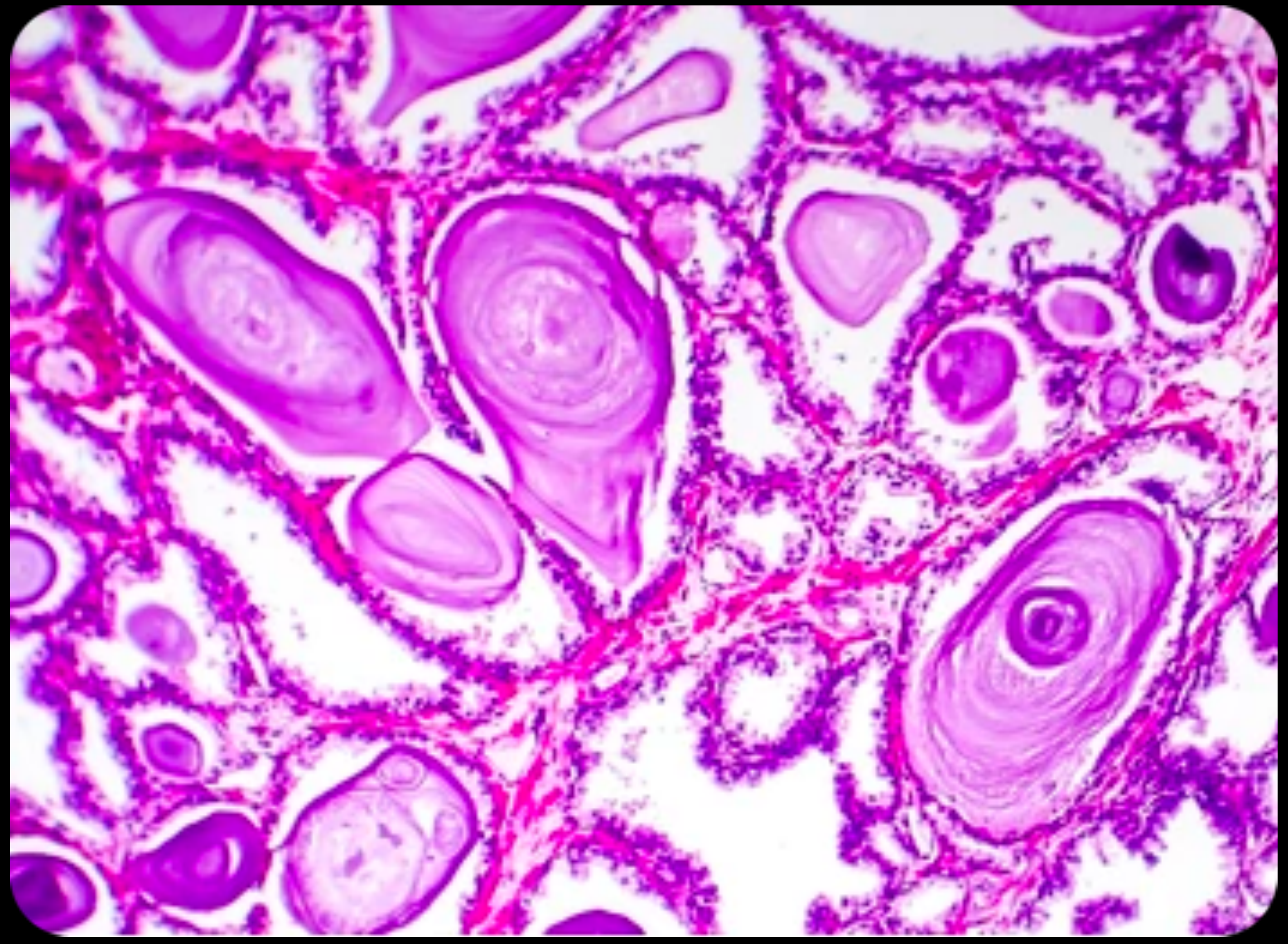
Breast Cancer Detection in Histopathology Images



Aastha Tandon
MSBA,CSU-East Bay

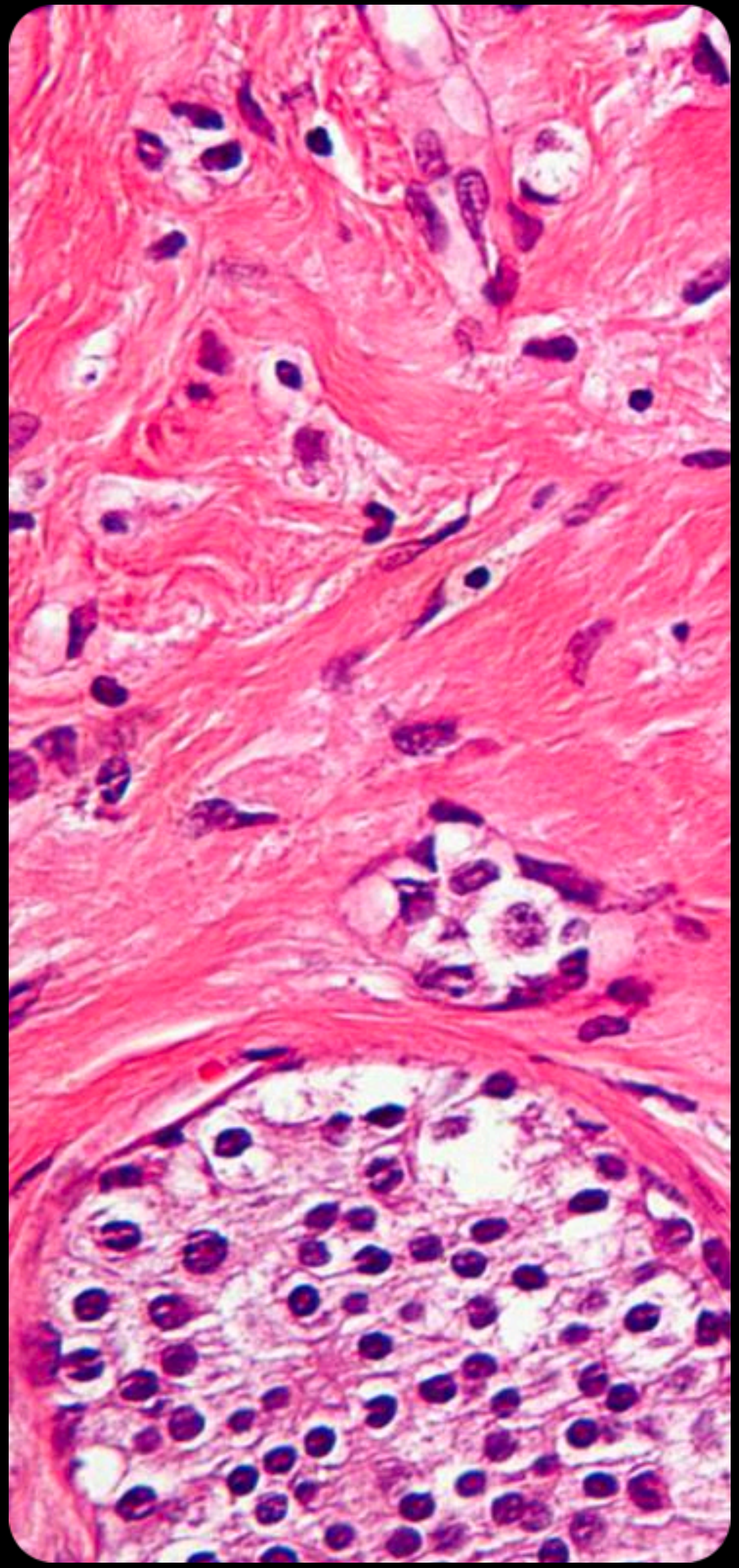
Agenda

Overview

- Breast cancer is the most common form of cancer in women
- IDC (Invasive Ductal Carcinoma) is the most common form of breast cancer
- Mammograms and ultrasound detect tumors, but histopathology determine malignancy at cellular level
- Computer-Aided Diagnostic system are needed to provide objective and consistent interpretation

