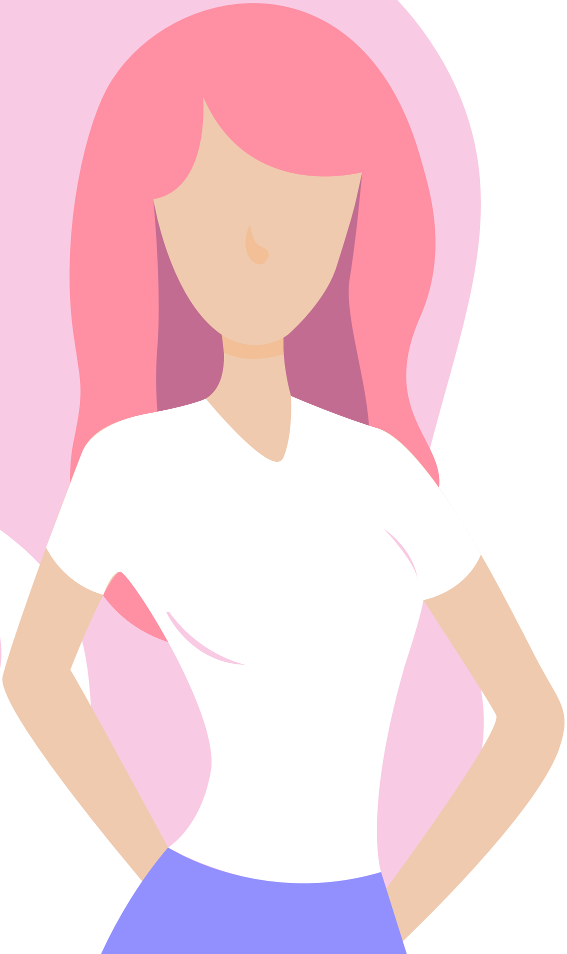


# Detecting Invasive Ductal Carcinoma (IDC) Presence in Breast Histopathological Images

*Building a binary image classification model to detect IDC presence*

# Presentation Outline

**01**

**Introduction & Problem Statement**

**02**

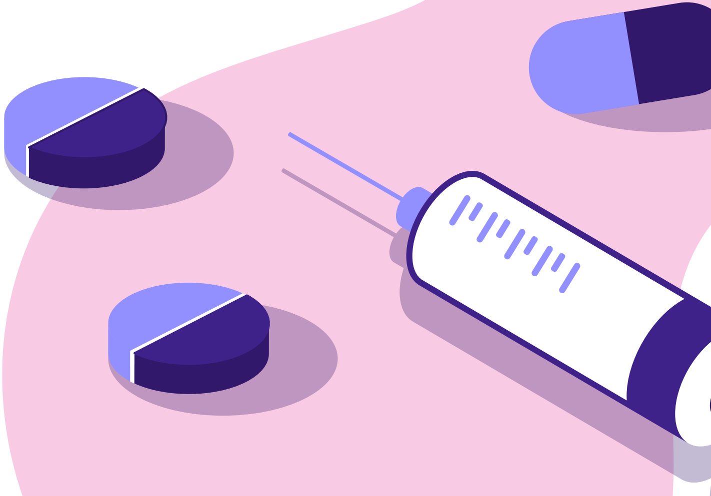
**Dataset Pre-processing**

**03**

**Model Evaluation & Summary**

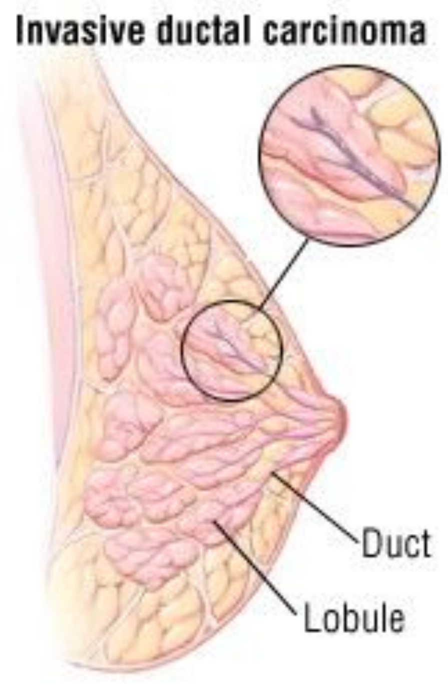
**04**

**Conclusion & Next Step**



# Introduction to Invasive Ductal Carcinoma (IDC)

- IDC accounts for 80% of all breast cancer diagnoses
- IDC grows in the milk duct and invades breast tissues outside the duct
- Pathologists identify IDC through biopsy, and examine the tissue for spread of IDC to assign it an aggressiveness grade

