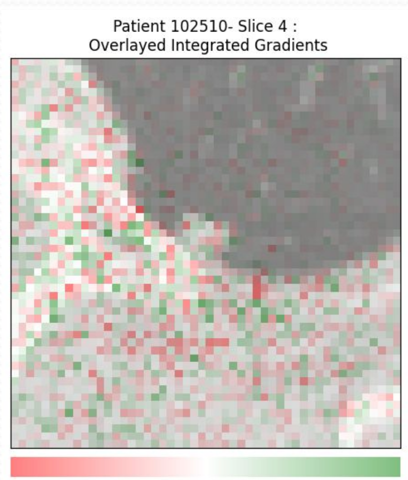
102510

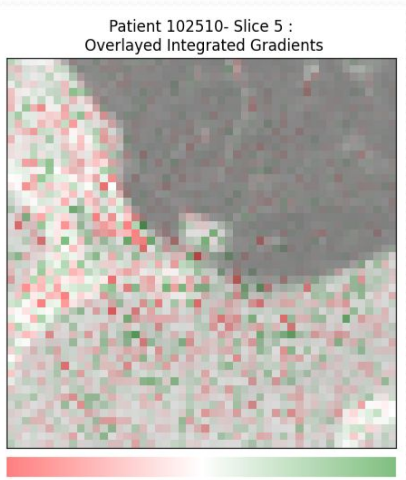


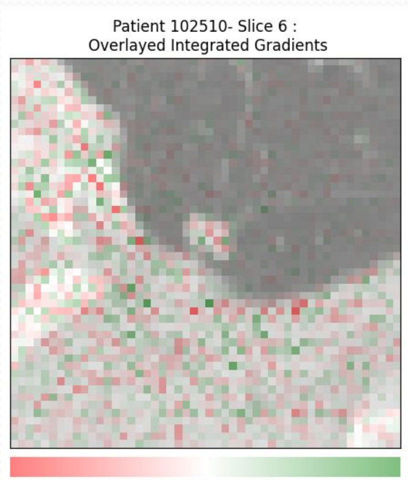


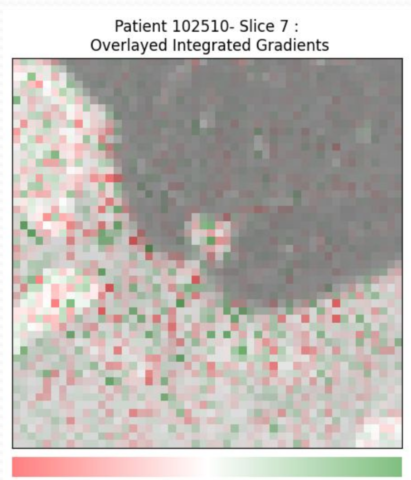






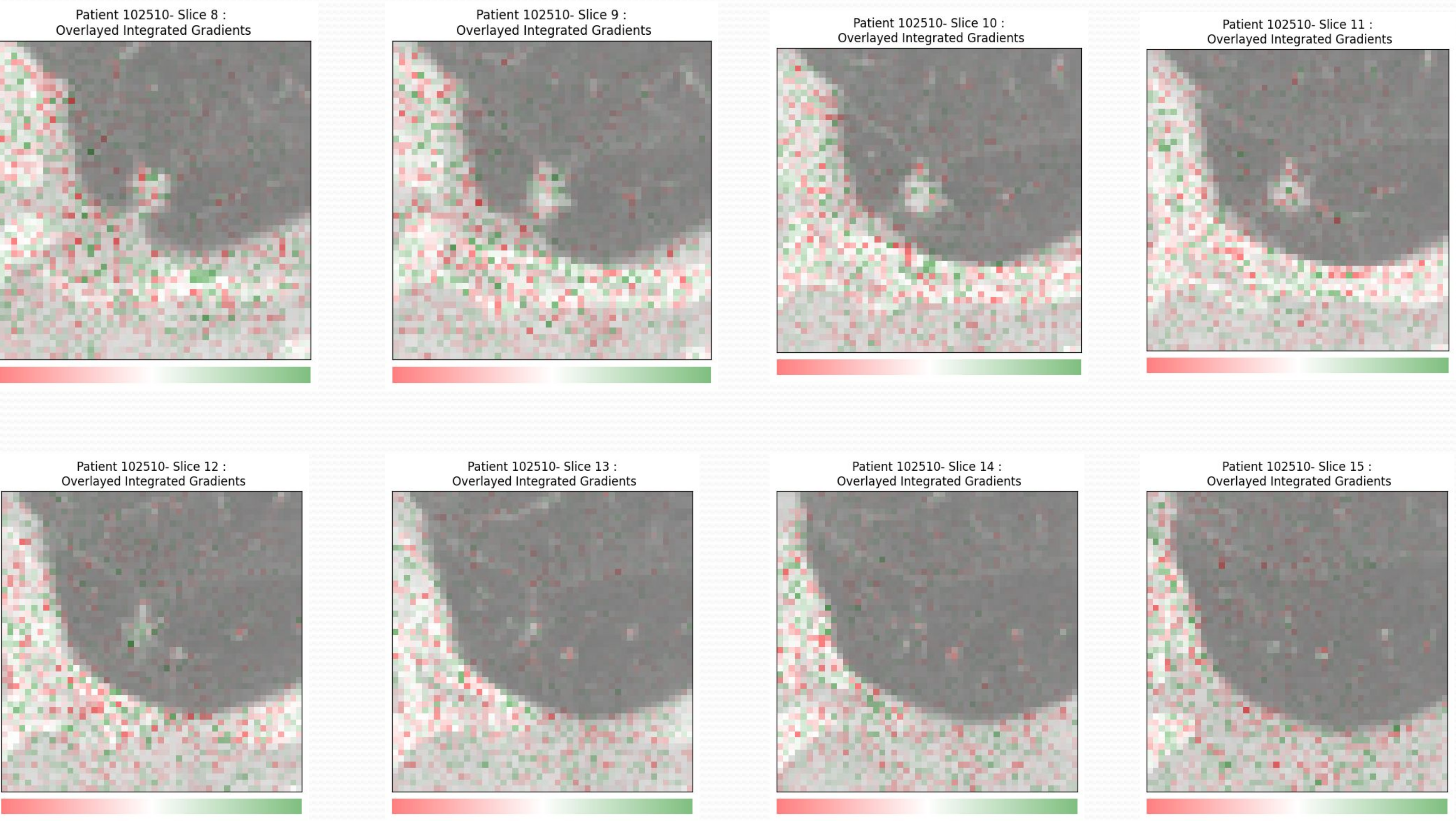






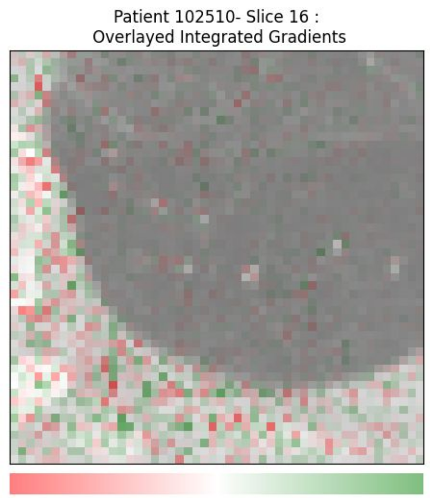
