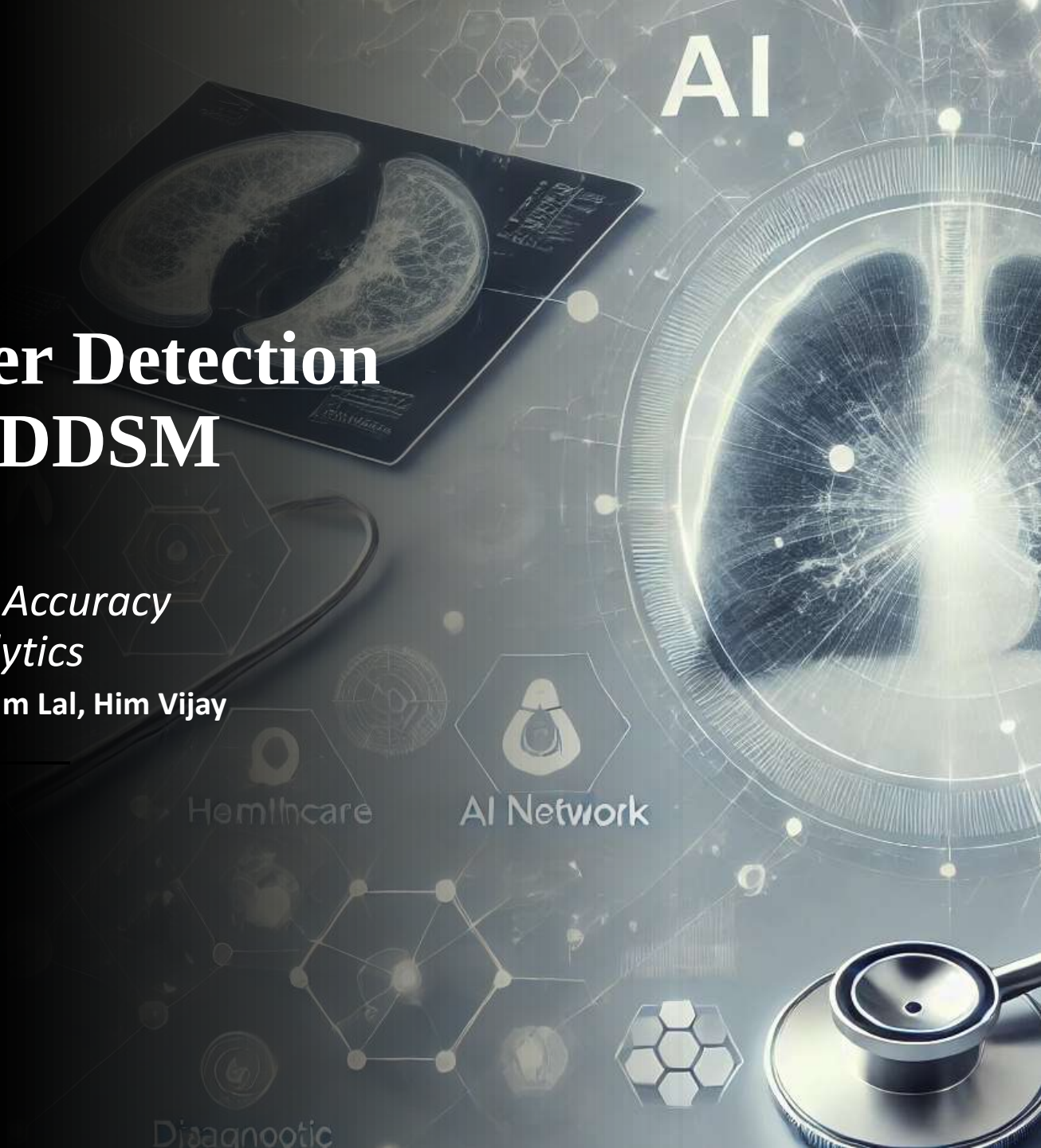
Breast Cancer Detection Using CBIS-DDSM Dataset

