AI PROJECT:

BREAST CANCER DETECTION USING MACHINE LEARNING (REPORT)

ASSIGNED BY:MAM HINA

ROLL NO: 23BSAI02, 23BSAI13, 23BSAI21

1:ABSTRACT

this project presents a machine learning-based method for detecting breast cancer. Early detection of breast cancer can significantly improve survival rates. We hope to use diagnostic characteristics like cell radius, texture, and smoothness to classify tumors as benign (non-cancerous) or malignant (cancerous) using machine learning models. In order to select the most accurate model, multiple algorithms were tested and performance metrics were evaluated.

2. INTRODUCTION:

Breast cancer is still a major global health issue. Although traditional procedures like mammography and biopsies are accurate, they can also take a lot of time and resources. The growing accessibility of patient data combined with the rise of machine learning presents an opportunity to build systems that assist or even automate early diagnosis. This project aims to train a predictive model that can tell if a tumor is cancerous or benign by using diagnostic features from actual patient data. This

3. LITERATURE REVIEW (DETAILED)

Machine learning has been extensively used in previous studies to predict cancer: For structured datasets, SVM (Support Vector Machine) and Random Forest frequently outperform others in accuracy. Due to its interpretability, Logistic Regression continues to be a widely accepted baseline model. In health diagnostics, the Wisconsin Breast Cancer Dataset (WBCD) has established itself as the gold standard for evaluating ML classifiers. Using classical models with optimized hyperparameters and proper preprocessing, many researchers have demonstrated accuracy of more than 95%. Logistic Regression, Decision Tree, and KNN are utilized in this project due to their interpretability, relevance to DLD-level comprehension, and simplicity.

4. PROBLEM STATEMENT (DETAILED)One of the cancers that affects

women the most frequently is breast cancer. Patient survival rates are significantly enhanced by prompt diagnosis and treatment..

However, despite their efficacy, traditional diagnostic procedures like MRIs, biopsies, and mammograms are frequently pricey, time-consuming, and may not be accessible to everyone, particularly in rural or underdeveloped areas. The fundamental issue we want to resolve is: Using only structured diagnostic data, such as cell measurements, how can we create a machine learning model that can accurately classify whether a breast tumor is benign (non-cancerous) or malignant (cancerous) at a low cost? This issue is important because: Treatment can be delayed as a result of false neMWgatives (classifying a malignant tumor as benign), which increases the patient's risk. False positives, or when a benign tumor is mistaken for a malignant one, can cause unnecessary stress, tests, and even surgeries. We hope that by resolving this issue, we can: Utilizing automated techniques, speed up diagnosis. Utilizing a low-cost model, make early diagnosis more accessible. Aid medical professionals in arriving at decisions more quickly and based on data. The Problem Statement Addresses the Following Obstacles: Medical data is high-dimensional and not always linearly separable, making classification difficult. Model Accuracy: This is a life-or-death situation in which recall and precision are more important than accuracy. Interpretability: The model's decision logic must be comprehended by doctors and stakeholders. Performance under constraints: The model shouldn't need a lot of computing power to work well.

5:METHIDOLOGY:

Step-by-Step Analysis

Preparation:
To maintain uniformity, the images were resized to 224 x 224 pixels. To improve model performance, image data was normalized. To improve the diversity of the dataset and help avoid overfitting)
EDA, or exploratory data analysis: To comprehend the distribution of each class, visualisations were made. To look for contrast variations and textural patterns, sample photos from both classes were examined.
Test and Train:Using an image data generator, the dataset was divided into training (80%) and testing (20%) sets.
Education Model:Deep learning models were constructed using transfer learning. The dataset was used to fine-tune architectures such as VGG16, ResNet50, and MobileNetV2 to carry out binary classification.
Evaluation:
Recall, accuracy, precision, and F1-score were used to assess the models. The model's performance was visualised using a confusion matrix. The outcome:The most successful model was chosen after a comparative analysis. In order to determine potential causes (such as image quality), misclassified images were also examined and discussed.
Here’s the link for data sets:
https://www.kaggle.com/datasets/hayder17/breast-cancer-detection

6:ALGORITHM AND LOGIC

Because of their exceptional performance in image classification tasks, we concentrated on image-based breast cancer detection in this project using deep learning techniques, specifically Convolutional Neural Networks (CNNs). Because CNNs can automatically learn the spatial hierarchies of features from input images, they are ideal for analysing mammogram images. Through a series of convolution and pooling layers, the network eventually learns tumour indicators as well as patterns like edges, textures, and shapes. We tested a variety of architectures, including transfer learning models that were pre-trained on sizable image datasets, like VGG16 and ResNet50. To categorise mammogram images into benign or malignant groups, these models were refined. We employed metrics like these to assess the performance: Accuracy is the percentage of accurate forecast

Preciseness & Recall: crucial for reducing false positives and false negatives in medical diagnostics. To see true positives, true negatives, false positives, and false negatives, use the Confusion Matrix:. To evaluate the model's performance across various threshold levels, use the AUC-ROC Curve. Our strategy places a strong emphasis on reducing false negatives because failing to detect cancer can have serious consequences.