Problem Statement

Identify breast cancer (IDC) in breast histopathology images using Convolution Neural Network

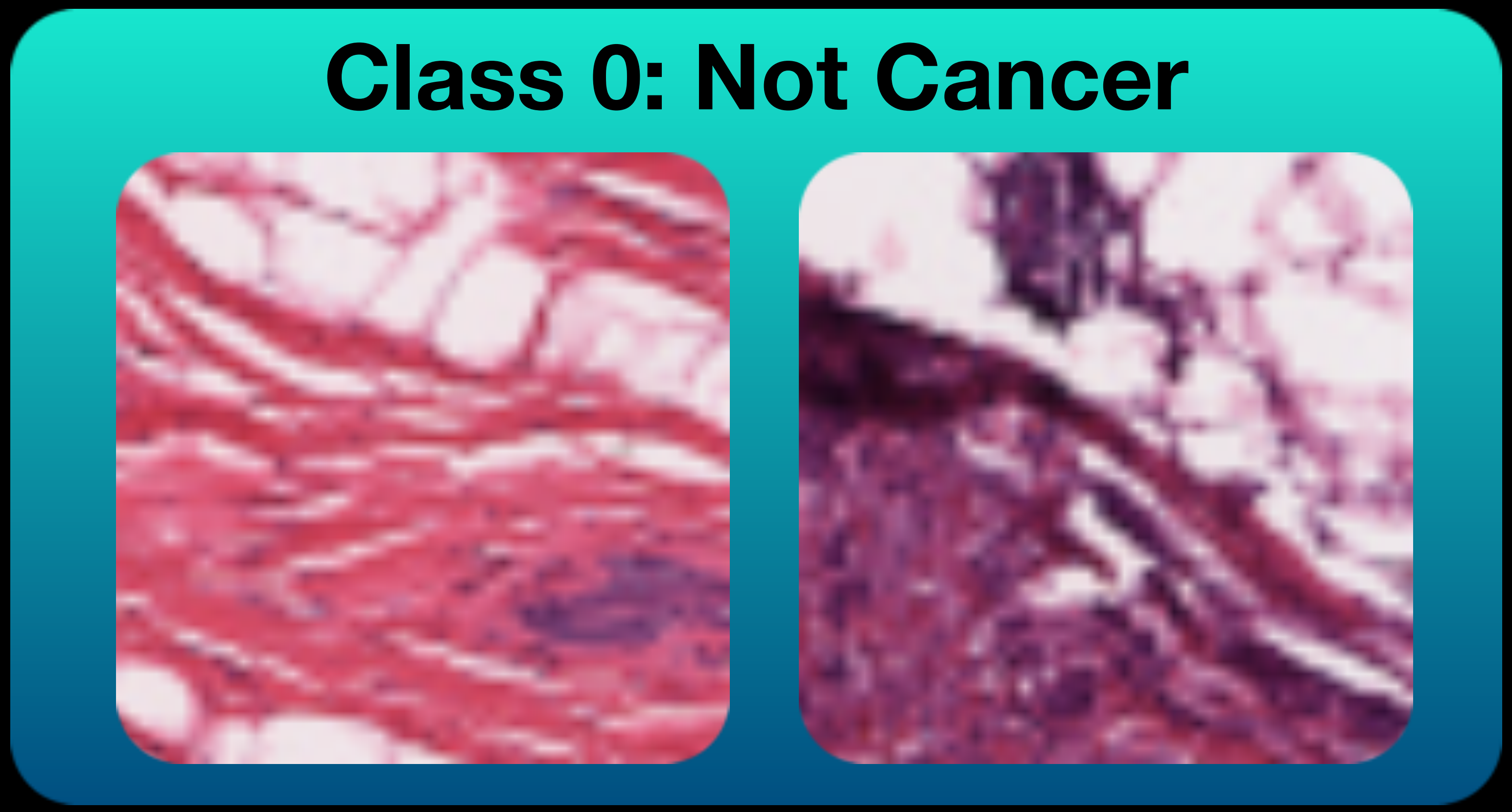


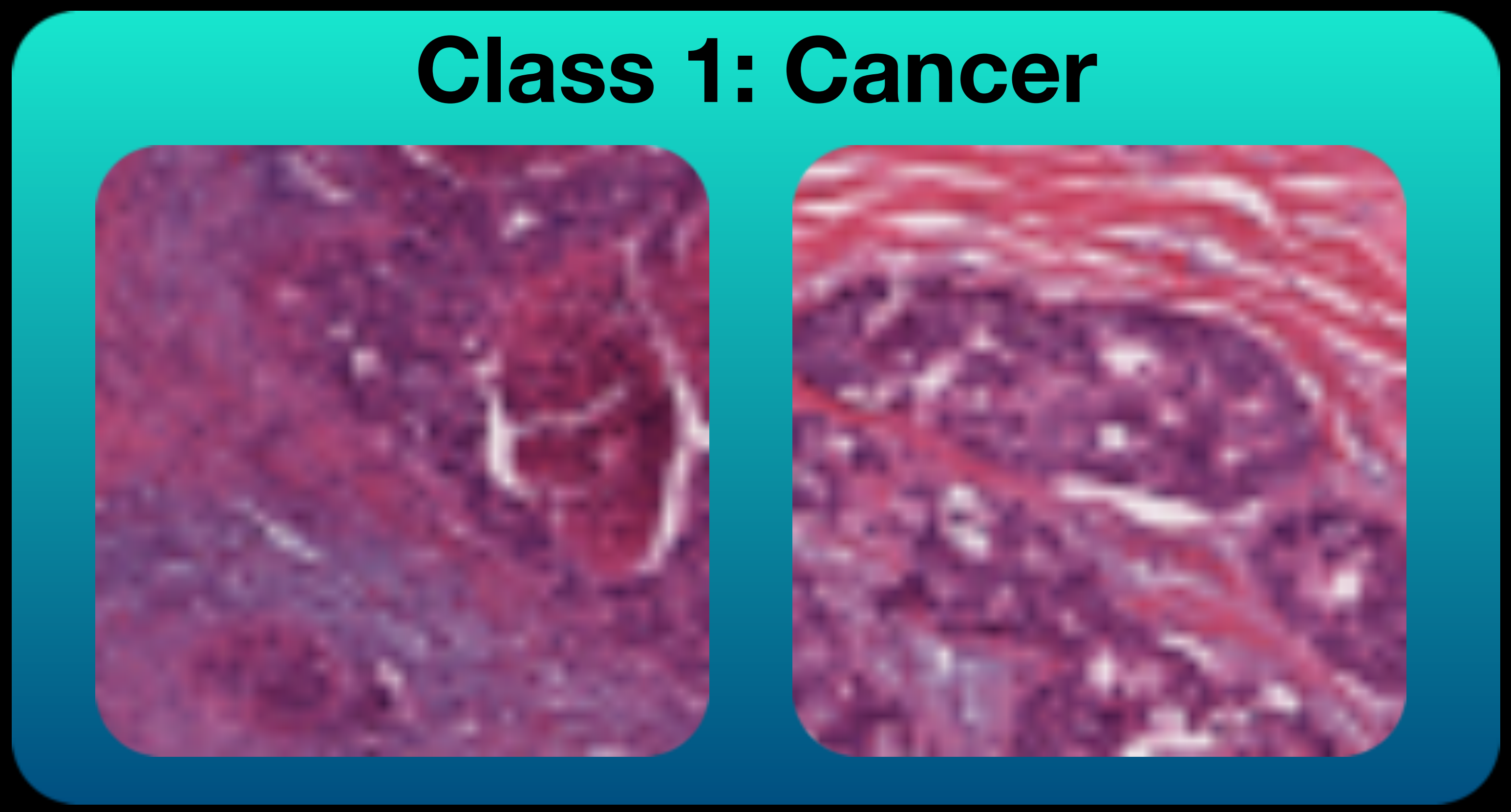
Motivation

- Early detection of Breast cancer
- Prognosis and Treatment Planning
- Assist medical professionals
- Enhance patient outcome and survival rates
- Cost effective solution for breast cancer screening

