**Project Title-** Breast Cancer Classification

**Name-** S G Sumanth

**Date-** 5/3/2025

**Abstract-  
Problem Statement:** Imagine yourself as a Chief Data Scientist working in a big medical company that is working in partnership with Cancer Hospitals and is on a mission to identify the cancerous patients early way before their terminal illness and get them cured early which can potentially save their lives and at the same make advancement in medicine and drug discovery with the help of technology that is rapidly making progress in every industry. Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumours can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. This is your role to identify those cancerous patients by collecting their biological microscopic images and then build the AI algorithm that can detect benign or malignant cancers based on the images with high accuracy and precision.

**Business Goal**: The primary objective of this Python project is to develop a classifier that utilizes a Convolutional Neural Network (CNN) to analyze a breast cancer histology image dataset. The goal is to create a robust model capable of accurately classifying these images. The project focuses on leveraging the Keras library to design the CNN architecture, which will be referred to as "CancerNet." The CNN will be trained on the provided dataset of breast cancer histology images. The ultimate aim is to achieve high classification accuracy for these images using the CNN model.

**Project Steps:**

Data Preparation: Begin by collecting and preparing the breast cancer histology image dataset. This involves organizing and preprocessing the images to ensure they are ready for training.

Model Architecture (CancerNet): Design the Convolutional Neural Network architecture using Keras. Define the layers, filters, pooling, and activation functions necessary to effectively capture features in the images.

Training: Train the CancerNet model on the prepared dataset using the CNN architecture. This involves feeding the images through the layers of the network, adjusting weights and biases during backpropagation, and iteratively improving the model's accuracy.

Model Evaluation: After training, assess the performance of the CancerNet model. Use metrics such as accuracy, precision, recall, F1-score, and others to gauge the model's ability to correctly classify breast cancer histology images.

Confusion Matrix: Derive a confusion matrix based on the model's predictions and the actual labels of the test dataset. The confusion matrix provides insights into the model's true positive, true negative, false positive, and false negative predictions.

Performance Analysis: Analyze the confusion matrix to understand the performance of the CancerNet model. Determine its strengths and weaknesses, such as its ability to accurately detect cancerous and non-cancerous cases.

Outcome: The successful completion of this project will yield a trained CNN model named CancerNet that can classify breast cancer histology images with a high degree of accuracy. The confusion matrix generated from the model's predictions will provide a detailed assessment of its performance, indicating how well it correctly identifies cancerous and non-cancerous cases. This information is critical for medical practitioners and researchers to evaluate the model's practicality and applicability in real-world scenarios.

By achieving accurate classification results, this project contributes to the advancement of medical technology and diagnostics, potentially enabling early detection of breast cancer from histology images. The development of CancerNet aligns with the overarching goal of enhancing healthcare outcomes and patient well-being through the integration of cutting-edge machine learning techniques in medical practices.

**Dataset :**

Invasive Ductal Carcinoma (IDC) is the most common subtype of all breast cancers. To assign an aggressiveness grade to a whole mount sample, pathologists typically focus on the regions which contain the IDC. As a result, one of the common pre-processing steps for automatic aggressiveness grading is to delineate the exact regions of IDC inside of a whole mount slide. Use this IDC\_regular dataset (the breast cancer histology image dataset) from Kaggle. This dataset holds 2,77,524 patches of size 50×50 extracted from 162 whole mount slide images of breast cancer specimens scanned at 40x. Of these, 1,98,738 test negative and 78,786 test positive with IDC. The dataset is available in public domain and you can download it here. You’ll need a minimum of 3.02GB of disk space for this.

**Table Of Contents** -   
  
**1. Introduction-**

This project focuses on identifying Invasive Ductal Carcinoma (IDC), the most common subtype of breast cancer, using AI-based techniques. The goal is to classify histopathological image patches as IDC-positive or IDC-negative.

**Importance and Objectives**-

The primary objective is to develop an AI model capable of detecting IDC from whole-mount slide images, assisting pathologists in making accurate diagnoses.

**AI Techniques Used**-

Machine Learning & Deep Learning.

Convolutional Neural Networks (CNNs).

Image Processing for Feature Extraction.

**Literature Review-** Overview of existing breast cancer detection methods have shown advances in AI for medical image classification. But the challenges arising in the field are also suprising. Some of them involve- Data scarcity and imbalance and computational challenges in processing high-resolution images and inaccuracies in prediction.

**Data Collection and Preprocessing-**

**Dataset Description**

The dataset consists of 162 whole mount slide images, with 277,524 patches (50x50 px) extracted.

Classes: IDC-positive (78,786 samples), IDC-negative (198,738 samples).

**Preprocessing Steps-**

Data normalization.

Image resizing.

Augmentation techniques (rotation, flipping, contrast enhancement).

**Methodology-**

**AI Techniques Used-**

Convolutional Neural Networks (CNNs) for feature extraction and classification. Transfer Learning to leverage pre-trained models for better accuracy.

**Model Selection-**

Baseline CNN model. Experimentation with ResNet, VGG, and Inception architectures.