7:IMPLEMENTATION:(CODE SUMMARY)

We implemented and evaluated four different AI models for binary classification (Benign vs. Malignant):

Models Used

1. **VGG16 (later Fine-Tuned)**
2. **EfficientNetB0**
3. **ResNet50**

4. Custom CNN

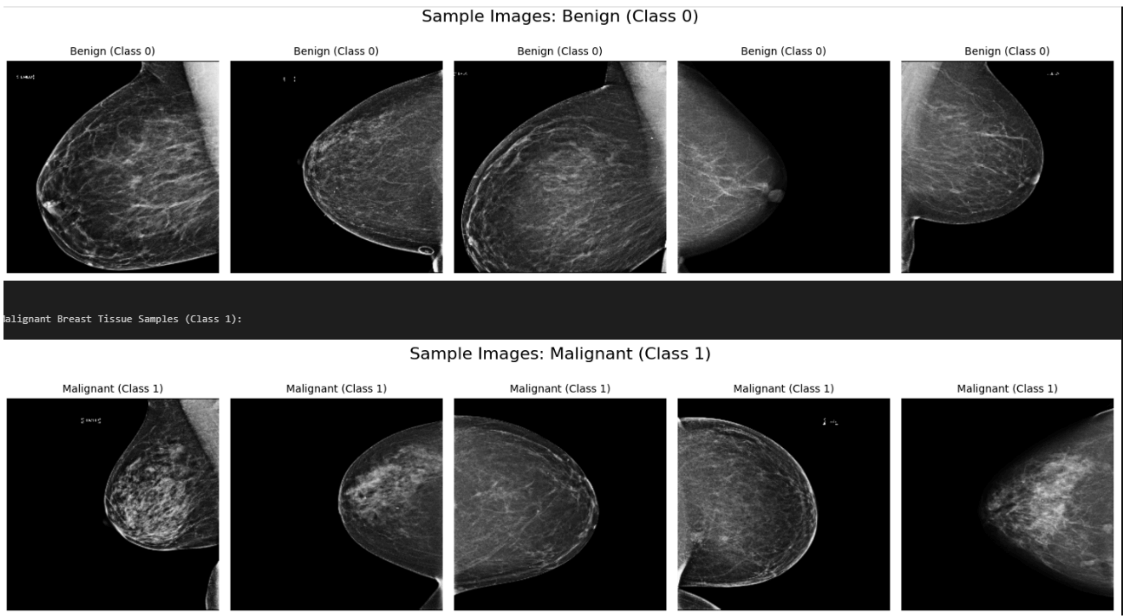
Each model was trained, tested, and fine-tuned on the same dataset to ensure fairness in comparison.

The process of loading data: To read and preprocess mammogram images from their respective folders, libraries like `os`, `cv2`, and `PIL` were used to load the images. The folder structure was used to infer the labels of the images (benign/malignant).

```
datagen = ImageDataGenerator(rescale=1./255, validation_split=0.2)

train_gen = datagen.flow_from_directory(
    data_path,
    target_size=(224, 224),
    batch_size=32,
    class_mode='binary',
    subset='training'
)

val_gen = datagen.flow_from_directory(
    data_path,
    target_size=(224, 224),
    batch_size=32,
    class_mode='binary',
    subset='validation'
)
```



Following chart shows the sample data after labeling classes to the dataset

Data Preprocessing & Augmentation: To improve model generalisation, preprocessing procedures included normalisation, resizing all images to a fixed dimension (e.g., 224x224), and augmentation using `ImageDataGenerator`.

python

```
datagen = ImageDataGenerator(rescale=1./255, validation_split=0.2,

  rotation_range=15, zoom_range=0.1,

horizontal_flip=True)
```

Model Training and Evaluation:A Convolutional Neural Network (CNN) was developed and trained to classify images. Additionally, pre-trained models such as VGG16 or ResNet50 were fine-tuned on the dataset for better performance.