Dataset

- Contains histopathology images of size 50 x 50.
- Dataset volume:
  - Total images ~ 277K
  - Negative example ~ 198K (not breast cancer)
  - Positive examples ~ 78K (breast cancer)
- Class 0: No Breast Cancer, Class 1: Breast Cancer





Dataset: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathologyimages

Dataset

Handling class imbalance:

- Oversampling from minority class
- Calculate class weight to penalize the model more if error occurs in the minority class
- Data augmentations like rotation, horizontal flip, width shift and height shift

Class 0: Non cancer
Class 1: Cancer

Methodology

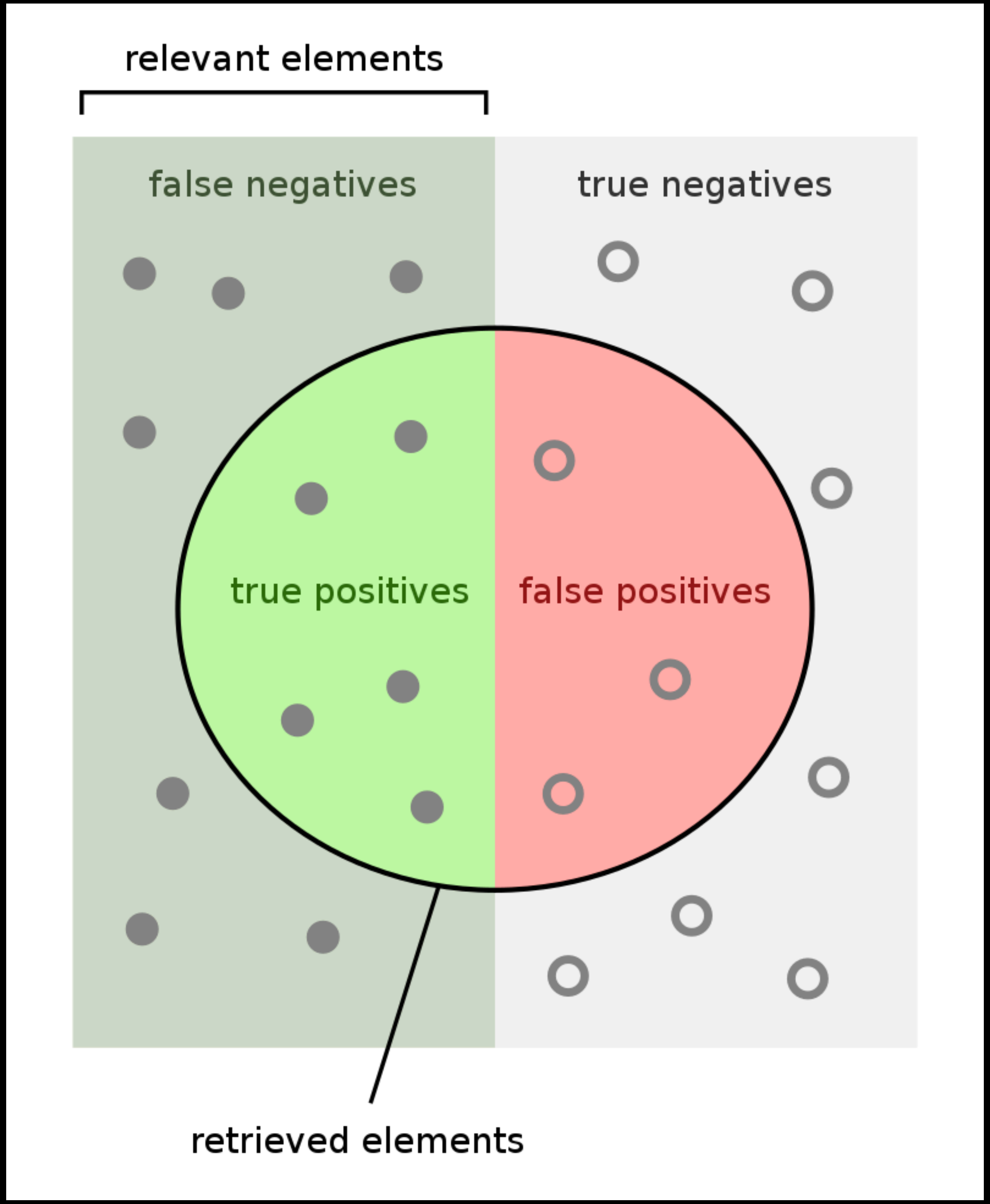
- Built Naive Model (baseline)- Randomly assigned class label based on class probability distribution in training data
- Aim to build model to beat the baseline model
- Created *three* more models :
  - Model 1 : CNN with Regularization
  - Model 2 : CNN with Residual Network
  - Model 3 : CNN with Transfer Learning

