Project Report

Mammography-Report Based Breast Cancer Detection

Submitted for

COGNITIVE COMPUTING (UCS420)

Submitted by:

Abhinav Dhir (102317123)

Devansh Kashyap (102317122)

Submitted to: Mr. Sukhpal Singh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY,
(DEEMED TO BE UNIVERSITY),
PATIALA, PUNJAB
INDIA

Session-Year (e.g. Jan-May, 2025)

Table of Contents:

Mammography Report-Based Breast Cancer Detection:

1. Abstract:

Breast cancer is a major global health concern, where early and accurate detection significantly improves treatment outcomes. This project develops an AI-driven breast cancer detection system using mammography reports, combining machine learning and deep learning techniques. Random Forest and XGBoost models are employed for structured clinical data, while deep learning techniques enhance image-based diagnosis. The integration of these methods enables a more robust and comprehensive classification approach. By leveraging both textual and imaging data, this system aims to support AI-assisted diagnostics, aiding medical professionals in early detection and reducing diagnostic errors.

2. Introduction:

Breast cancer is one of the most common cancers affecting women worldwide, and early detection significantly improves survival rates. This project aims to develop a machine learning model to classify breast cancer findings using both textual and image data from mammography reports. By analyzing structured data extracted from radiology reports, our goal is to assist radiologists in decision-making and enhance diagnostic efficiency.

3. Dataset Overview:

The dataset used for this project is CBIS-DDSM (Curated Breast Imaging Subset of DDSM), which contains digitized film mammography images annotated with information crucial for breast cancer detection. The dataset includes ROI (Region of Interest) annotations, lesion segmentation masks, and pathology-confirmed labels for abnormalities such as calcifications and masses. The project directory consists of multiple CSV annotation files and a folder containing the mammogram images in DICOM format converted to jpg. The dataset is structured as follows:

jpeg: Contains mammography images in DICOM format converted to jpg, categorized based on cases of calcifications and masses and patient specific cases.

calc_case_description_train.csv & calc_case_description_test.csv: Metadata for calcification cases, containing:study_id: The encoded study identifier.

- • `patient_id`: Unique identifier for the patient.
- • `image_path`: Path to the corresponding DICOM image.
- • `breast_density`: Density category of the breast (1-4).
- • `abnormality_type`: Type of abnormality (Calcification).
- • `abnormality_id`: Unique identifier for the abnormality.
- • `assessment`: BI-RADS assessment of the abnormality.
- • `subtlety`: Visibility of the abnormality (1-5).

- • `pathology`: Confirmed diagnosis (Benign / Malignant).

mass_case_description_train.csv & mass_case_description_test.csv: Metadata for mass cases, with attributes:

- • `patient_id`: Unique identifier for the patient.
- • `image_path`: Path to the corresponding DICOM image.
- • `breast_density`: Density category of the breast.
- • `abnormality_type`: Type of abnormality (Mass).
- • `abnormality_id`: Unique identifier for the abnormality.
- • `assessment`: BI-RADS assessment of the abnormality.
- • `subtlety`: Visibility of the abnormality.
- • `pathology`: Confirmed diagnosis (Benign / Malignant).
-
- ROI (Region of Interest) Mask Files: These files provide segmentation masks highlighting the regions containing masses or calcifications.