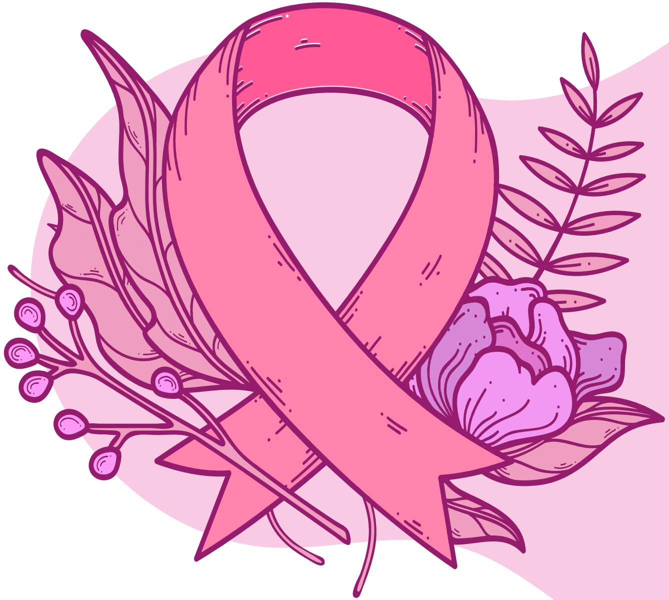


# Difficulty in Analysing Histopathological Images

- The analysis of breast cancer histopathological image is time-consuming and inaccurate
- Barriers to accurate image analysis:
  - Go through swathes of benign regions to identify areas with IDC
  - Variability of appearance in H&E stained areas
- Machine learning approach will increase efficiency in detecting IDC

# Problem Statement

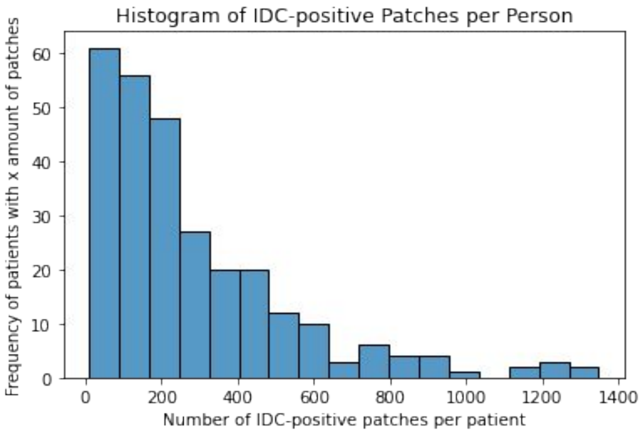
A research firm has hired a team of data scientists to use deep learning methods to help **diagnose the presence of IDC in breast histopathology images**.  The objective is to be to **classify IDC presence/absence** and obtain a reasonably high **Balanced Accuracy** and **Recall** Score.