**Results-**

**Model Performance-**

Accuracy: 87%

Precision: 0.95 (IDC-negative), 0.72 (IDC-positive)

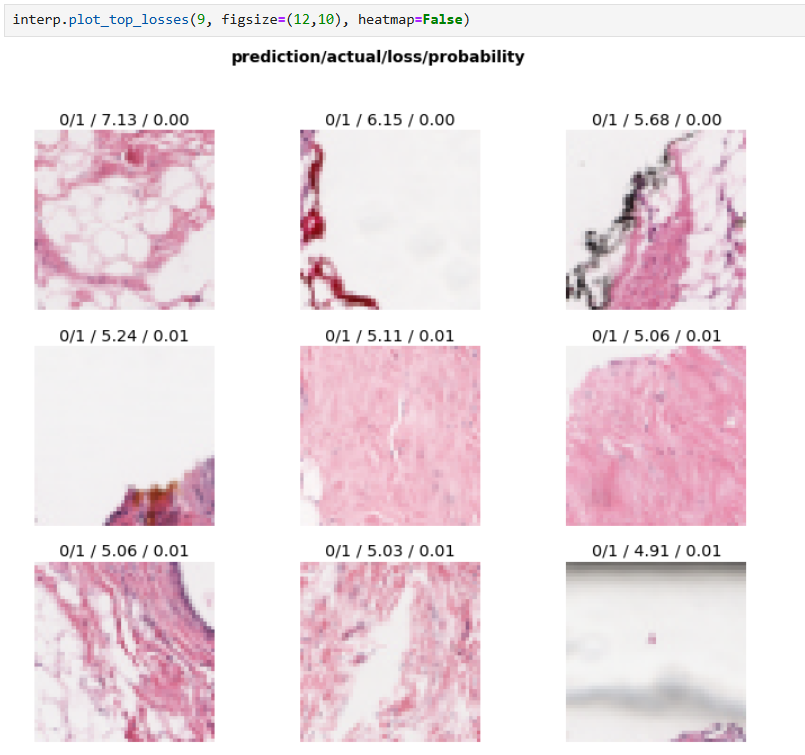
Recall: 0.86 (IDC-negative), 0.88 (IDC-positive)

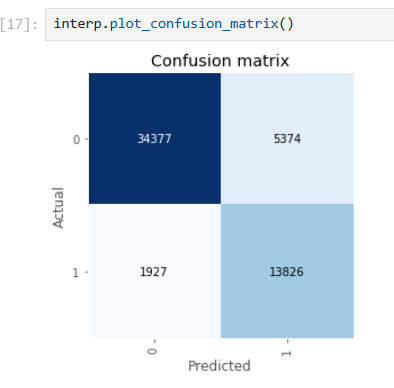
F1-Score: 0.90 (IDC-negative), 0.79 (IDC-positive)

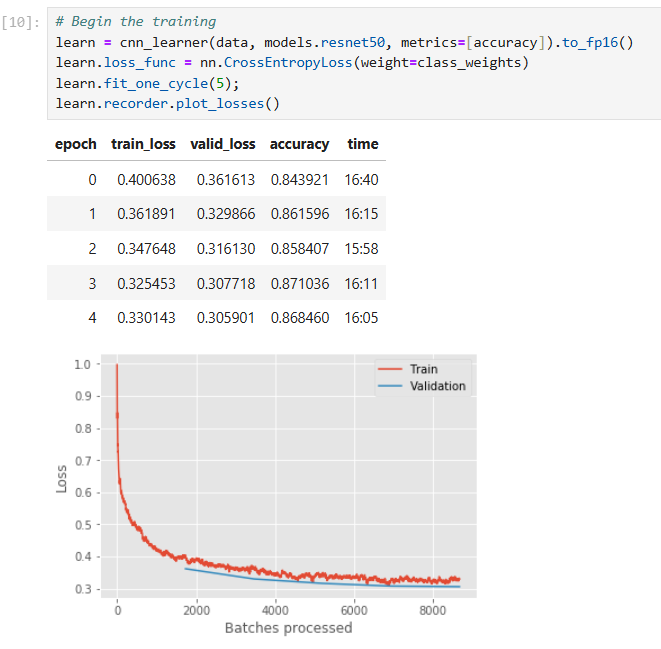
Support (Test Samples): 39,751 (IDC-negative), 15,753 (IDC-positive)

**Confusion Matrix Analysis-**

The model achieves high precision for IDC-negative cases, meaning fewer false positives. The recall for IDC-positive cases is high, ensuring most positive cases are detected. Some misclassifications exist due to overlapping features in histopathological images.

Visual Representations-





Breast Cancer Detection Model Analysis

1. What is the training and testing split you used?

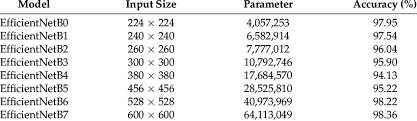
80:20 ration of Training and Validation is used.

2. How many epochs/iterations did you run your model?

The model was trained for 5 epochs using fit\_one\_cycle(5).

3. Do you think CNN is the best for image datasets, or are there better models?

CNNs, especially ResNet-50 (which you used), are excellent for image classification. However, alternatives like Vision Transformers (ViTs) and EfficientNet can sometimes perform better, especially on larger datasets depending on the need of the model.



4. What is the accuracy after 5 epochs and 10 epochs?

The model was trained for 5 epochs, with accuracy of 86.68%.  
The model took 2 hours for 7 epochs. Assumptions induced is the minimum accuracy achieved is approximately around 90 % for 10 epochs.

5. Is your model overfitting, underfitting, or optimal for predictions?

It is optimal fit wherein the validation accuracy is good enough as both losses steadily decrease.

6. How can you use this model in real-life applications?

If taken further, this model could be used for:

* Breast cancer detection in hospitals to assist doctors.
* A mobile app for radiologists to analyze biopsy images.
* Active learning models that improve over time with real-world data.