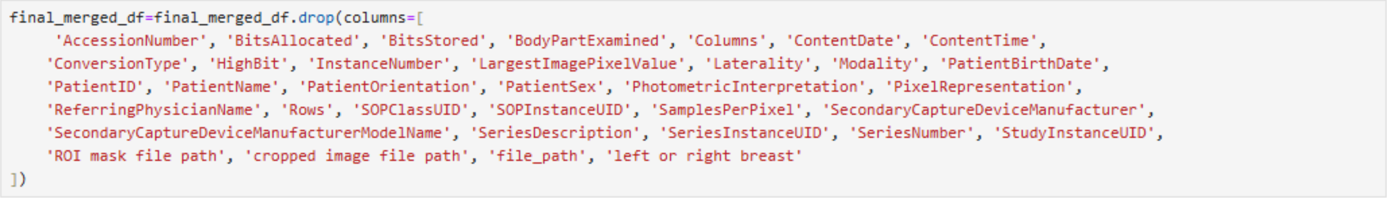
File Directory Structure:

```
-   ├──CBIS-DDSM
-      ├── csv
-        ├── calc_case_description_train.csv
-        ├── calc_case_description_test.csv
-        ├── mass_case_description_train.csv
-        ├── mass_case_description_test.csv
-        ├── dicom_info.csv
-        ├── meta.csv
-      ├── jpeg
-          ├── 1.3.6.1.4.1.9590.100.1.2.100018879311824535125115145152454291132
-                ├── Image_1.jpg
-                ├── Image_2.jpg
-          ├── 1.3.6.1.4.1.9590.100.1.2.100131208110604806117271735422083351547
-                ├── Image_1.jpg
-                ├── Image_2.jpg
-          ├── 1.3.6.1.4.1.9590.100.1.2.100522099512256189513864912954167862869
-                ├── Image_1.jpg
-
-   .
-   .
-   .
-   .
-   .
-         ├──1.3.6.1.4.1.9590.100.1.2.999339456125048761365874498818889466699
-                ├── Image_1.jpg
-                ├── Image_2.jpg
```

4. Data Preprocessing:

a. Feature Selection and Cleaning

The dataset contained numerous attributes, many of which were redundant or irrelevant for model training. The following columns were dropped to streamline the dataset and retain only the most useful features:



```
final_merged_df=final_merged_df.drop(columns=[
     'AccessionNumber', 'BitsAllocated', 'BitsStored', 'BodyPartExamined', 'Columns', 'ContentDate', 'ContentTime',
    'ConversionType', 'HighBit', 'InstanceNumber', 'LargestImagePixelValue', 'Laterality', 'Modality', 'PatientBirthDate',
    'PatientID', 'PatientName', 'PatientOrientation', 'PatientSex', 'PhotometricInterpretation', 'PixelRepresentation',
    'ReferringPhysicianName', 'Rows', 'SOPClassUID', 'SOPInstanceUID', 'SamplesPerPixel', 'SecondaryCaptureDeviceManufacturer',
    'SecondaryCaptureDeviceManufacturerModelName', 'SeriesDescription', 'SeriesInstanceUID', 'SeriesNumber', 'StudyInstanceUID',
    'ROI mask file path', 'cropped image file path', 'file_path', 'left or right breast'
])
```

b. Handling Missing Values

Missing values in the dataset were imputed using the mode of the respective column. This approach ensures that categorical features remain consistent while minimizing data loss.

c. Encoding Categorical Variables

Categorical attributes such as BI-RADS assessment, breast density, and abnormality types were converted into numerical representations using label encoding and one-hot encoding where applicable. This transformation ensures compatibility with machine learning models.

5. Machine Learning Models

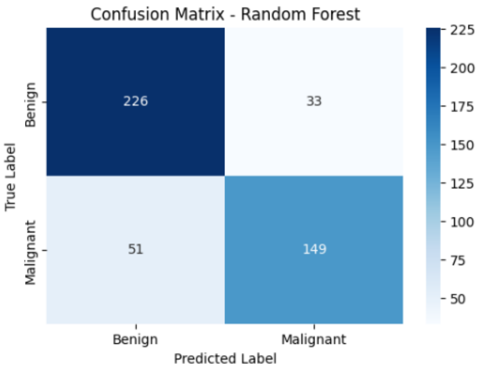
a. Random Forest Classifier

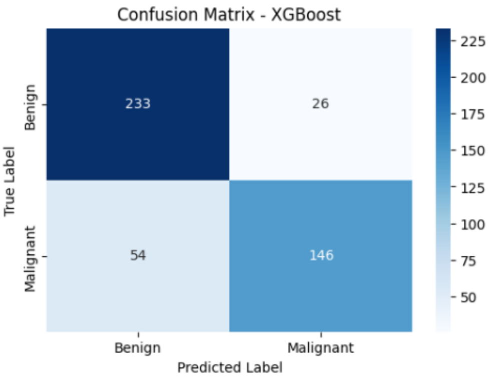
Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their outputs for improved accuracy. It is particularly effective for structured data classification tasks.