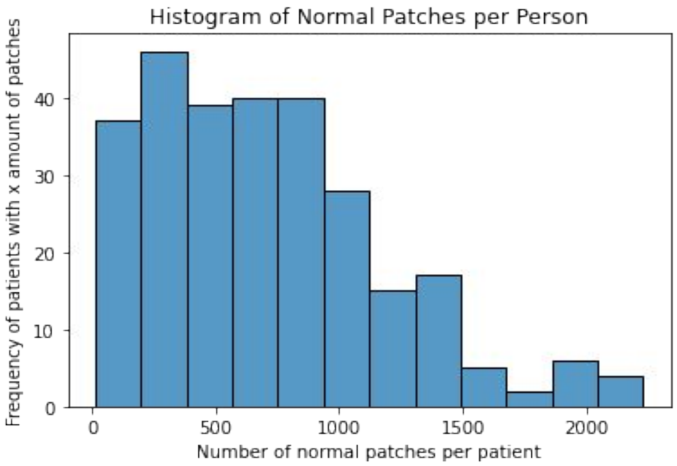


# Data Pre-processing

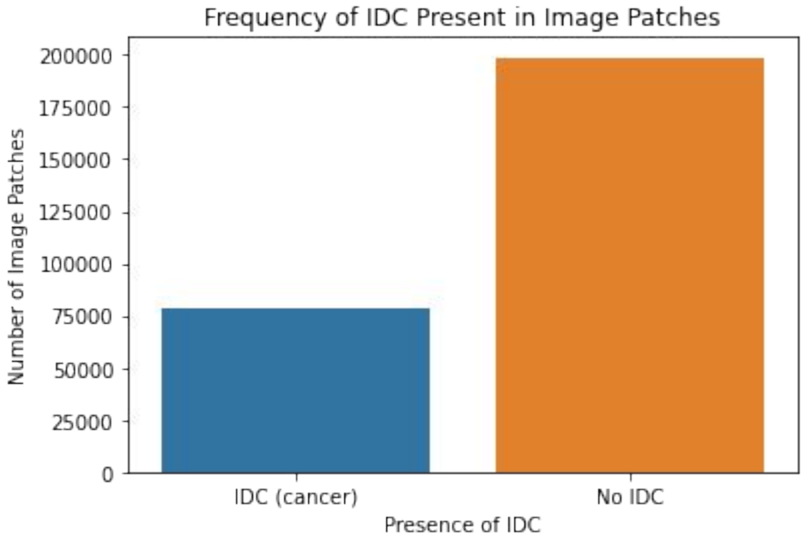
# Image Patch Extraction

- Dataset consist relevant coloured image patches (50 by 50) extracted from Whole Slide Images
- Image are either annotated “IDC-positive” or “normal”
- Each patient had a mix of “IDC-positive” and normal image patches