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_all

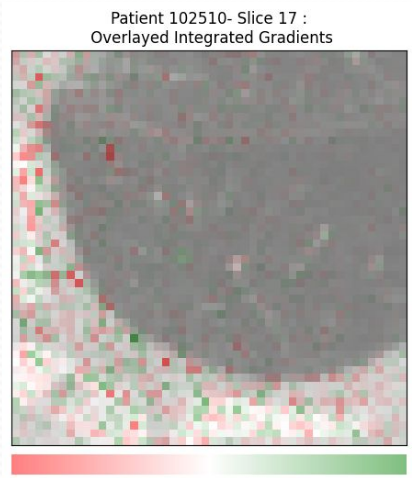
102510

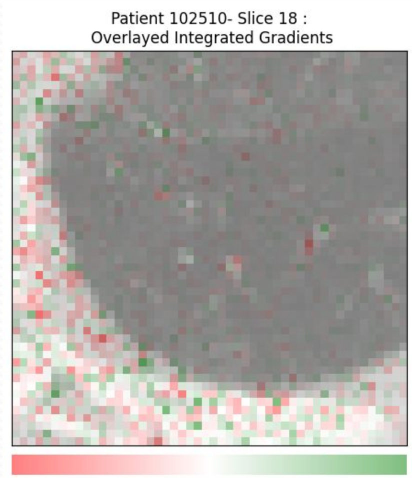


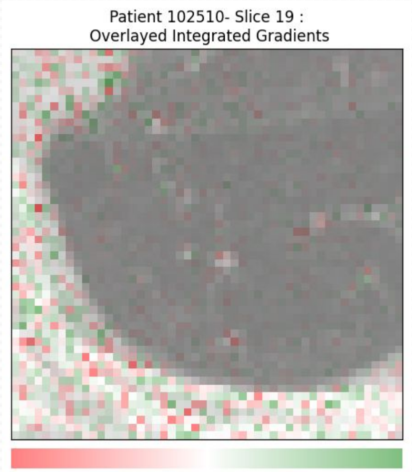
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_all