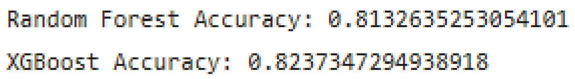
b. XGBoost Classifier (Extreme Gradient Boosting)

XGBoost is an optimized gradient boosting algorithm that improves accuracy and computational efficiency. It is widely used in predictive modeling due to its ability to handle missing values and extract feature importance

By using a combination of accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC, we ensure a thorough evaluation of our breast cancer detection models. Among these, recall is the most critical metric since it directly impacts the early detection of breast cancer and reduces the risk of missing malignant cases.







```
Random Forest Accuracy: 0.8132635253054101
XGBoost Accuracy: 0.8237347294938918
```

c. Hyperparameter Tuning

Hyperparameter tuning was performed using Random Search to optimize the parameters of both Random Forest and XGBoost models. The best-performing configurations are being evaluated, and the final accuracy metrics will be updated accordingly.

6. Deep Learning Models

a. EfficientNet for Image Classification

EfficientNet is a state-of-the-art convolutional neural network (CNN) optimized for high accuracy with reduced computational cost. It uses compound scaling to balance network depth, width, and resolution for optimal performance.

b. Data Preprocessing for Deep Learning

1. Image resizing to 224×224 pixels for EfficientNet input requirements.
2. Pixel normalization to scale intensity values between 0 and 1.
3. Data augmentation using horizontal flipping, rotation, and contrast adjustment to enhance model robustness.

c. Model Implementation and Training

The EfficientNet model was implemented using TensorFlow . The architecture includes:

- Pretrained EfficientNet-B0 backbone (trained on ImageNet).
- Global Average Pooling (GAP) layer to extract high-level image features.
- Dense layers with ReLU activation for feature learning.
- Dropout layers to reduce overfitting.
- Sigmoid activation for binary classification (malignant vs. benign tumors).

7. Conclusion and Future Work

The combination of machine learning models (Random Forest, XGBoost) and deep learning (EfficientNet) provides a comprehensive approach for breast cancer detection. While machine learning models offer interpretability and efficiency in handling structured data, deep learning models excel at feature extraction from mammographic images.

Future updates may include incorporating advanced natural language processing (NLP) techniques, such as BioBERT or ClinicalBERT, to improve the extraction and analysis of relevant textual features from radiology reports.

Further improvements can be made by optimizing hyperparameters, increasing dataset size, and incorporating advanced techniques such as Grad-CAM for explainability in deep learning predictions.

8. Expected Impact

This project enhances early breast cancer detection by leveraging machine learning (Random Forest, XGBoost) and deep learning (EfficientNet with TensorFlow)**.** The AI-driven system improves diagnostic accuracy (~82%)**,** reduces false negatives**,** and assists radiologists with faster, standardized assessments.

- Assist radiologists in decision-making by automating the classification process.
- Reduce diagnostic errors by using machine learning to analyze structured mammography report data.
- Improve early breast cancer detection through efficient data-driven predictions.

9. Conclusion

This project demonstrates the potential of machine learning (Random Forest, XGBoost) and deep learning (EfficientNet with TensorFlow) in breast cancer detection using the CBIS-DDSM dataset. By leveraging feature engineering, hyperparameter tuning, and deep feature extraction, we achieved promising classification accuracy (~82%), aiding in early detection and diagnosis.

The integration of Grad-CAM further enhances explainability, providing visual insights into the model's decision-making process, which is crucial for clinical adoption. The system can serve as a decision-support tool for radiologists, improving diagnostic accuracy, reducing false negatives, and accelerating early intervention.

Moving forward, this AI-driven approach can be optimized with larger datasets, transfer learning, and federated learning to enhance robustness, scalability, and real-world deployment. Ultimately, this project contributes to the advancement of AI-powered healthcare, making breast cancer screening more accurate, efficient, and accessible.