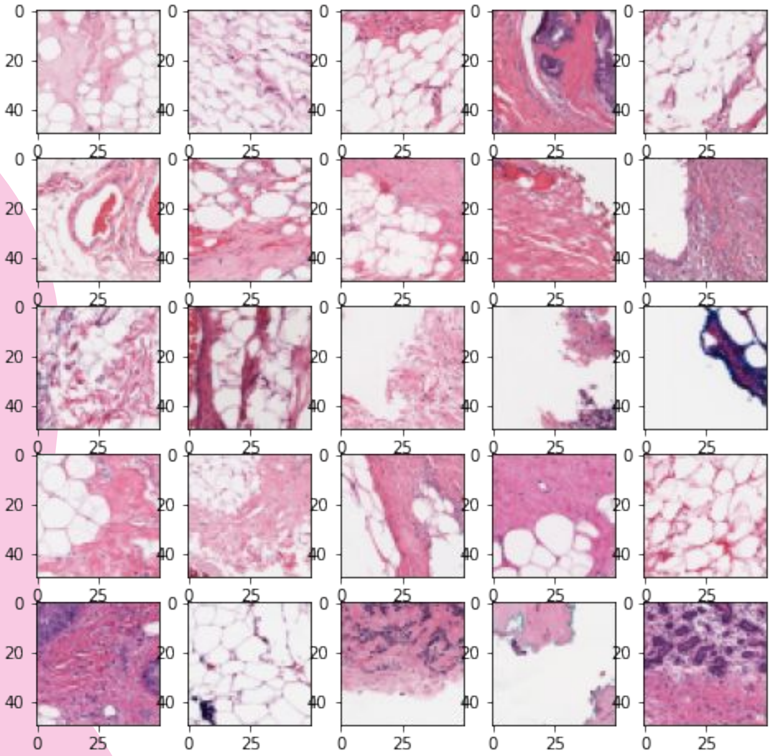



# Class Imbalance

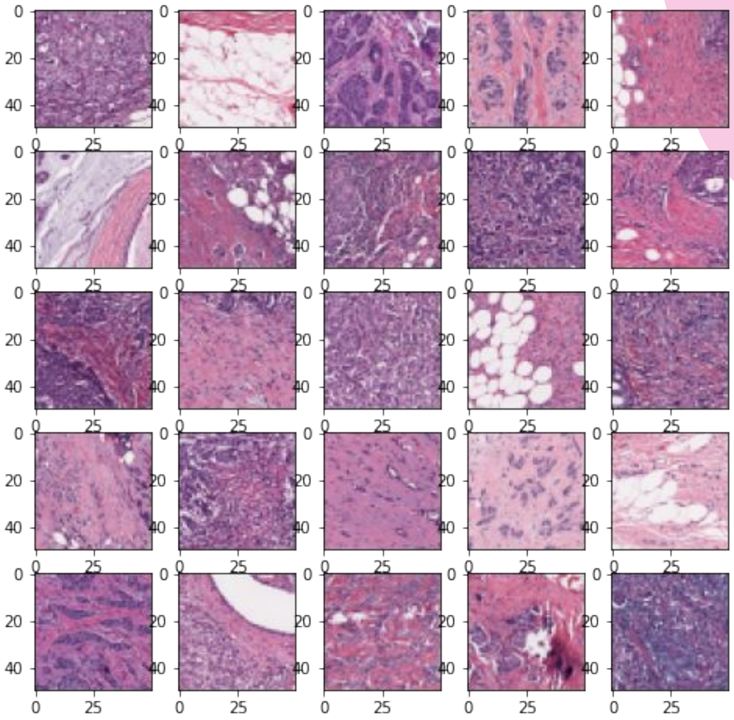


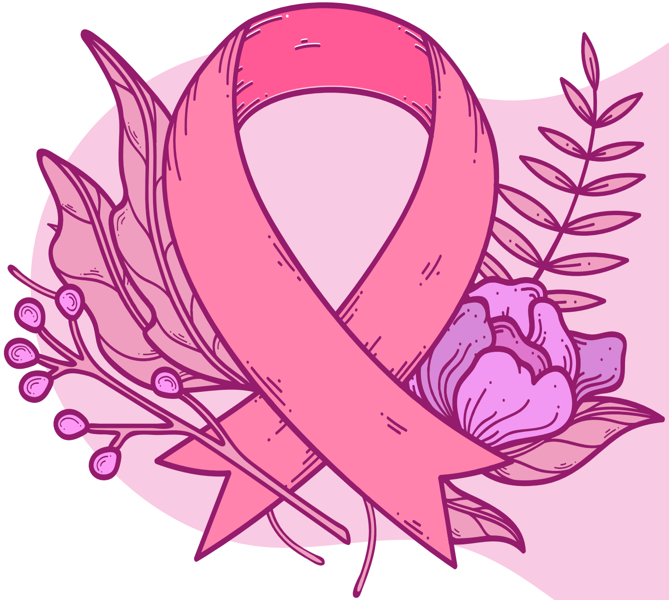

- Number of Scans with IDC is 39.6% that of scans with no IDC
- Downsampling was chosen:
  - Large dataset
- Randomly sampled equal number of patches from both classes and shuffled them before splitting and training the model

# Normal Image Patches



# IDC Patches





# Model Evaluation

# Model Workflow

**Pre-processed Data**

**Baseline Model**

1 convolutional layer

2 fully connected dense layer

**VGG16**

5 convolutional layers

3 fully connected dense layers

**Custom Model**

2 convolutional layers

3 fully connected dense layers

**Resnet50**

48 convolutional layer

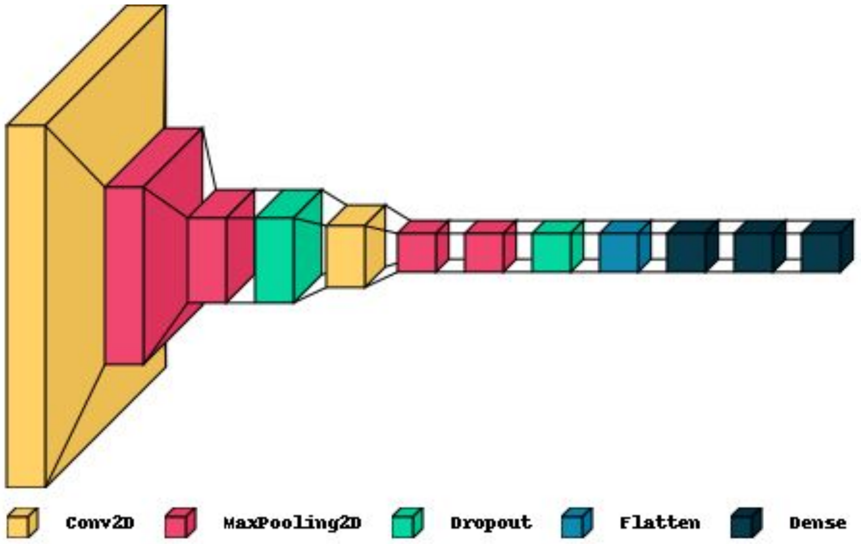
1 fully connected dense layer

**Custom model is chosen and further tuned and trained with training dataset**

# Model Comparison

| Baseline | Custom Model (Chosen Model) | VGG16 | Resnet50 |
|---|---|---|---|
| 73.2% | 81.5% | 80.2% | 70.0% |
| Balanced Accuracy | Balanced Accuracy | Balanced Accuracy | Balanced Accuracy |
| • Baseline<br>• Recall : 0.49<br>• Specificity: 0.97 | • Best recall: 0.76<br>• Specificity: 0.87 | • Recall: 0.66<br>• Specificity: 0.94 | • Worst recall: 0.43<br>• Best specificity: 0.97 |