102510



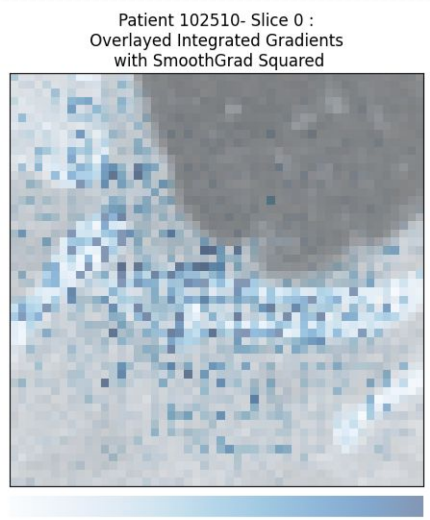


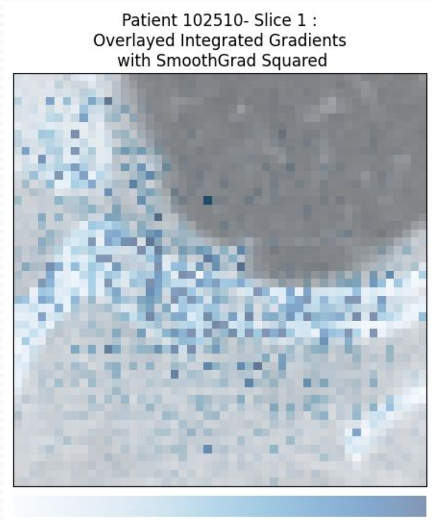


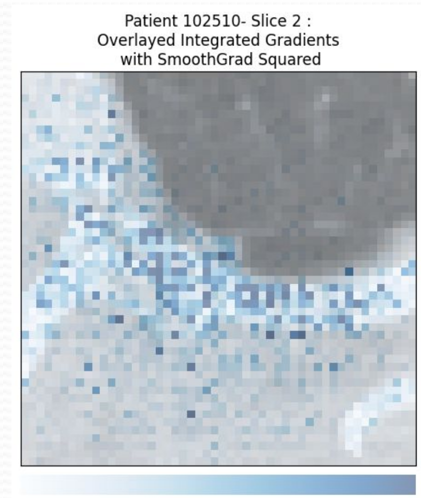


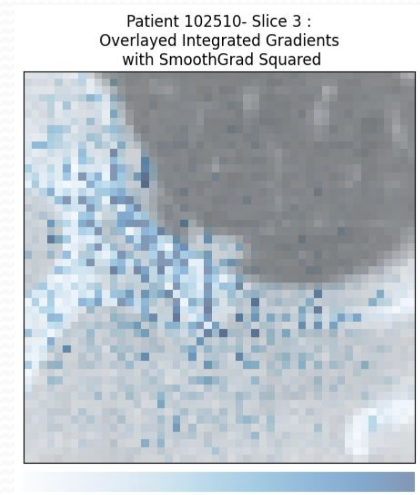
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_abs

