# Explainable AI

DA – 2

Name: Yadukrishnan

Reg. number: 21BAI1210

Slot: F1

Certainly! Let's break down the theoretical aspects of your assignment and project, which involves applying Layer-wise Relevance Propagation (LRP) to segmentation of breast cancer cells using a UNet architecture.

1. Breast Cancer Segmentation:

Breast cancer segmentation involves the process of identifying and delineating regions of interest (such as cancerous tumors) within medical images of breast tissue. Segmentation plays a crucial role in diagnosis, treatment planning, and monitoring of breast cancer.

2. UNet Architecture:

UNet is a convolutional neural network (CNN) architecture designed for biomedical image segmentation tasks. It consists of a contracting path, which captures context, and an expansive path, which enables precise localization. UNet's architecture is characterized by symmetric encoder and decoder paths, with skip connections between corresponding encoder and decoder layers. This design helps in capturing both global context and fine-grained details, making it well-suited for medical image segmentation tasks.

3. Layer-wise Relevance Propagation (LRP):

LRP is an interpretability technique used to understand the decisions made by neural networks. It assigns relevance scores to input features, indicating their contribution to the network's output. LRP propagates the output relevance backwards through the network layers to assign relevance scores at the input layer.

4. LRP-ε (LRP Epsilon):

LRP-ε is an extension of the standard LRP algorithm designed to handle cases where the denominator of the relevance propagation equation approaches zero. This situation typically arises when the sum of weights or activations in a layer is close to zero, potentially leading to numerical instability.

5. LRP-Z Rule:

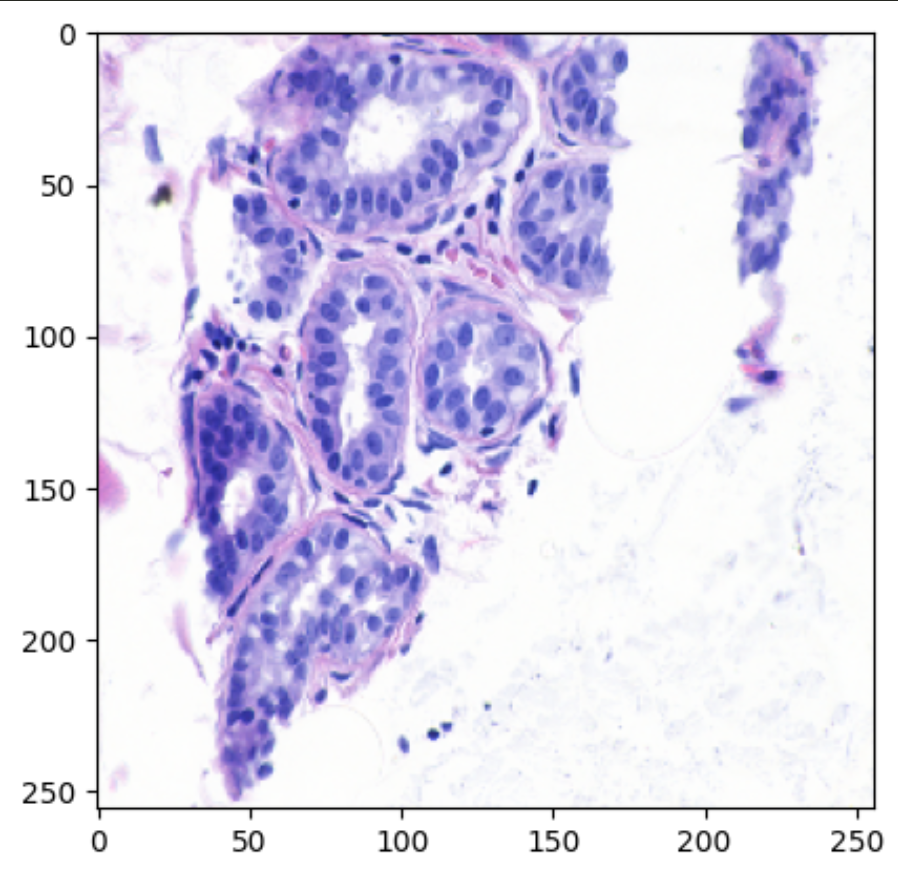
LRP-Z rule is another variant of LRP designed to handle cases where the denominator of the relevance propagation equation is zero or close to zero. This rule provides a systematic approach for propagating relevance in such situations.

6. Deep Taylor Decomposition

Deep Taylor Decomposition (DTD) is a technique used for interpreting the decisions made by deep neural networks (DNNs). It is based on the principle of Taylor decomposition, which approximates a function as a sum of its derivatives at a particular point. In the context of neural networks, DTD decomposes the network's output with respect to its input features. DTD attributes relevance to input features based on their influence on the network's output. It analyzes the gradients of the output with respect to the input features, separating relevance into positive and negative contributions to indicate how each input feature contributes to increasing or decreasing the network's output.

7. Implementation validation

**Input**



**Predicted cancer cells**

**A purple square with yellow dots

Description automatically generated**

**LRP (Epsilon rule)**

**A graph with numbers and a dot

Description automatically generated**

**LRP (Z rule)**

**A graph with numbers and a dot

Description automatically generated**

**Deep Taylor**

**A graph with numbers and a red dot

Description automatically generated**

The above plots highlight the pixels in the image that contributed the most in predicting the output.