# Model Evaluation

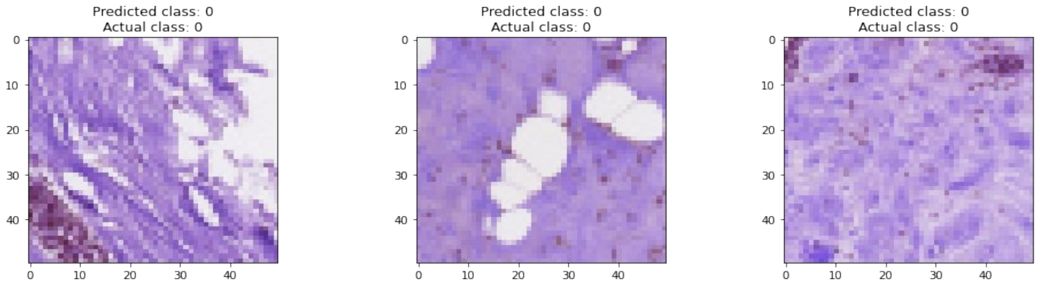



Best scores were obtained when:

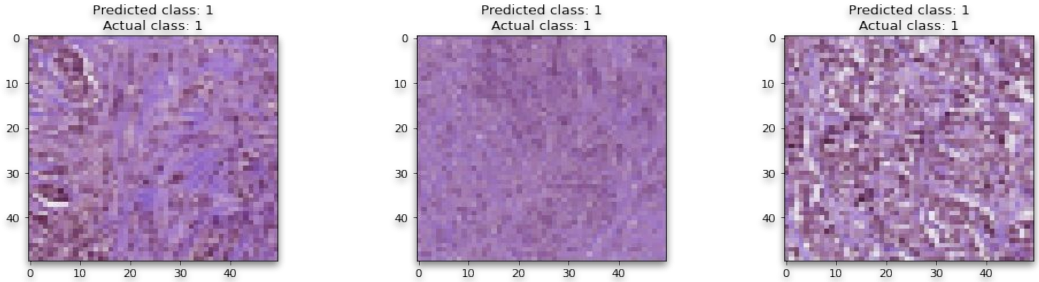
- Image was augmented slightly
- Dropout rate for last dropout layer increased from 0.25 to 0.5
- Dense layers = 3 as opposed to 2 or 4
- Absence of Batch Normalization
- Batch size = 512