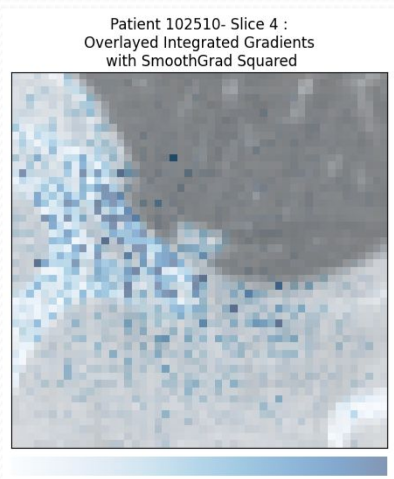
102510

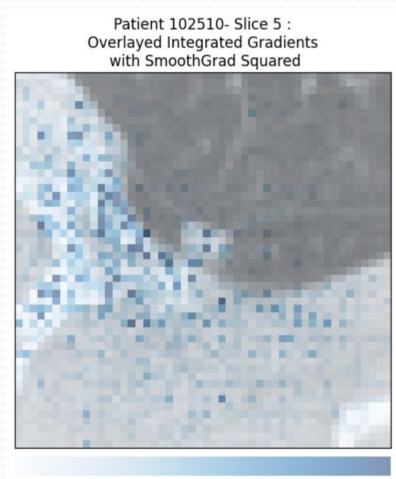


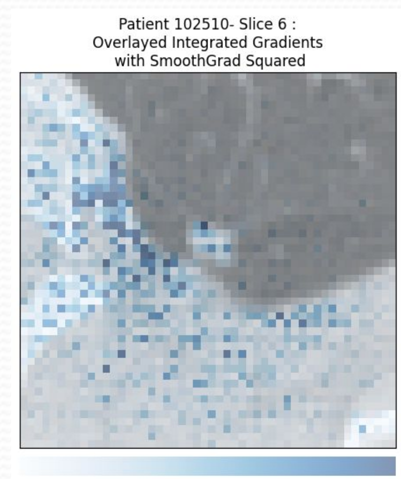


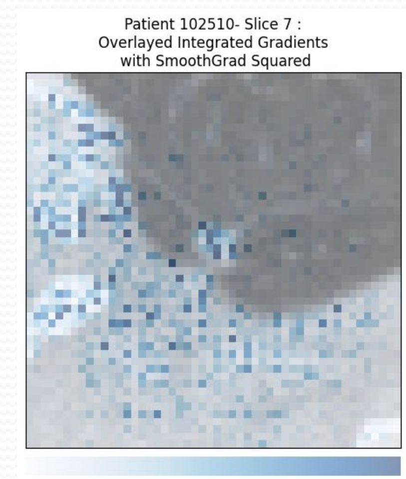






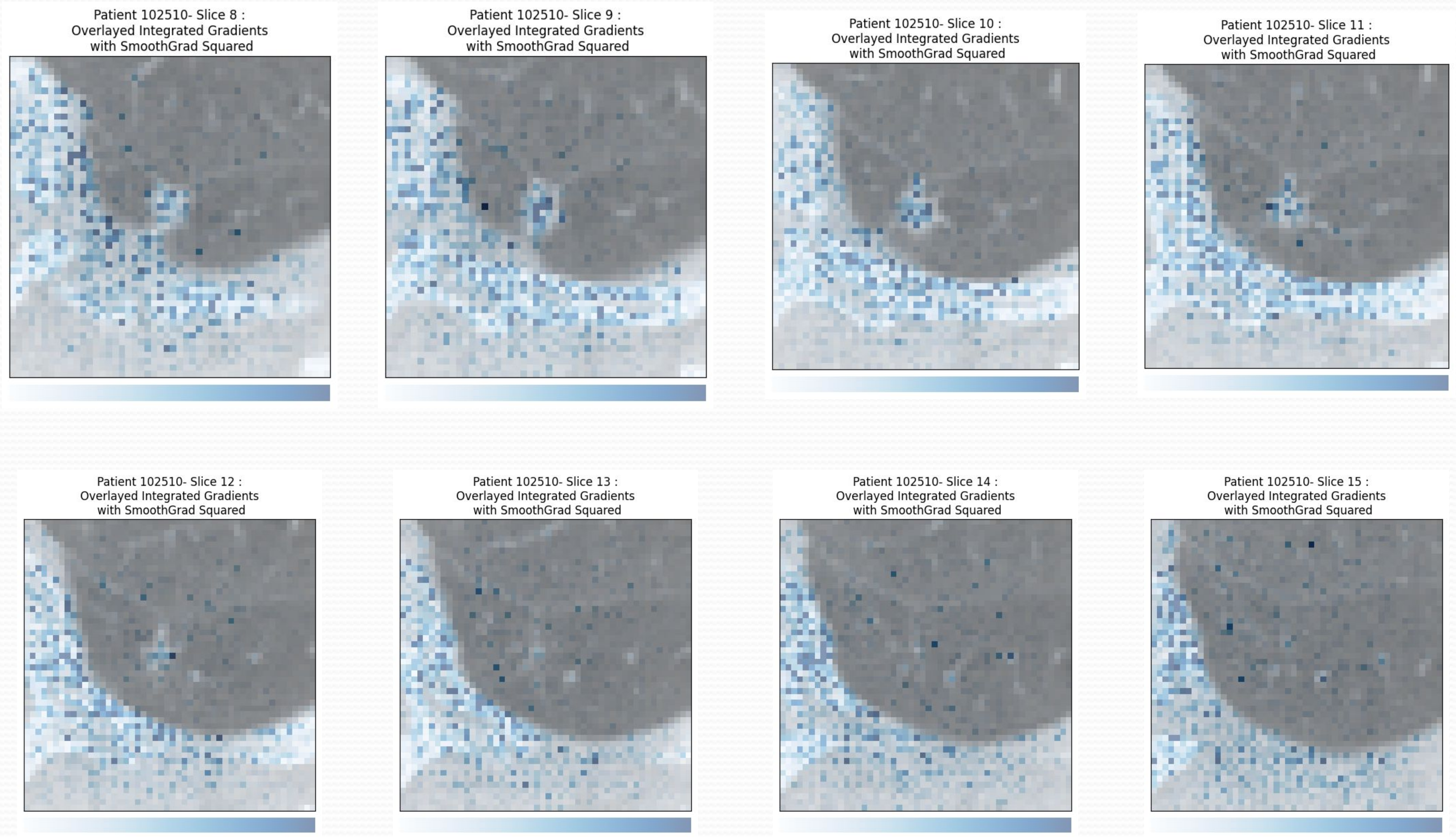






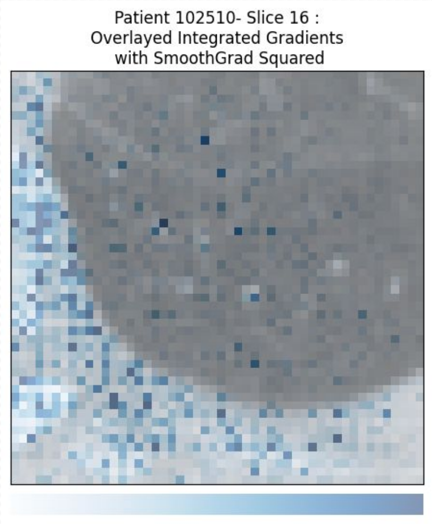
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_abs

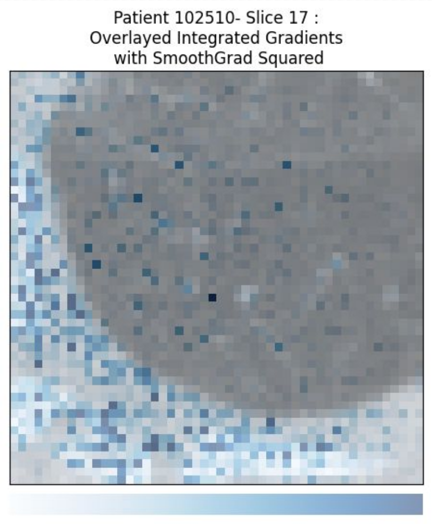
102510

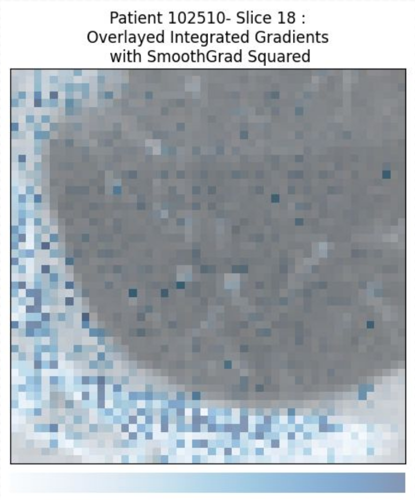


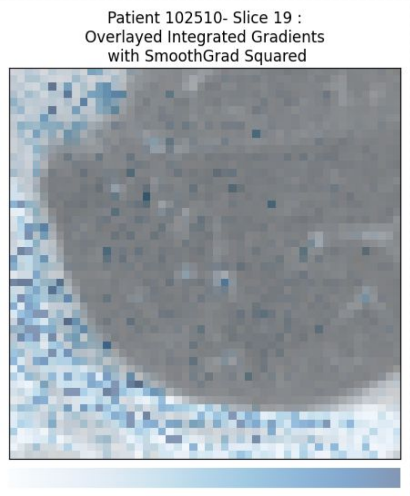
Y_true [False] & Y_pred = [False] -> Size= 112
Ig_abs

102510









References

[1] ANGELOV, Plamen P. et al. Explainable artificial intelligence: an analytical review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 11, n. 5, p. e1424, 2021.

[2]https://towardsdatascience.com/explainable-ai-xai-a-guide-to-7-packages-in-python-to-explain-your-models-932967f0634b

[3]http://xai-tools.drwhy.ai/

[4]A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging