Statistics and Data Science Final Project Report
Binary Classification of Mammogram Images Using VinDr-Mammo Dataset

Kuratov Almas/Selivankin Kirill BDA2304

1. **Introduction**
   This report presents the classification of mammogram images into normal (BI-RADS 1) and abnormal (BI-RADS 2, 3, 4, 5) categories. We used the VinDr-Mammo dataset, which includes both metadata and mammogram images, to train and evaluate various machine learning models for classification.

2. **Data Preprocessing**

- Extracted breast-level annotations and converted BI-RADS categories into a binary target variable.
- Extracted image features such as mean intensity and standard deviation using OpenCV to capture texture details from the images.
- Encoded categorical variables (breast density, laterality, view position) into numerical representations.
- Handled missing values by imputing with column means.
- Applied SMOTE (Synthetic Minority Over-sampling Technique) for class balancing.
- Standardized numerical features using StandardScaler for consistency across models.

3. **Feature Extraction Methodology**

- Extracted metadata features from the dataset, including image height, width, breast density, laterality, and view position.
- Computed image intensity statistics (mean and standard deviation) from the mammogram images to capture essential texture information.
- Incorporated pixel-level features and metadata for comprehensive analysis and improved model performance.

4. **Model Implementation and Evaluation**
   The following models were implemented and evaluated:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree Classifier
- Random Forest Classifier
- K-Nearest Neighbors (KNN)

5. **Performance Comparison**

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5370	0.3453	0.4511	0.3912	0.518
SVM	0.4880	0.3484	0.6353	0.4501	0.536
Decision Tree Classifier	0.5218	0.3366	0.4640	0.3902	0.503
Random Forest Classifier	0.5012	0.3472	0.5823	0.4350	0.543
K-Nearest Neighbors	0.5235	0.3277	0.4230	0.3693	0.490

6. **Model Selection and Discussion**
   Recall is the primary metric, as correctly identifying abnormal cases is crucial for medical diagnostics.

- **SVM** and **Random Forest** performed best in terms of recall (63.53% and 58.23%, respectively), demonstrating their stronger ability to detect abnormal cases.
- **Logistic Regression** and **Decision Trees** showed moderate performance, while **KNN** had the lowest effectiveness in recall.
- The overall performance of all models is suboptimal, indicating that further improvements are needed.

7. **Confusion Matrix Analysis**
   Each model's confusion matrix was plotted to visualize misclassifications. Most models struggled with correctly identifying abnormal cases, leading to lower recall scores. This is critical in medical diagnostics where false negatives could lead to undetected conditions.

8. **Conclusion & Future Work**
   The recall scores indicate that while the models can detect a reasonable number of abnormal cases, there is significant room for improvement in terms of clinical applicability.
   Potential enhancements:

- **Feature engineering** using advanced deep learning-based feature extraction methods.
- **Hyperparameter tuning** for SVM and Random Forest models to fine-tune their performance.
- Exploring **CNN-based models** for better image representation and classification.
  Future work should focus on improving the classification performance using more advanced techniques like Convolutional Neural Networks (CNNs) and additional data sources to refine the model for clinical use.

This report provides an overview of the methodology, results, and potential improvements for binary classification of mammogram images using both metadata and image features. Further refinements are necessary to achieve clinically acceptable performance.