# Model Predictions

## Correctly classified as normal patches

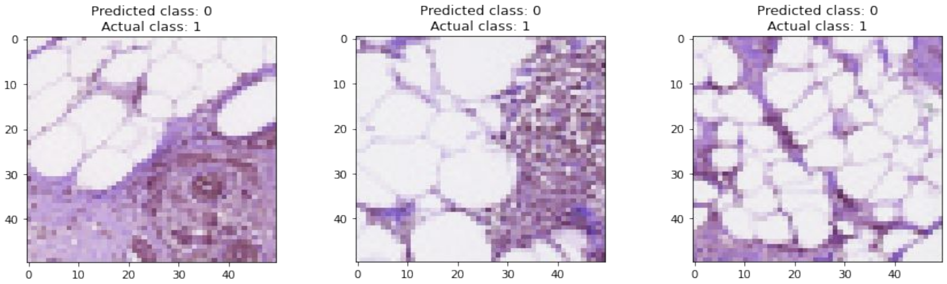



## Correctly classified as IDC-present patches

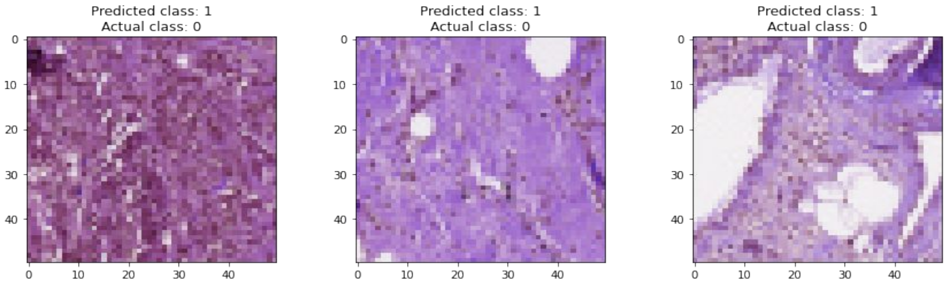


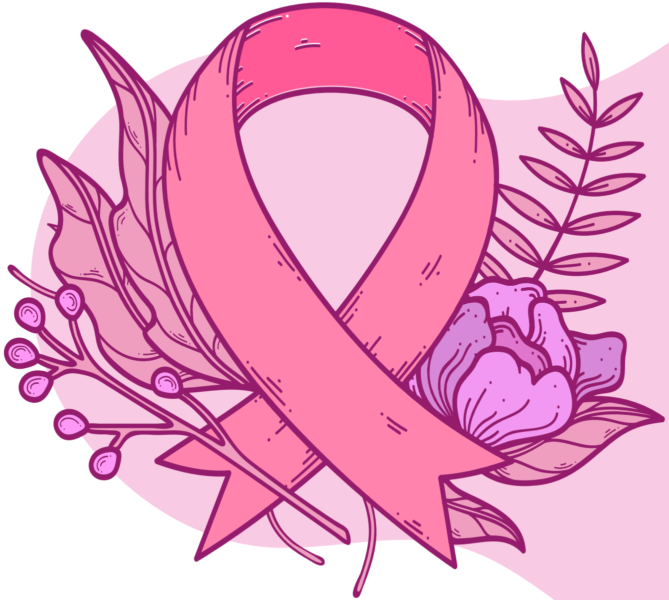
# Misclassification Analysis

## Predicted normal when IDC is present




## Predicted IDC-present when image patch is normal





# Conclusion & Next Steps

# Conclusion and Limitations

- Model achieved a **81.5% balanced accuracy score** and a 76.0% recall score
- **Improve productivity**
  - < than 1 minute to predict for IDC-presence in 39,000 images

Challenges

- Dataset Limitations
  - Unable to train on features connecting image patches
  - Absence of data on normal patients

# Next Steps

### Other Methodologies

- Other state-of-the-art models (ensemble CNN)
- Image Segmentation (segmenting regions of image for meaningful analysis)

### Whole Slide Imagery

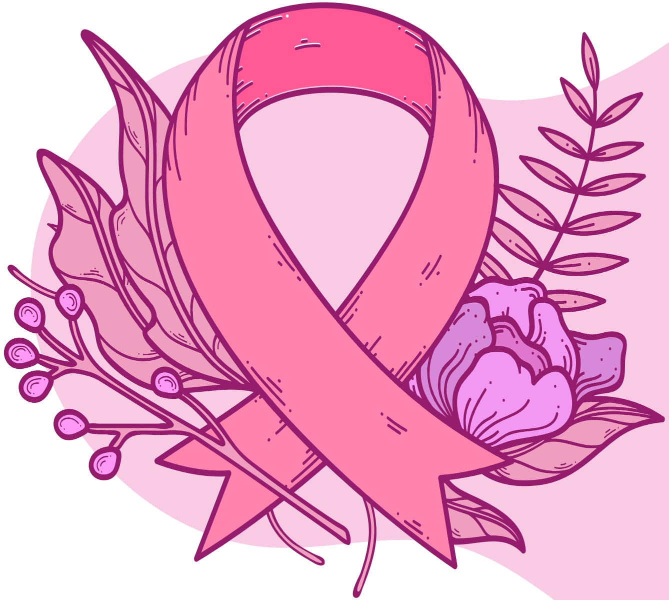
- Utilise whole slide imagery as opposed to relevant fragments