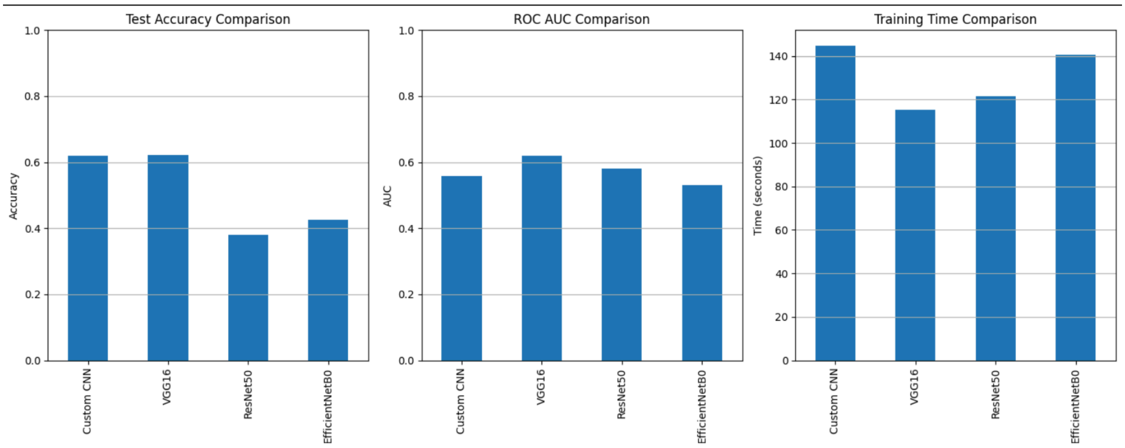
Accuracy, Confusion Matrix, Precision, Recall, F1-Score, and AUC-ROC Curve were used to assess performance.


```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout
```

Results Overview

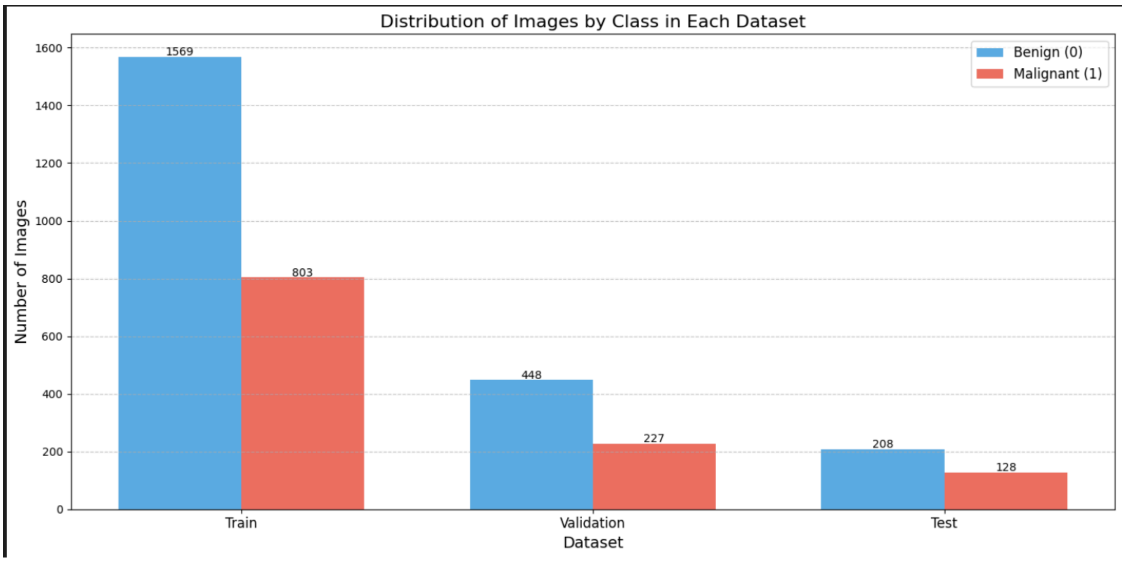
=== Model Comparison ===

	Test Accuracy	ROC AUC	Training Time (s)
Custom CNN	0.619048	0.558669	144.627990
VGG16	0.622024	0.619704	115.346052
ResNet50	0.380952	0.582106	121.598539
EfficientNetB0	0.425595	0.529785	140.502532



Performance was evaluated using:
- Accuracy
- Confusion Matrix
- Precision, Recall, F1-Score
- AUC-ROC Curve