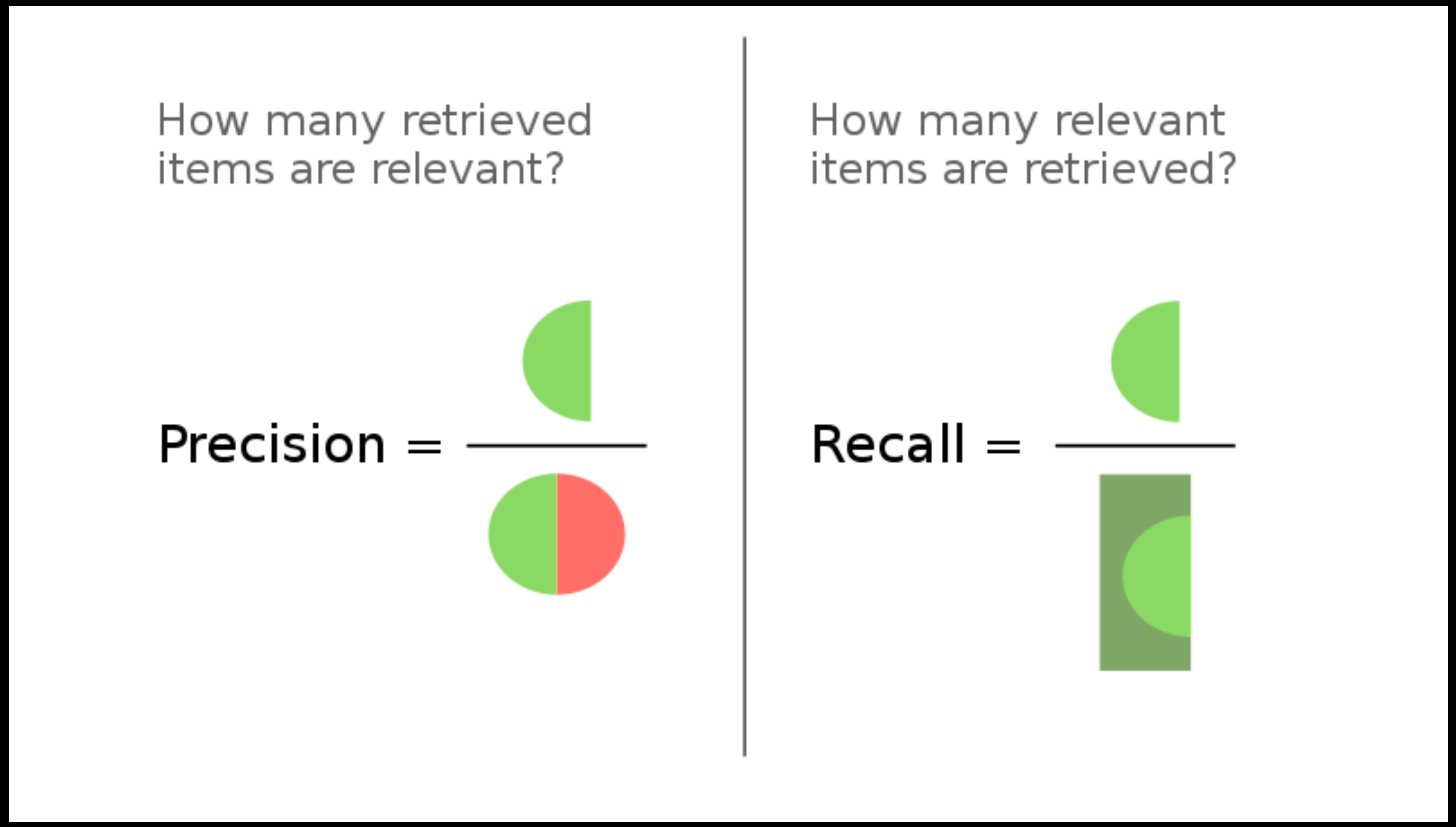
Methodology

- Used Cosine learning rate decay - Gradually decreases learning rate, helps model converge faster to global minimum, and prevents oscillations
- We used most commonly Adam optimizer
- Loss function used: Binary Cross Entropy
- Metrics used - Accuracy, Precision, Recall and F1 score
- Since this is a binary classification problem, we are using 1 neuron in the last Fully connected(FC) layer and applying sigmoid activation to get probabilities value for cancer

