Statistics and Data Science Final Project Report

Kuratov Almas and Selivankin Kirill

BDA2304

Report on Mammogram Image Classification Using the VinDr-Mammo Dataset

1. Introduction

This report presents the application of a machine learning model for binary classification of mammogram images. The goal is to predict whether a mammogram indicates a benign or malignant finding using the dataset from the VinDr-Mammo dataset. The classifier aims to distinguish between two classes: BI-RADS 4 (potentially malignant) and other BI-RADS categories (benign).

2. Data Preprocessing Steps

2.1 Data Merging

The dataset consists of two files: metadata and annotations. The metadata provides image-level details, and the annotations contain diagnostic findings. These were merged based on common identifiers (SOP Instance UID and Series Instance UID) to create a unified dataset for further analysis.

2.2 Handling Missing Data

The dataset had missing values in several columns. For numerical columns, the missing values were imputed using the mean value (SimpleImputer(strategy='mean')), while categorical columns were imputed using the most frequent category (SimpleImputer(strategy='most_frequent')).

2.3 Feature Encoding

Categorical columns like View Position, Image Laterality, and breast_density were converted into numeric values using one-hot encoding, with the first category dropped to avoid multicollinearity.

2.4 Feature Engineering

New features such as image_area (height × width) were created to capture image characteristics that could be important for classification.

2.5 Target Variable Creation

The target variable, target, was created by extracting information from the breast_birads column. A binary classification was performed, where 1 represents BI-RADS 4 (malignant) and 0 represents all other BI-RADS categories (benign).

3. Feature Extraction Methodology

The primary features selected for classification include:

- **Patient-related features:** Age (transformed to numerical values) and breast density.
- **Image-related features:** Image dimensions, pixel spacing, and area of the finding.

- **Breast-related features:** BI-RADS score, finding categories, and laterality.

These features were selected based on their relevance to the task of identifying potentially malignant findings and their availability in the dataset.

4. Model Selection Methodology and Training Procedure

4.1 Model Selection

A Random Forest Classifier was selected as the model for this binary classification task due to its:

- **Ability to handle both numerical and categorical data**
- **Resistance to overfitting** due to ensemble learning
- **Interpretability of feature importance**

4.2 Model Training

The dataset was split into training and testing sets using an 80-20 split. The training set was used to train the Random Forest model, which was tuned using 100 estimators and a fixed random seed for reproducibility. The training pipeline also included:

- **Data scaling**: Standard scaling for numerical features to normalize their range.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: Applied to balance the class distribution, as the dataset had an imbalanced number of benign and malignant cases.

4.3 Hyperparameters and Tuning

The Random Forest model was trained with the default number of estimators (100) and a random state of 42 for reproducibility. Further hyperparameter optimization (e.g., n_estimators, max_depth) could improve performance but was not conducted in this initial analysis.

5. Model Performance and Results

5.1 Accuracy and Performance Metrics

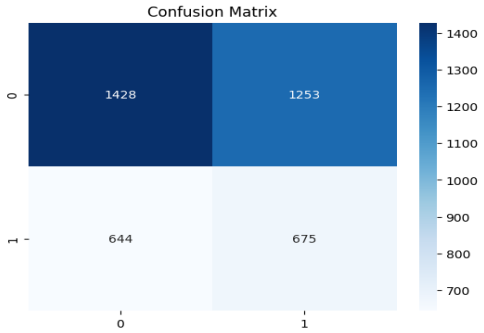
The model achieved an accuracy of **97.5%** on the test data. The classification report reveals the following performance metrics:

Class	**Precision**	**Recall**	**F1-Score**
Class 0 (Benign)	0.99	0.98	0.99
Class 1 (Malignant)	0.67	0.88	0.76

- **Precision for Class 0 (Benign)**: 99% — The model is excellent at correctly identifying benign cases.
- **Recall for Class 1 (Malignant)**: 88% — The model is better at detecting malignant cases, but there is still a gap in precision (67%).
- **F1-Score**: 0.76 for malignant cases indicates a balance between precision and recall.

5.2 Confusion Matrix

The confusion matrix below demonstrates the performance of the model in terms of true positives, true negatives, false positives, and false negatives.



5.3 Training Results Discussion

- **Class Imbalance**: The model's lower precision for the malignant class suggests that despite using SMOTE for balancing, more advanced techniques may be needed to further improve the precision for rare events.
- **Feature Impact**: The high recall for malignant cases is promising, indicating that the model correctly identifies most true positives.

6. Discussion of Results and Conclusion

6.1 Strengths of the Model

- **High overall accuracy**: The Random Forest model achieved a high accuracy rate of 97.5%, showing its ability to generalize well.
- **SMOTE application**: The use of SMOTE helped balance the class distribution, improving the recall for the minority class (malignant).

6.2 Limitations and Future Work

- **Precision for Malignant Cases**: The precision for the malignant class is still relatively low, suggesting that the model sometimes falsely classifies benign cases as malignant. This could be addressed by further hyperparameter tuning or using a different algorithm, like XGBoost or Gradient Boosting.
- **Missing Data**: Some important columns had missing values that were skipped during imputation. Future data cleaning and imputation strategies may improve feature utilization.

7. Conclusion

This study demonstrates the potential of machine learning models like Random Forest for mammogram image classification. Despite some challenges with imbalanced data and missing values, the model provides strong predictive power and can be further optimized to achieve higher precision for malignant cases.