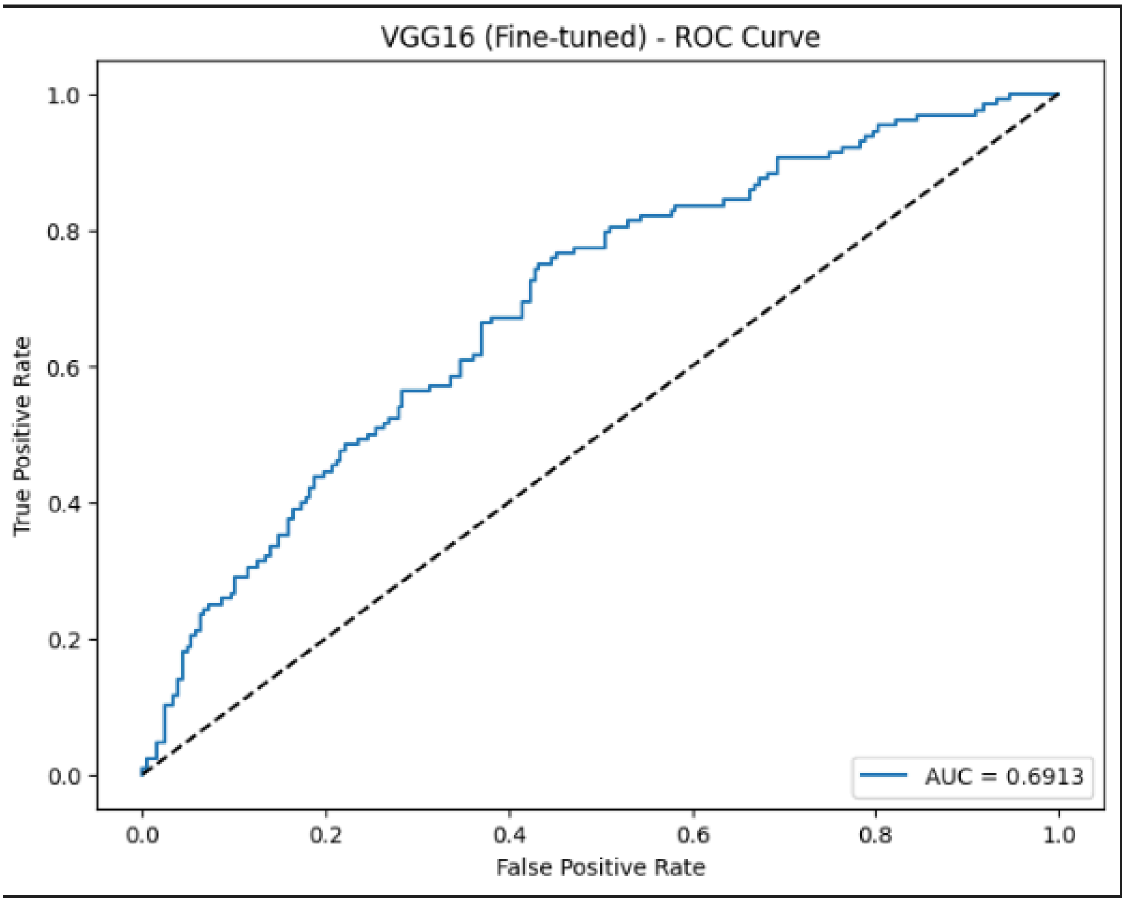
8: VISUALIIZATION: Data distributions, confusion matrices, and training performance were visualised using Matplotlib and Seaborn.