Methodology

Evaluation metrics



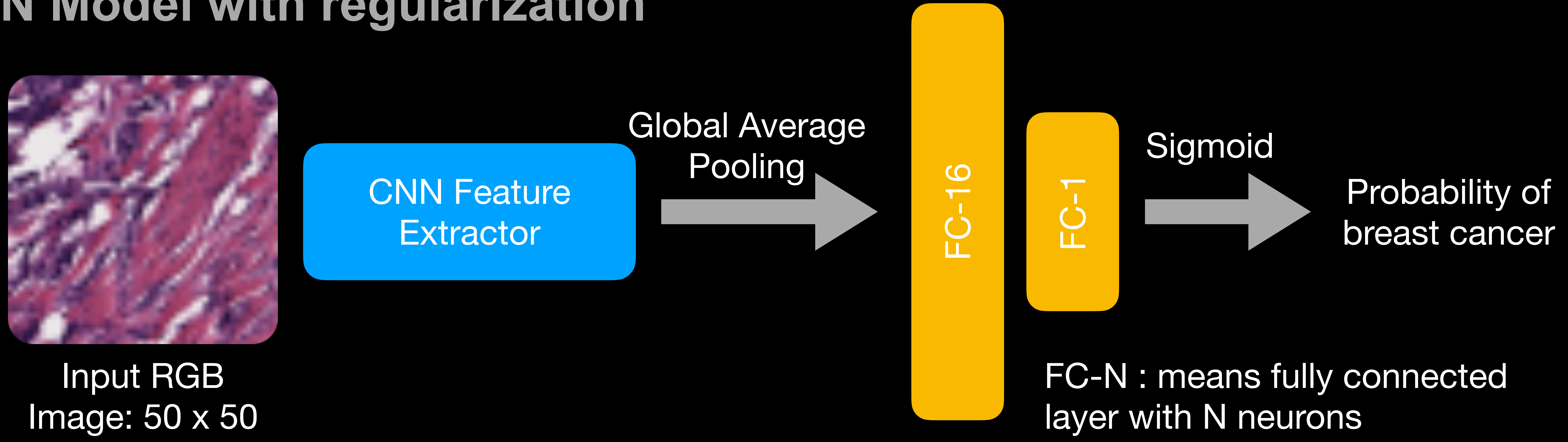


$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Image source: Wikipedia

Model 1

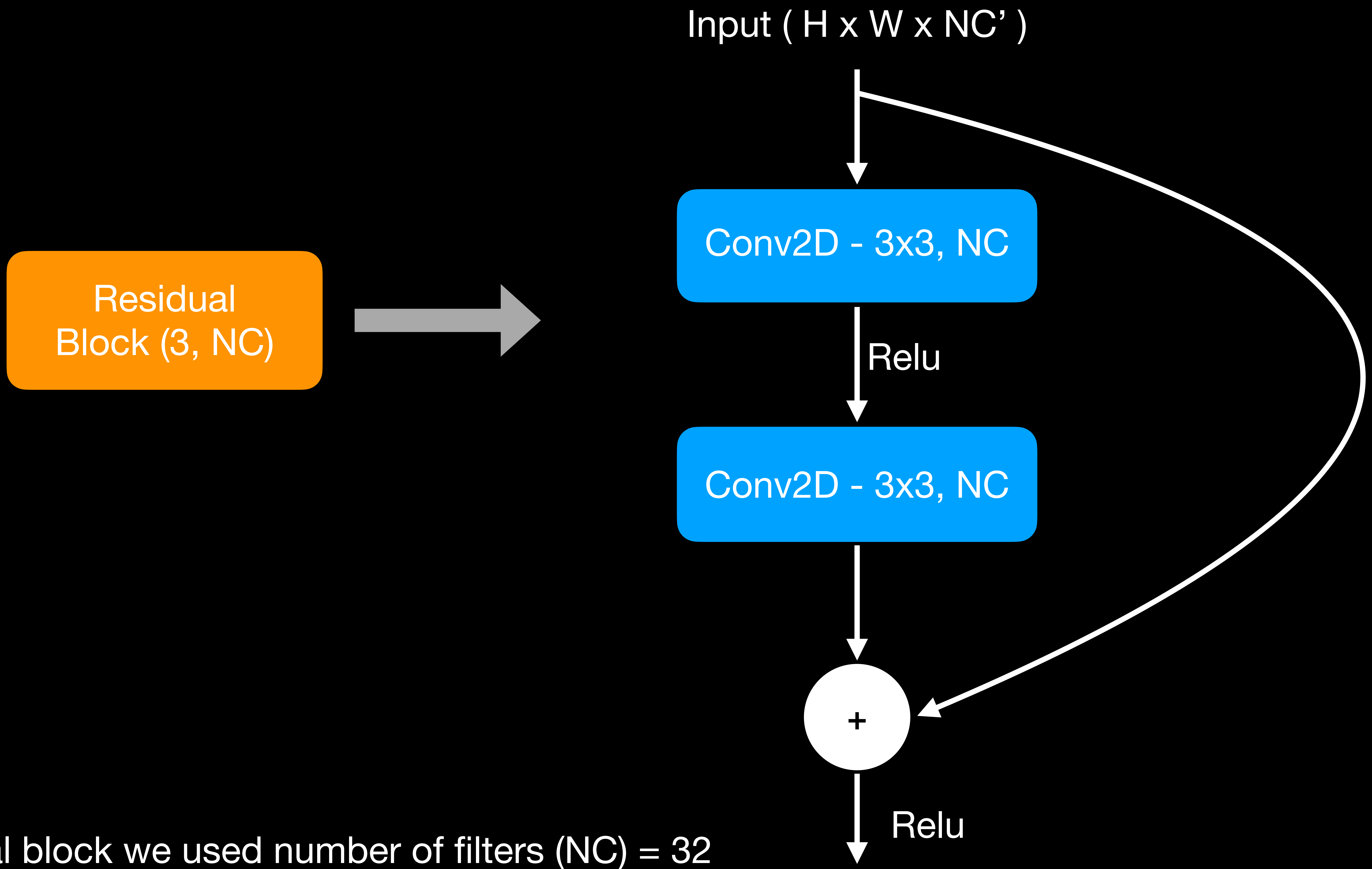
CNN Model with regularization



- Our CNN Feature Extractor consists of ***four*** Conv2D layers with 32 filters each and kernel size of 3 x 3
- BatchNormalization to facilitate faster convergence, Dropout for regularization
- Include ReLU activation, and GlobalAveragePooling2D for feature aggregation
- Model is trained with L2 Regularization