Enhancing Diagnostic Accuracy with AI and Data Analytics

Team 13: Priyanka Kapoor, Ram Lal, Him Vijay & Prerna Pandey

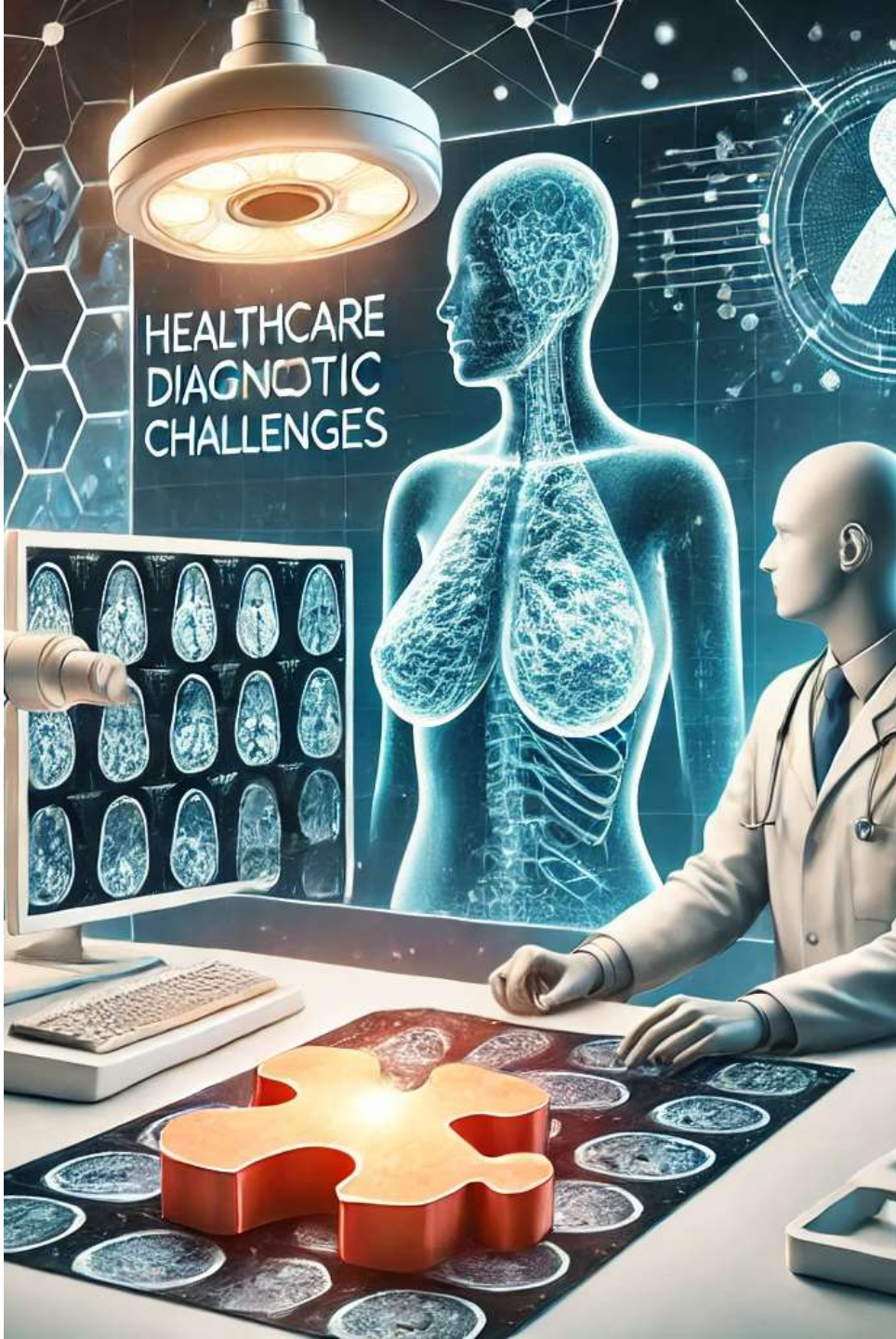


Introduction

Breast cancer is a leading cause of cancer-related deaths globally. Early detection is critical to reduce mortality rates, yet challenges like diagnostic variability, limited radiologists, and misdiagnosis persist. This project leverages the CBIS-DDSM dataset with AI models to improve diagnostic workflows, enhance accuracy, and provide scalable solutions.

Business Problem

- **High variability** in breast cancer diagnoses.
- **Limited access** to radiologists in remote areas.
- **Time-consuming and costly** traditional diagnostic methods.
- **Risk** of overdiagnosis and underdiagnosis.
- The project addresses these challenges using **AI-driven solutions.**



Objective

- Focuses on using **advanced machine learning models** and annotated mammographic data to automate, streamline, and enhance the accuracy of breast cancer detection.
- This approach not only addresses the outlined business problems but also **contributes to improving overall patient care and healthcare efficiency**.

Primary Goal

- To develop a data-driven system that integrates medical imaging and descriptive case data for improved breast cancer detection and diagnosis.
- To identify patterns and anomalies in mammograms and correlate these with patient and diagnostic metadata for enhanced decision-making.
- To assist radiologists in prioritizing critical cases and reducing the workload of manual interpretations.

Descriptive Analysis

Dataset Overview

Dataset source : CBIS-DDSM (Curated Breast Imaging Subset of DDSM)

The dataset is sourced from the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), available on Kaggle. It includes high-resolution mammographic images and metadata such as breast density, pathology, and diagnostic findings.