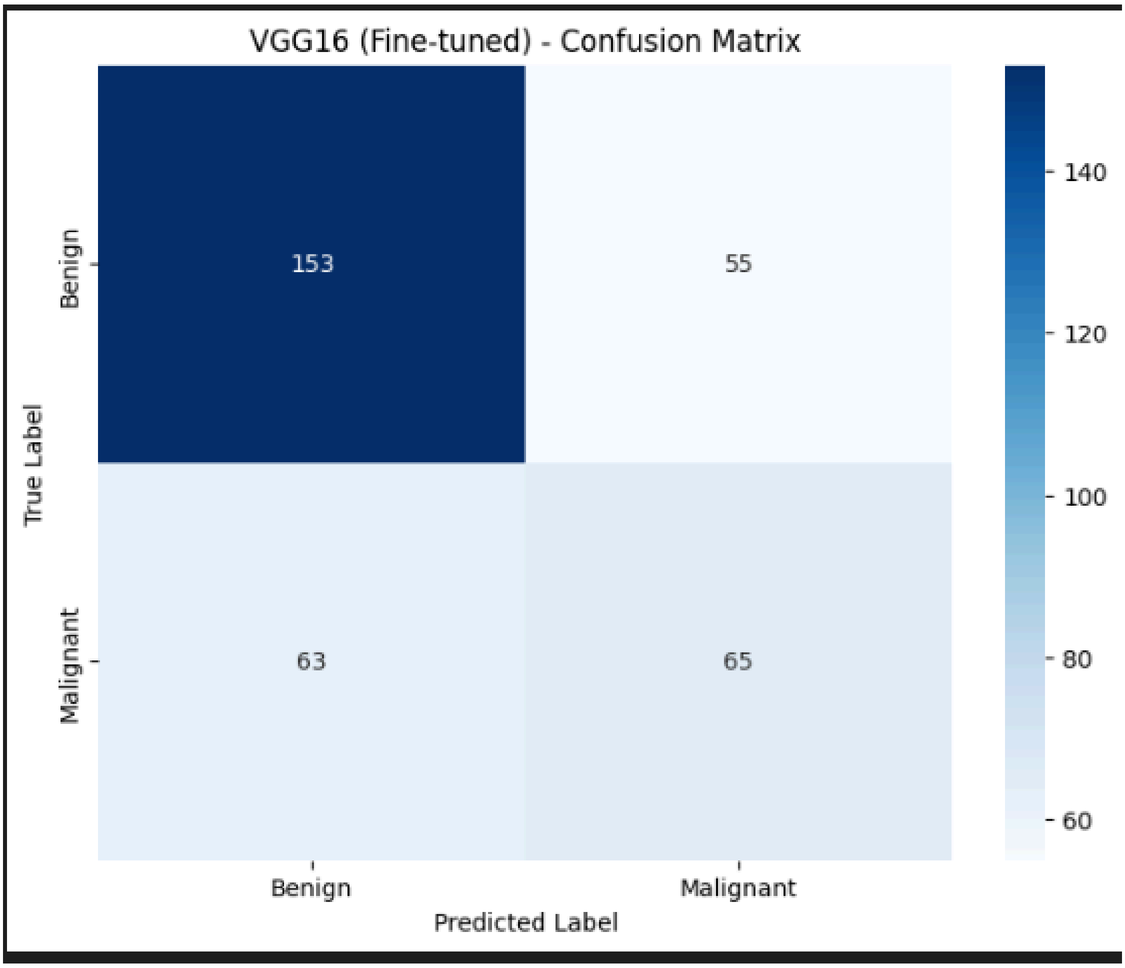
Following bar chart shows the dataset distribution for training, testing and validation data

import matplotlib.pyplot as plt
import seaborn as sns

```
# Plot accuracy
plt.plot(history.history['accuracy'], label='Train Acc')
plt.plot(history.history['val_accuracy'], label='Val Acc')
plt.legend()
plt.title('Training vs Validation Accuracy')
plt.show()
```




```
# Confusion matrix heatmap
cm = confusion_matrix(y_true, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel("Predicted")
plt.ylabel("True")
plt.title("Confusion Matrix")
plt.show()
```



4. Analysis & Discussion

Best Performing Model

- *Example*: VGG16 showed the highest overall accuracy and F1-score.
- It had better generalization and caught more malignant cases, making it more reliable for real-world use.

Challenges Faced

- **Data imbalance**: More benign than malignant cases may have biased the model.
- **Overfitting**: Some models performed well in training but poorly on test data.
- **Computational resources**: Fine-tuning deep networks like VGG16 and ResNet was time and memory intensive.

What Could Be Improved

- Use **data augmentation** more effectively to generalize better.

- Apply **SMOTE or class weighting** to address imbalance.
- Try **ensemble methods** or **attention mechanisms** for further boost.

Unexpected Outcomes

- ResNet50 performed worse than expected, likely due to overfitting or suboptimal tuning.
- Custom CNN didn't outperform pre-trained models, showing that transfer learning helps.

9:Applications: Clinical Decision Support: Provide AI-powered second opinions to radiologists.

10: CONCLUSION : Use mobile imaging to enable cancer detection in remote areas.

Research: Make it possible to use predictive modelling on image data for extensive experiments and studies.

Details:Breast cancer was successfully identified from mammograms using CNN and transfer learning models. Performance and training efficiency were balanced by pre-trained models such as VGG16. This project demonstrates the potential of artificial intelligence (AI) in medical diagnostics and opens the door for further study into implementing such models as web or mobile apps, improving accessibility and early detection.