Model 2

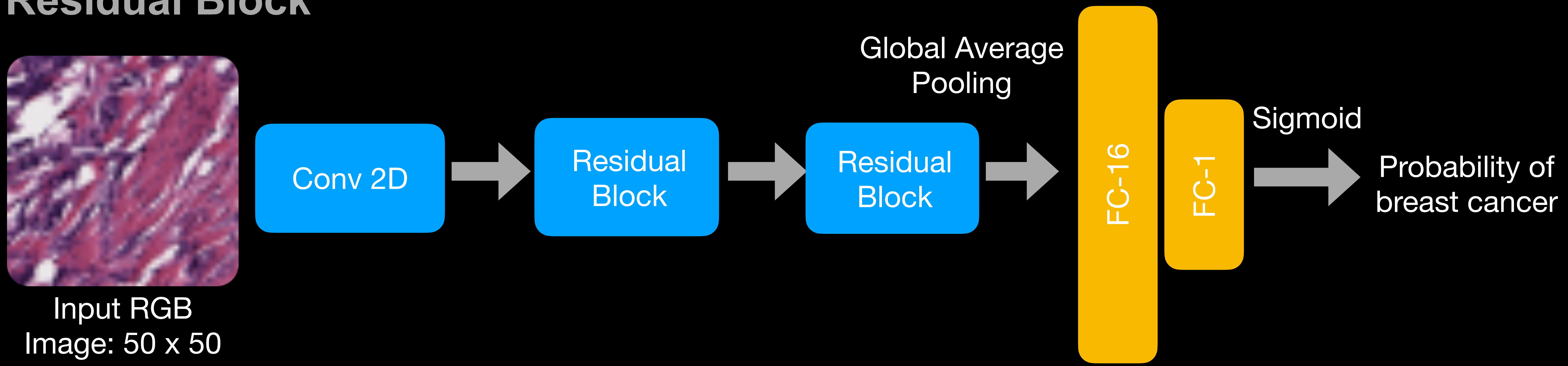
Residual Block



For each Residual block we used number of filters (NC) = 32

Model 2

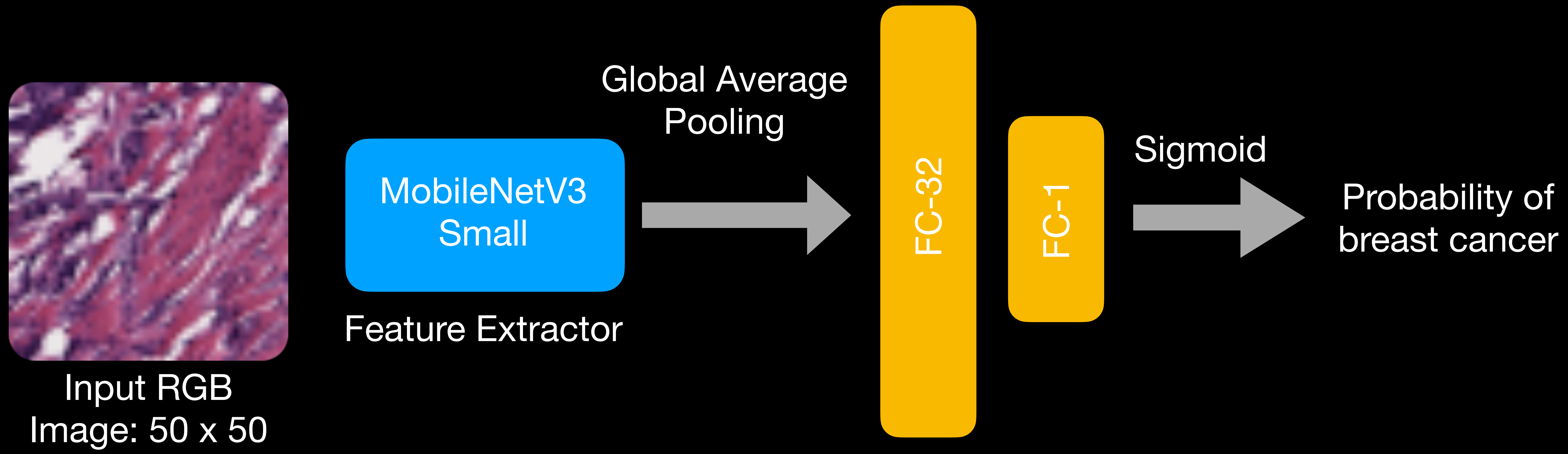
Residual Block



Conv2D layer uses kernel size of 3 x 3 followed by Maxpooling2D

Model 3

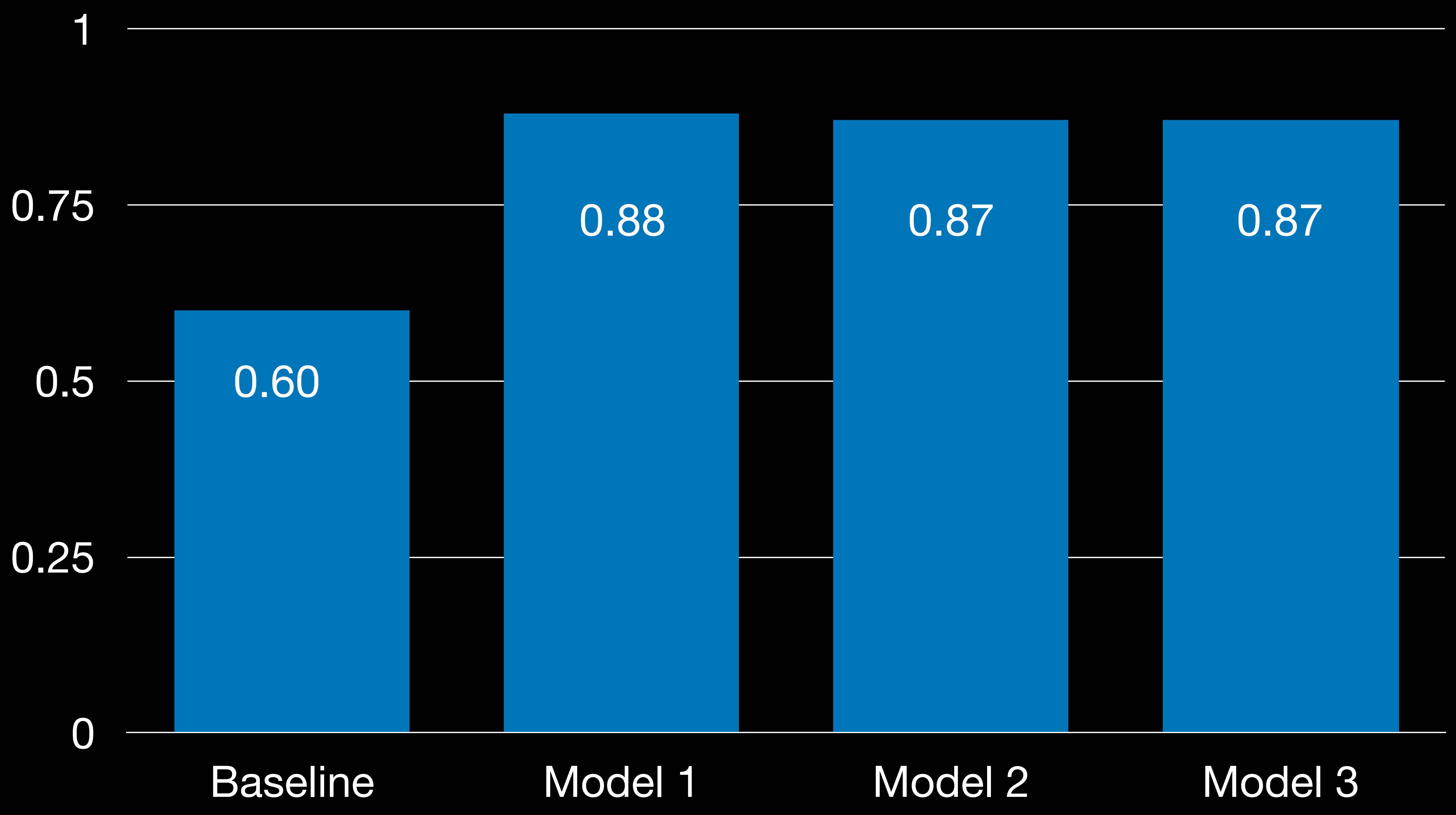
CNN with transfer learning



- ***MobileNetV3 Small*** model is very lightweight and mobile friendly and has total around 2M training parameters.
- We use pre trained weights from imagenet and train MobileNet backbone along with FC layers