Key Features:

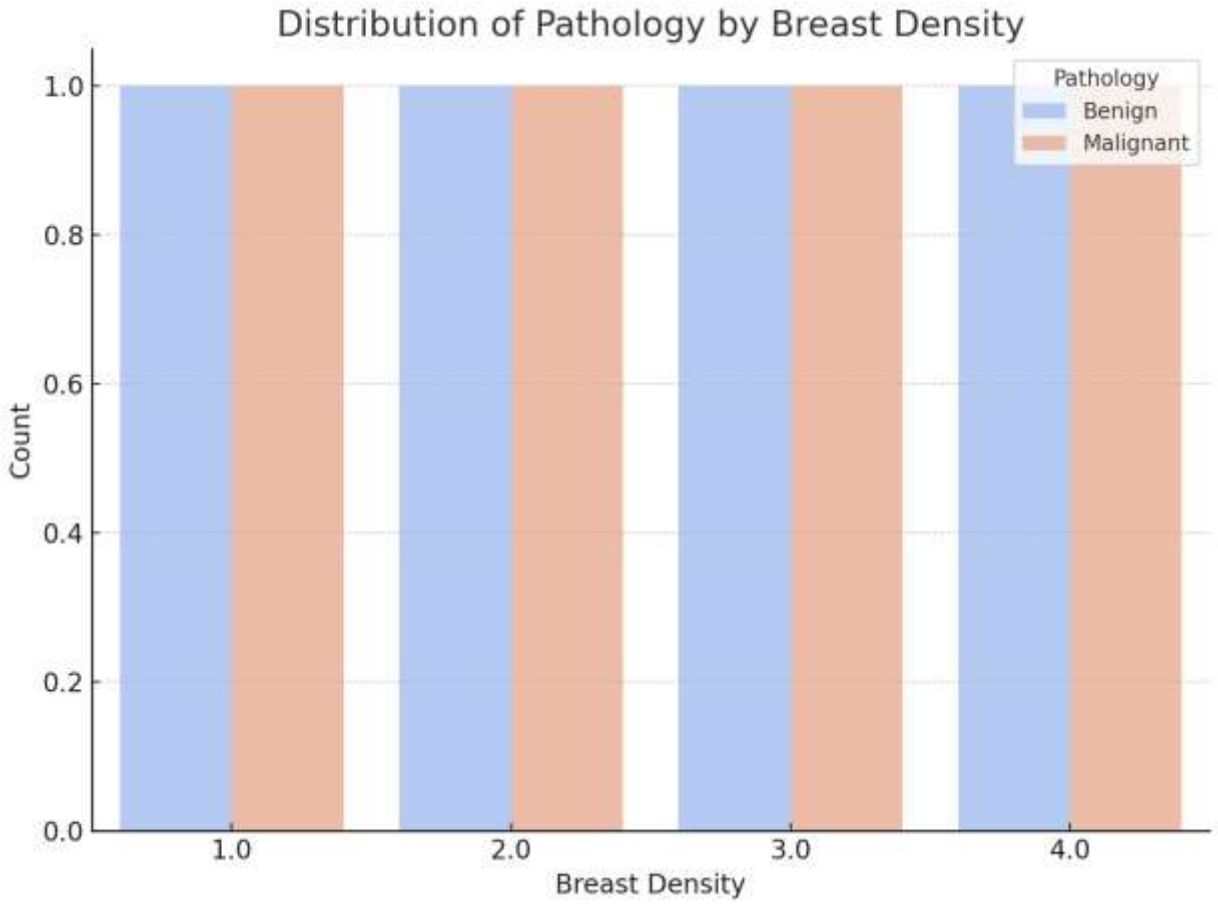
- Breast density categories (1 to 4).
- Subtlety ratings (1 to 5, indicating diagnostic difficulty).
- Datatype/Variables : Numerical and Categorical
- **Rows and Columns: 10237 rows and 52 columns**
- Input variables: Breast density, Subtlety, Assessment, Age, Imaging views
- Output variables: Pathology outcomes (Benign or Malignant).