### Aggressiveness Rating

- Automatically assign ratings after determining IDC presence

# Thanks

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik.**

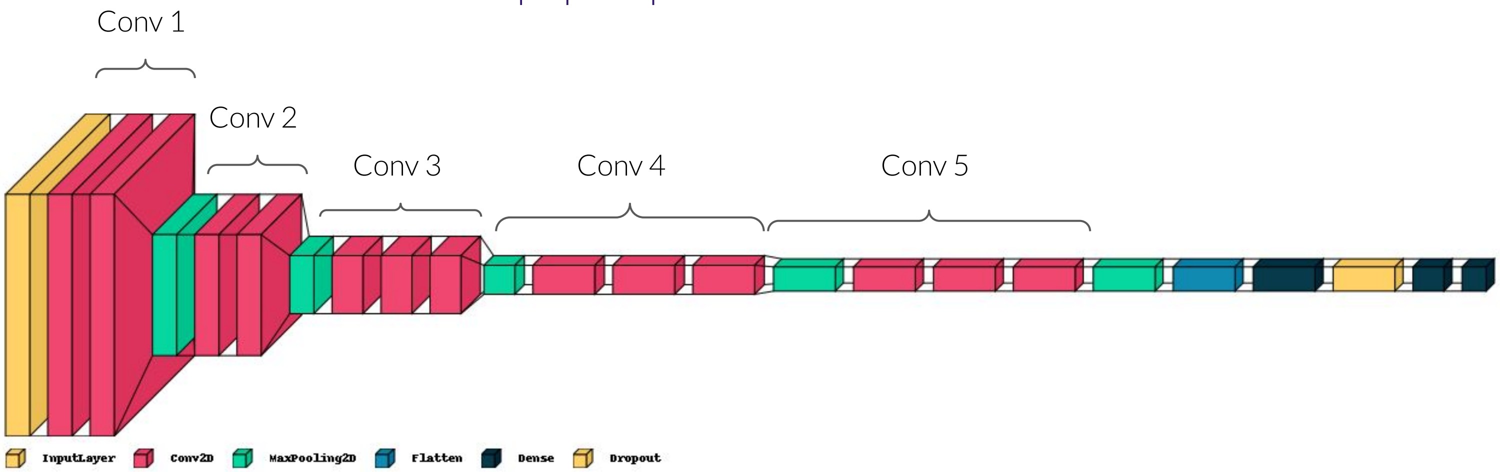


# Appendix

# Custom VGG16

Consists of:

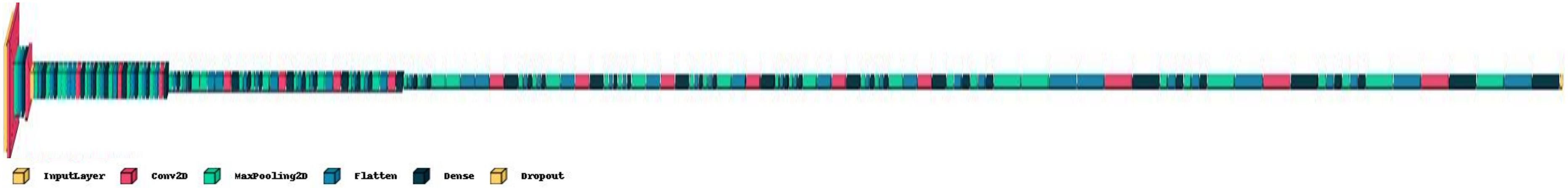
- 5 convolutional layers
- 3 fully connected dense layers
- Epochs = 10
- Steps per epoch = 300



# Resnet50

Consists of:

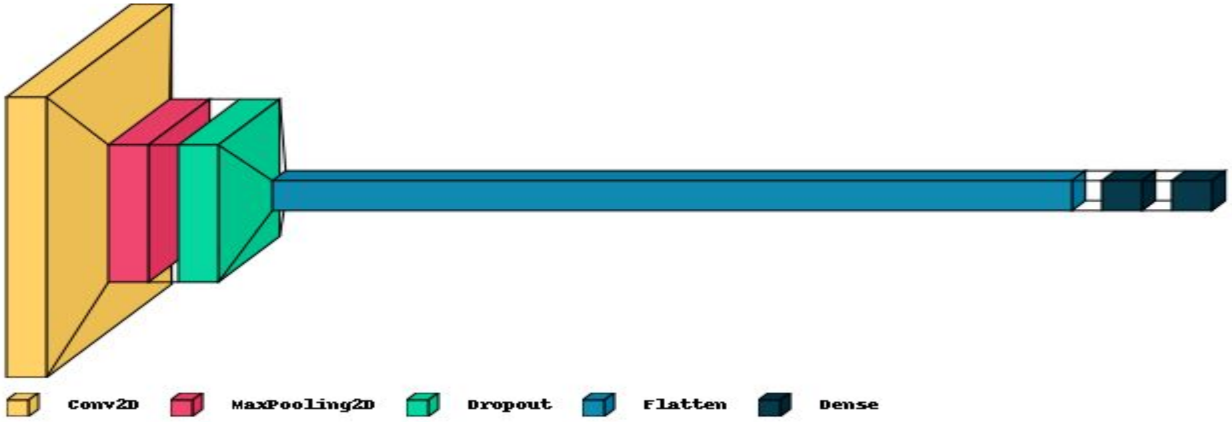
- 48 convolutional layers
- 1 max pool
- 1  fully connected dense layer
- Steps per epoch = 300
- 15 epochs