Results

Accuracy

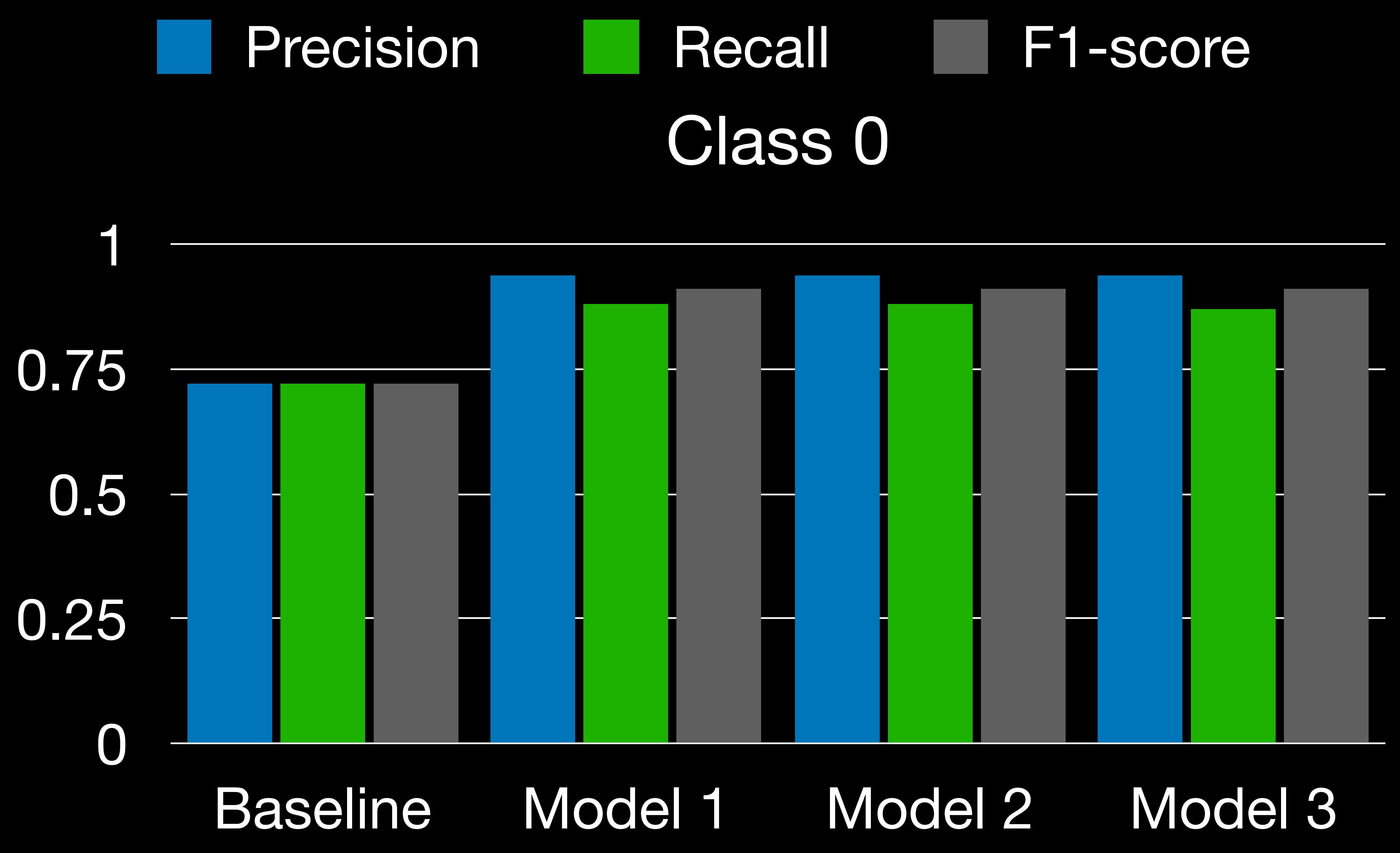


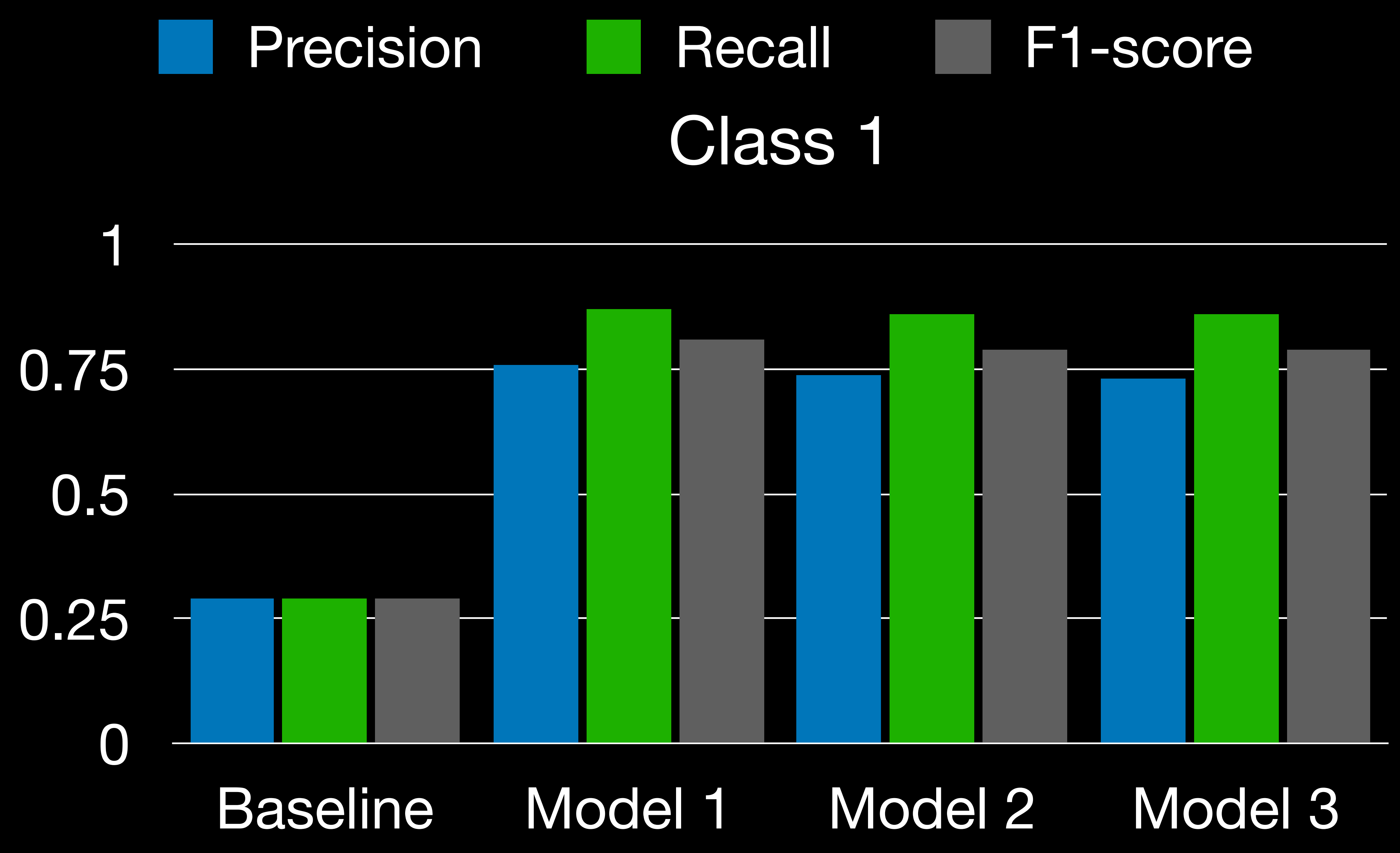
Baseline : Naive model
Model 1 : CNN Regularized Model
Model 2 : CNN with Residual Block
Model 3 : CNN with transfer learning

We observe Model 1 achieves the highest accuracy

Results

Precision/Recall/F1-Score @ 0.5 threshold



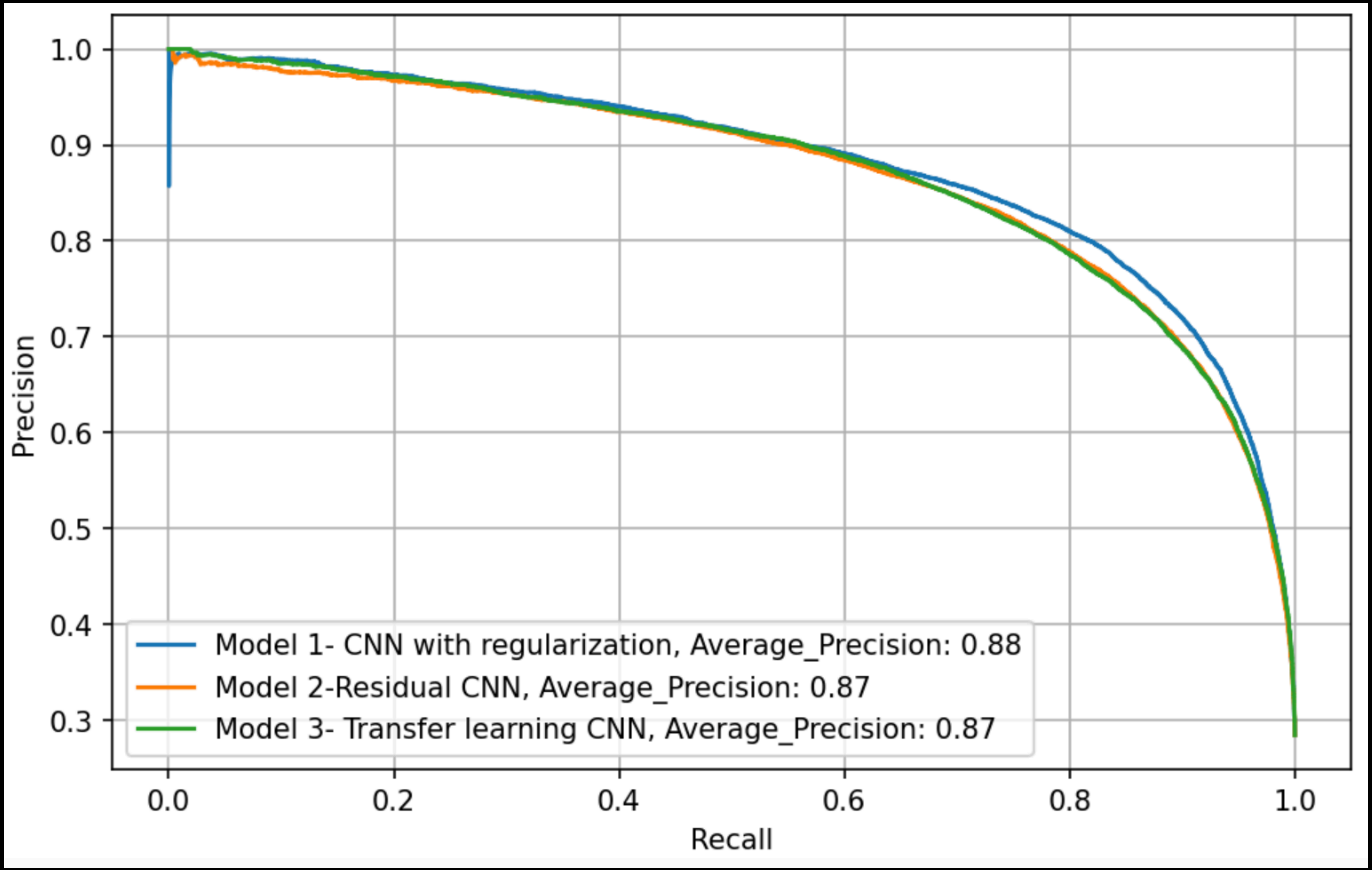


Baseline : Naive model
Model 1 : CNN Regularized Model
Model 2 : CNN with Residual Block
Model 3 : CNN with transfer learning

For Class 1, Model 1 has highest precision (0.76), recall (0.87) and F1- score(0.81)