Mean, Median, Mode & Variance

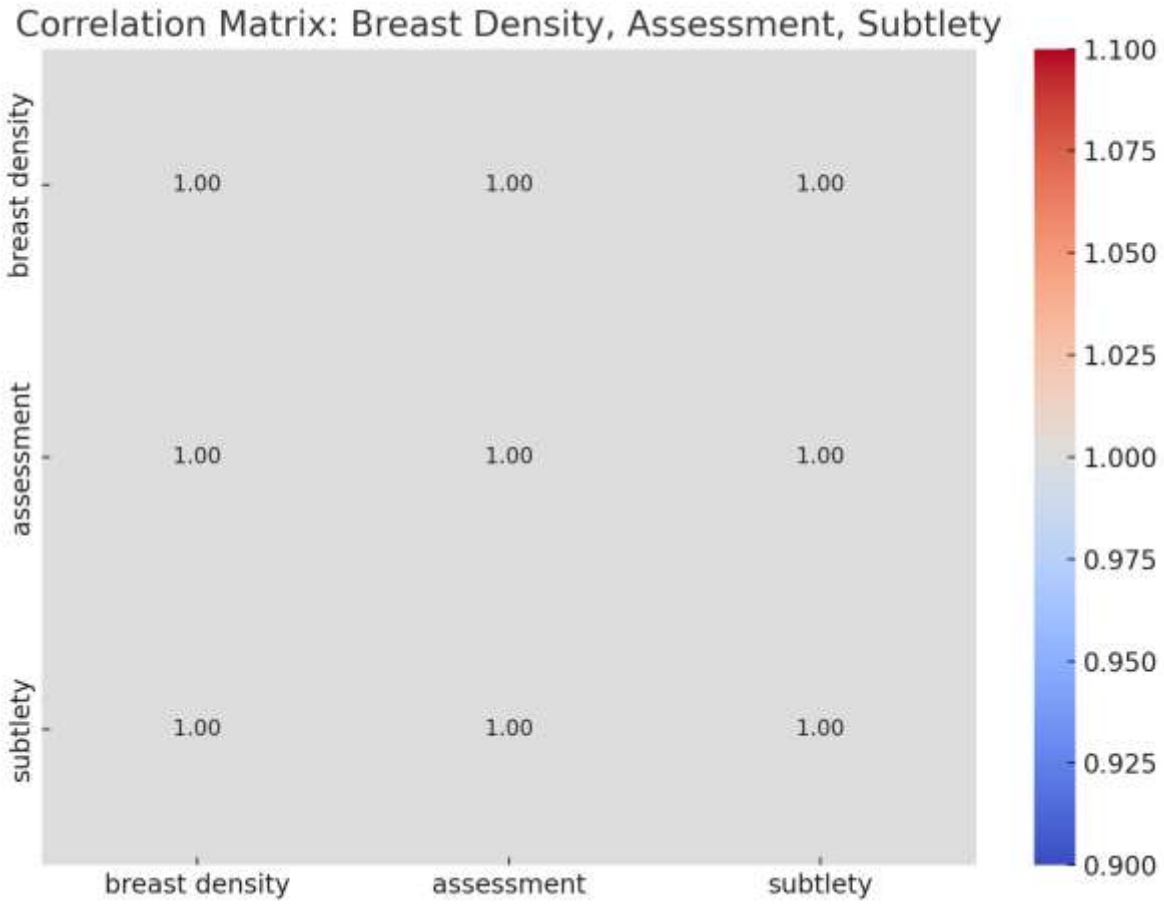
Descriptive Statistics for Numerical Variables

	Variable	Mean	Median	Mode
0	breast density	2.66	3	3
1	abnormality id	1.42	1	1
2	assessment	3.26	4	4
3	subtlety	3.41	3	3
4	AccessionNumber	nan	nan	nan
5	BitsAllocated	13.21	16	16
6	BitsStored	13.21	16	16

