Analysis of Breast Cancer Classification Methodologies using Transfer Learning (VGG16)

Deep Learning Task Analysis

1 Introduction

Breast cancer remains one of the leading causes of mortality worldwide. Early and accurate detection is critical for improving survival rates. This report analyzes two distinct deep learning approaches implemented in the provided task, both leveraging the VGG16 architecture through transfer learning to classify malignant and benign cases.

2 Task Description

The primary objective of the task is to perform binary classification (Malignant vs. Benign). The notebook explores two different data sources:

- **Approach 1:** Classification based on clinical numerical features (Tabular data).
- **Approach 2:** Classification based on histopathology image tiles (Image data).

3 Methodology Comparison

3.1 Approach 1: Numerical Feature Transformation

This code uses the `load_breast_cancer` dataset from Scikit-Learn. It contains 30 numerical features for 569 patients.

- **Data Preprocessing:** The features are scaled using `MinMaxScaler`. To make them compatible with a Convolutional Neural Network (CNN), the 30 features are padded to 1024 and reshaped into a $32 \times 32 \times 3$ "pseudo-image."
- **Model Architecture:** VGG16 base (frozen) followed by a Dense layer (64 units), Dropout (0.2), and a Sigmoid output.

3.2 Approach 2: Histopathology Image Classification

This code utilizes the `Breast Histopathology Images` dataset, consisting of actual tissue image patches.

- **Data Preprocessing:** 20,000 images are sampled. Images are resized to 50×50 pixels and normalized. `ImageDataGenerator` is used for structured loading.
- **Model Architecture:** VGG16 base (frozen) followed by a Dense layer (256 units), Dropout (0.5), and a Sigmoid output. The optimizer is Adam with a custom learning rate of 0.0001.

Table 1: Comparison of Implementation Details

Feature	Approach 1 (Numerical)	Approach 2 (Images)
Data Source	Sklearn Breast Cancer Dataset	Kaggle Histopathology Images
Input Nature	Tabular (transformed to image)	Raw PNG Image Patches
Input Size	$32 \times 32 \times 3$	$50 \times 50 \times 3$
Sample Size	569 records	20,000 images
Dense Layer	64 units	256 units
Dropout Rate	0.2	0.5
Final Accuracy	**97%**	**82.7%**

4 Comparative Analysis

5 Conclusion

The results show that Approach 1 achieved higher accuracy (97%). However, it is important to note that Approach 1 deals with a much smaller and ”cleaner” set of pre-calculated structural features. Approach 2 deals with raw medical imagery, which is significantly more complex and reflective of real-world clinical diagnostic tasks. While Approach 2 has lower accuracy, it demonstrates the ability of VGG16 to extract meaningful patterns from biological tissue textures without manual feature engineering.