# Baseline

Consists of:

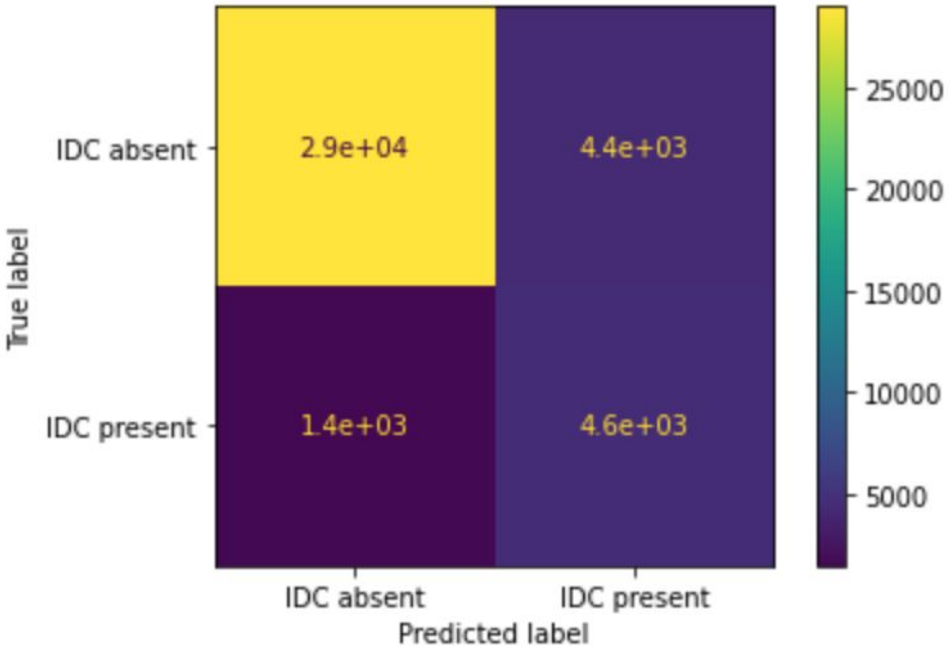
- 1 convolutional layer
- 1 max pool
- 2  fully connected dense layer
- Steps per epoch = 200
- 60 epochs



# Fine-tuning Custom Model

| Model Versions | Image Augmentation | Dropout Layer Value | Hidden Layer Node | Batch Normalization | Balanced Accuracy | Specificity | Recall | Remarks |
|---|---|---|---|---|---|---|---|---|
| **v0** | No | [0.25,0.25] | [128,64,2] | No | 0.806 | 0.889 | 0.722 | |
| **v1** | Yes | [0.25,0.25] | [128,64,2] | No | 0.812 | 0.843 | 0.781 | |
| **v2** | Yes | [0.5,0.5] | [128,64,2] | No | 0.727 | 0.514 | 0.940 | Large number of False Positives |
| **v3** | Yes | [0.25,0.5] | [128,64,2] | No | 0.815 | 0.869 | 0.761 | **Chosen model** |
| **v4** | Yes | [0.25,0.5] | [128,64,2] | Yes (after conv 2) | 0.535 | 0.971 | 0.098 | |
| **v5** | Yes | [0.25,0.5] | [64,2] | No | 0.807 | 0.883 | 0.731 | |
| **v6** | Yes | [0.25,0.5] | [32,16,2] | No | 0.815 | 0.852 | 0.778 | Large number of False Positives |
| **v7** | Yes | [0.25,0.5] | [64,32] | No | 0.818 | 0.843 | 0.793 | Large number of False Positives |
| **v8** | Yes | [0.25,0.5] | [128,64,32,2] | No | 0.802 | 0.777 | 0.828 | |
| **v9** | Yes | [0.25,0.5] | [128,64,2] | Yes (before pooling) | 0.792 | 0.880 | 0.704 | |

# Confusion Matrix




```
array([[28967,  4366],
       [ 1448,  4612]])
```