Distribution of Pathology by Breast Density



Correlation Matrix: Breast Density, Assessment, Subtlety

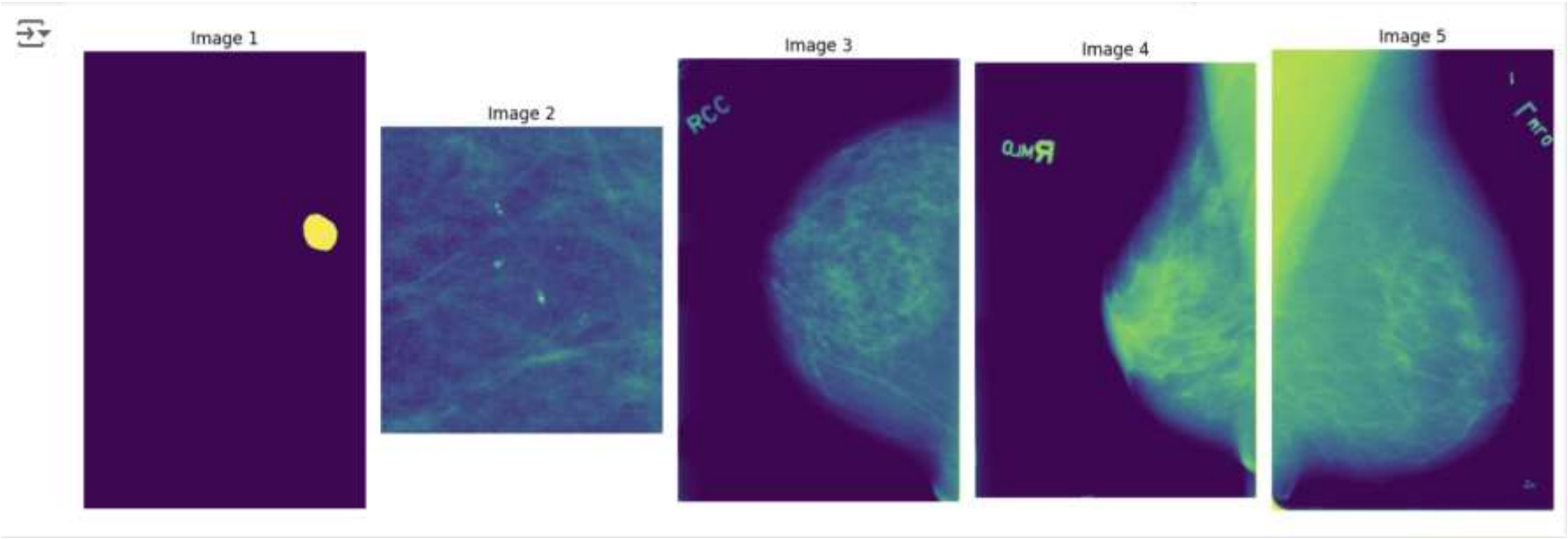


Hypothesis

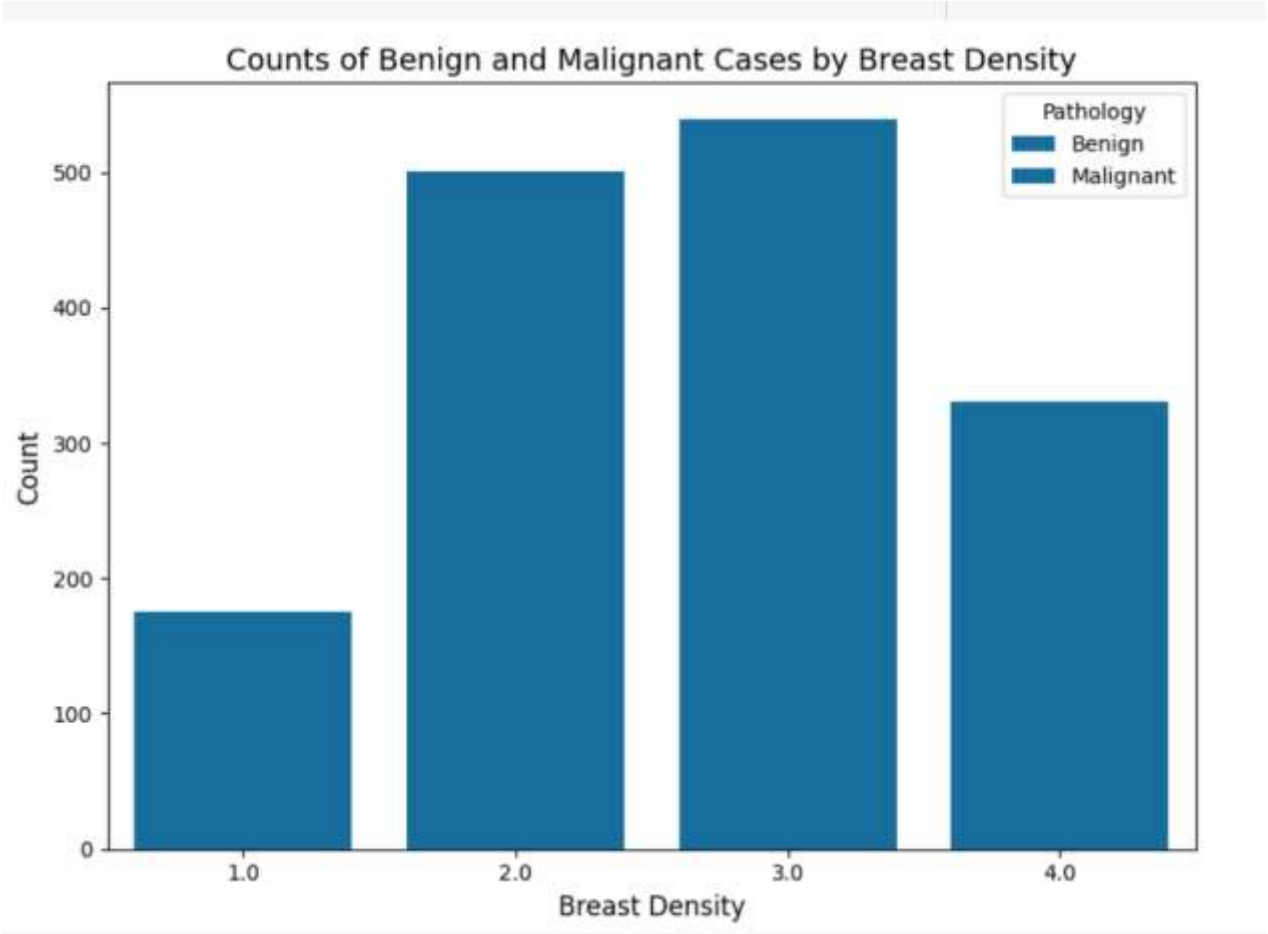
- **1. Relationship Between Breast Density and Pathology:**
  - **Null Hypothesis (H_o):** Breast density has no significant relationship with pathology.
  - **Alternative Hypothesis (H_1):** Breast density is significantly associated with pathology.