Results

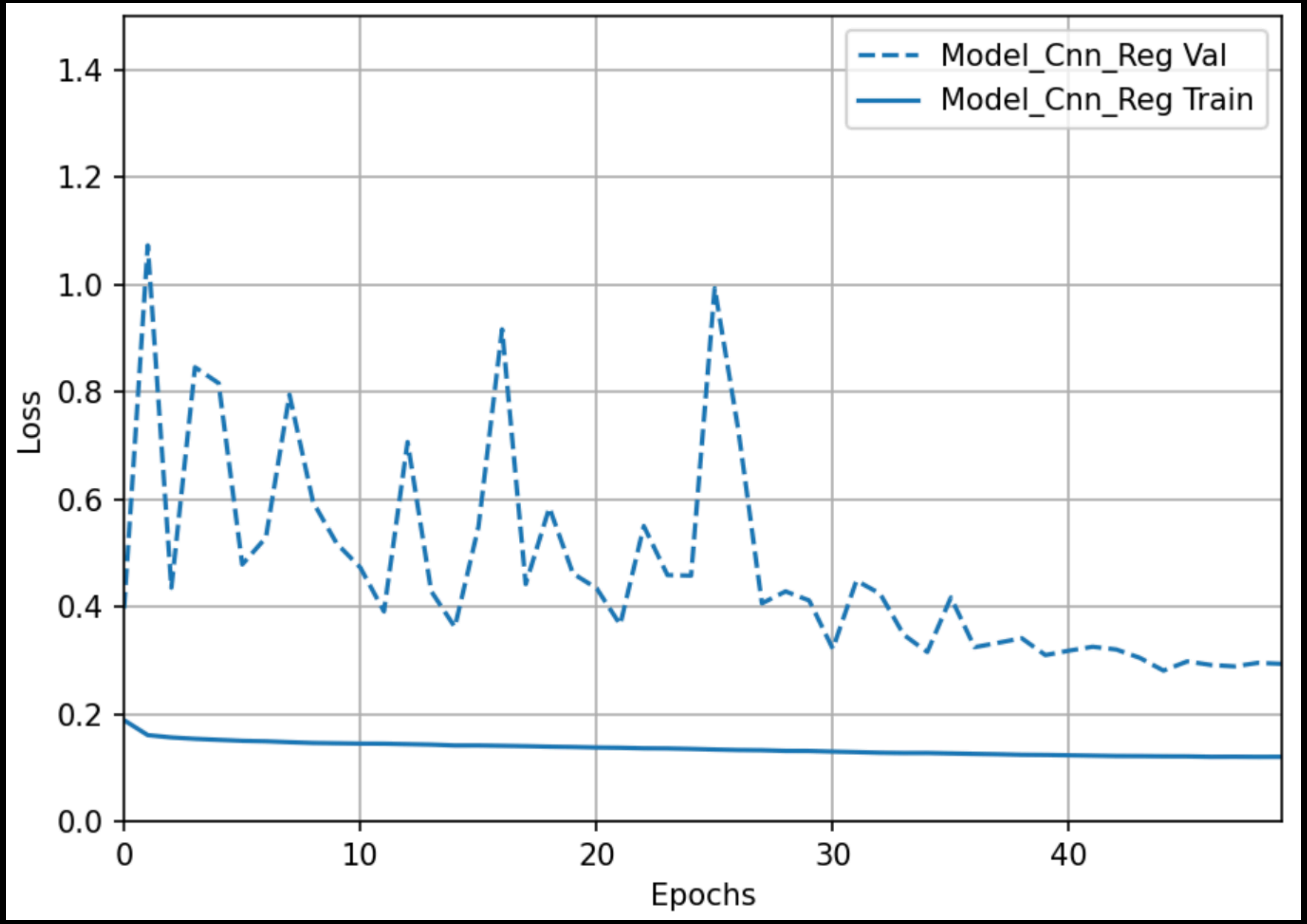
Precision Recall Curve

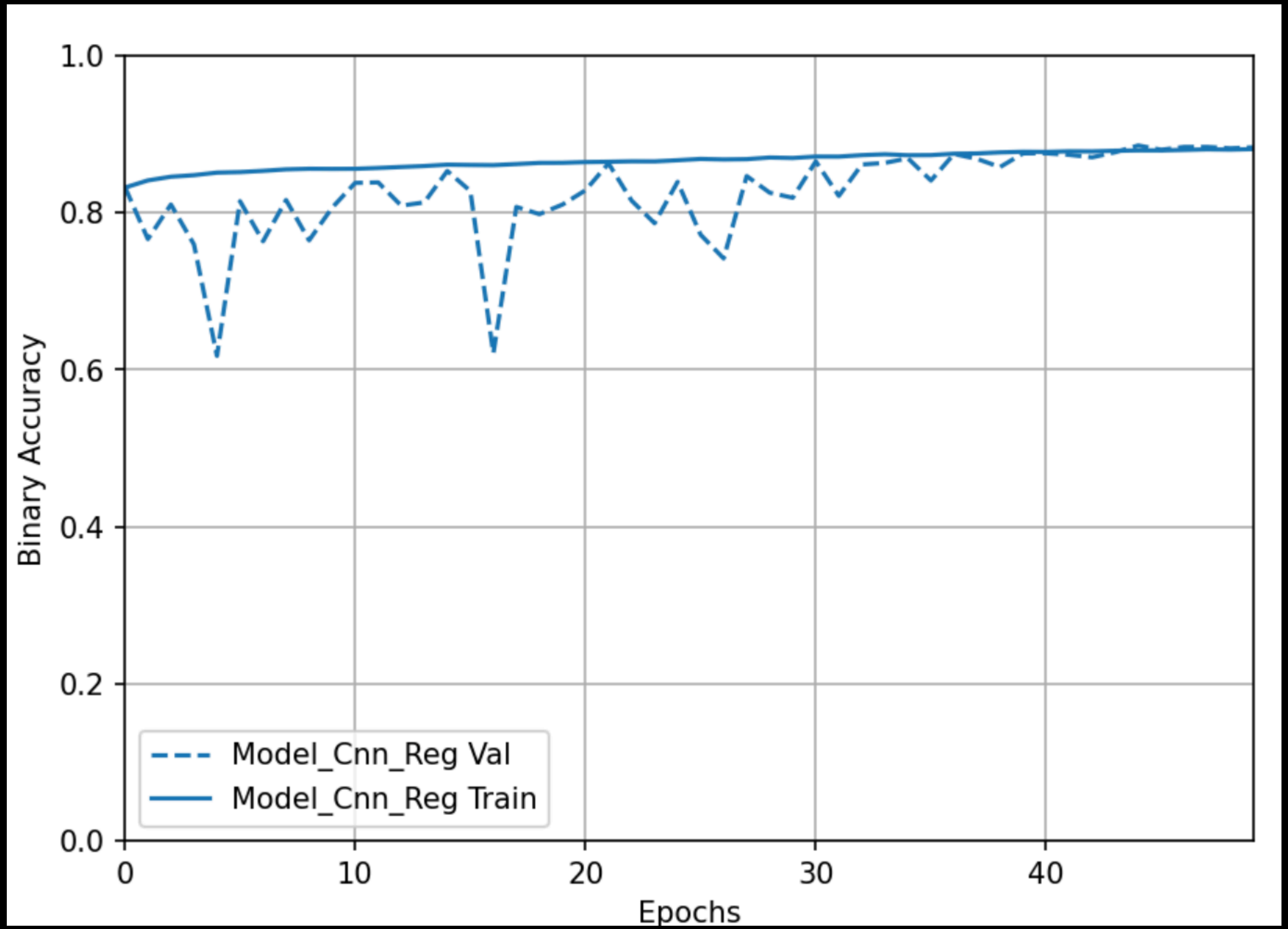


Model 1 has highest average precision

Training curve of best model (Model 1)

Loss and Accuracy





Best Model - CNN Regularized Model

Trained over 50 epochs, model tends to converge in both Loss and Binary Accuracy graph

Conclusion

- CNN Regularized (L2) Model is the best model achieving highest precision, recall and accuracy compared to other candidate models
- Based on PR curve, we observe at an operating threshold of 0.64 model has precision of 0.8 and recall of 0.8
- Finally we would like to deploy our best model at this operating point

Future Work

- We can further improve the performance of the model by using other feature extractor backbones like VGG, InceptionNet etc.
- Can implement multi-class classification to identify subtypes of breast cancer (e.g., lobular carcinoma, DCIS, IDC)
- Perform segmentation to locate cancerous regions in breast tissue

Thanks !

Q & A