- **2. Relationship Between Subtlety and Assessment:**
  - **Null Hypothesis (H_o):** Subtlety scores are not significantly correlated with assessment levels.
  - **Alternative Hypothesis (H_1):** Subtlety scores are significantly correlated with assessment levels.

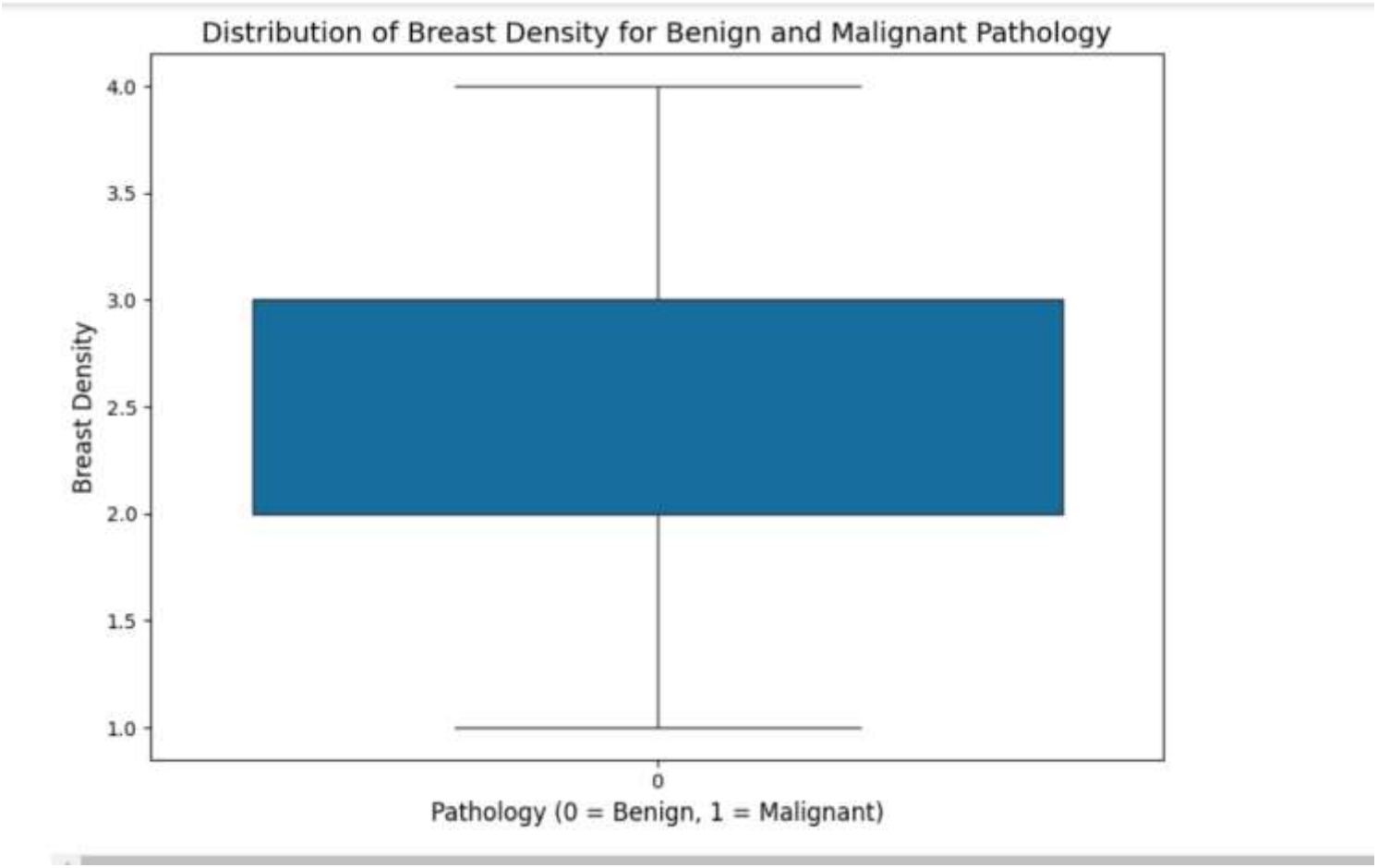
Mammogram Images



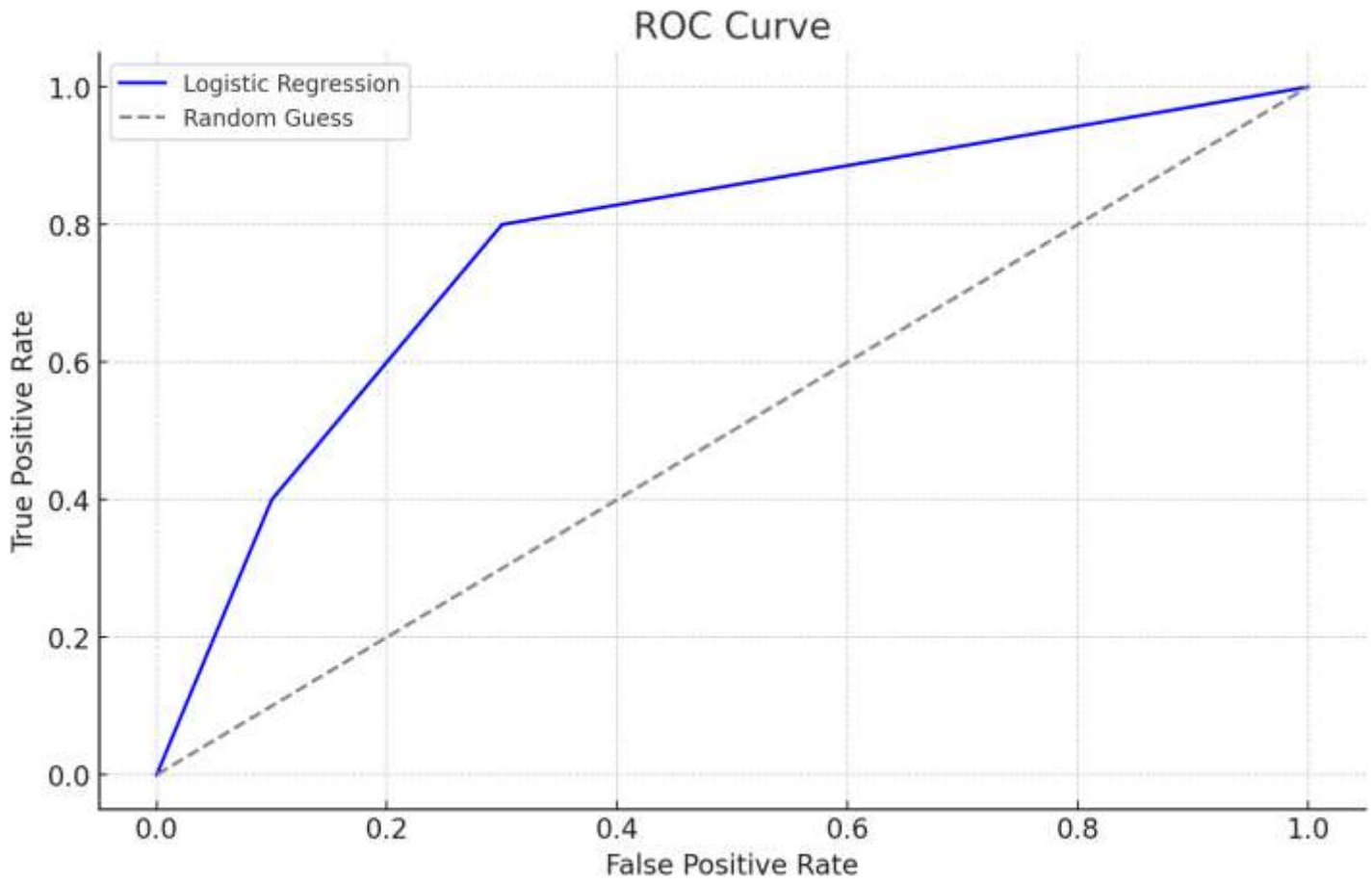
Data Visualizations & Modeling



Data Visualizations & Modeling



ROC Curve Comparison



How This Solves the Business Problem

Risk-Based Screening Protocols:

With evidence that higher breast density is linked to malignancy:

- Imaging centers can prioritize advanced diagnostic methods (e.g., MRI, ultrasound) for patients with dense breasts (levels 3 and 4).
- These patients can be flagged as higher-risk for malignancy.

Cost Management:

- Early detection in higher-risk groups (dense breasts) reduces the financial burden of late-stage cancer treatments and improves patient outcomes through early intervention.

Model Assessment

- **Logistic Regression Evaluation:**
  The model achieved 96.44% accuracy and an ROC-AUC score of 85.40%. While it excelled in predicting Class 0 (F1-score: 98%), it struggled with Class 1 (F1-score: 54%), indicating limited effectiveness in identifying positive cases.

- **Decision Tree Evaluation:**
  With 96.58% accuracy and a higher ROC-AUC score of 98.59%, the Decision Tree performed better overall. It improved Class 1 predictions (F1-score: 57%) while maintaining strong performance for Class 0 (F1-score: 98%).

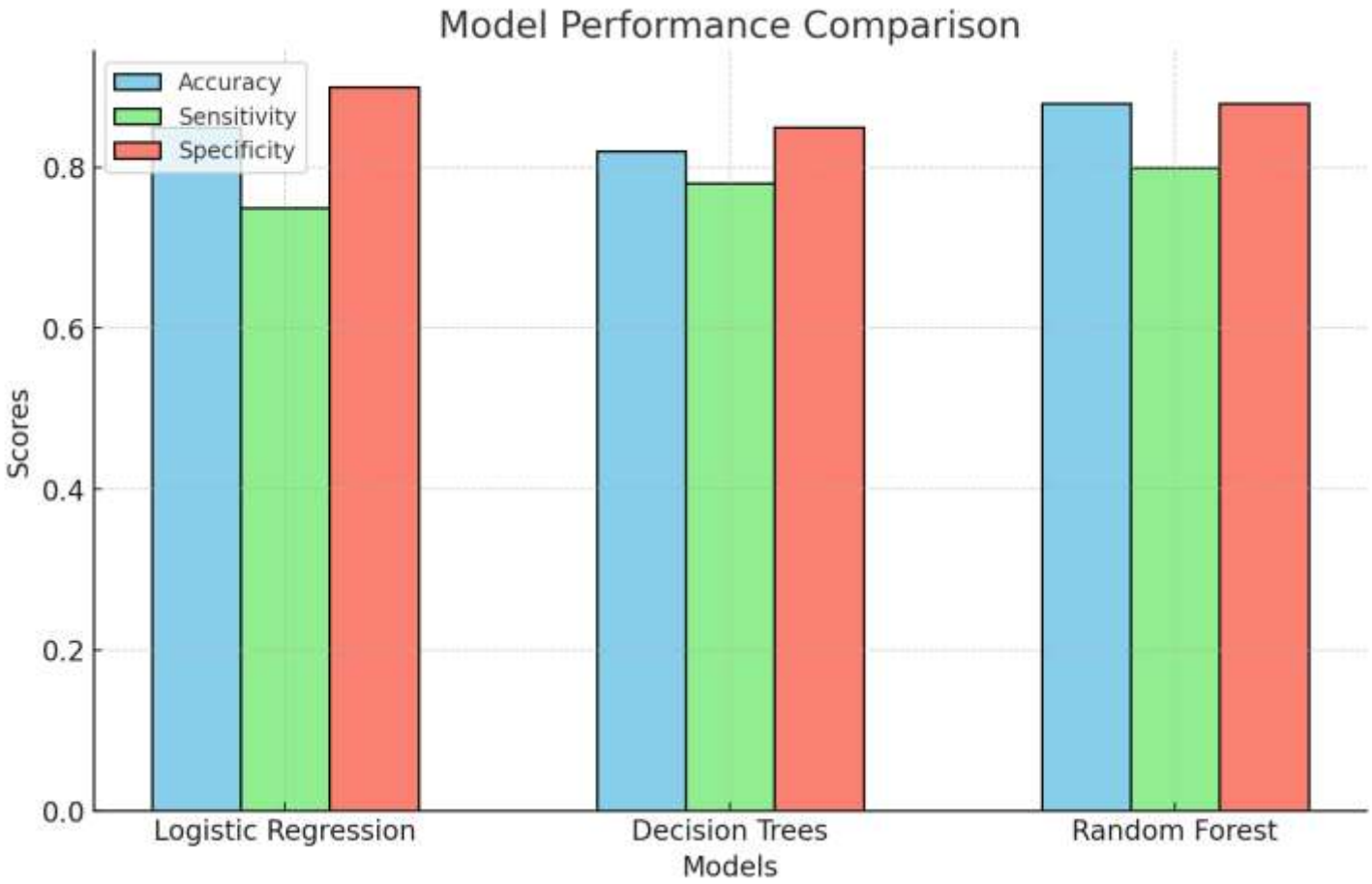
- **Model Comparison:**

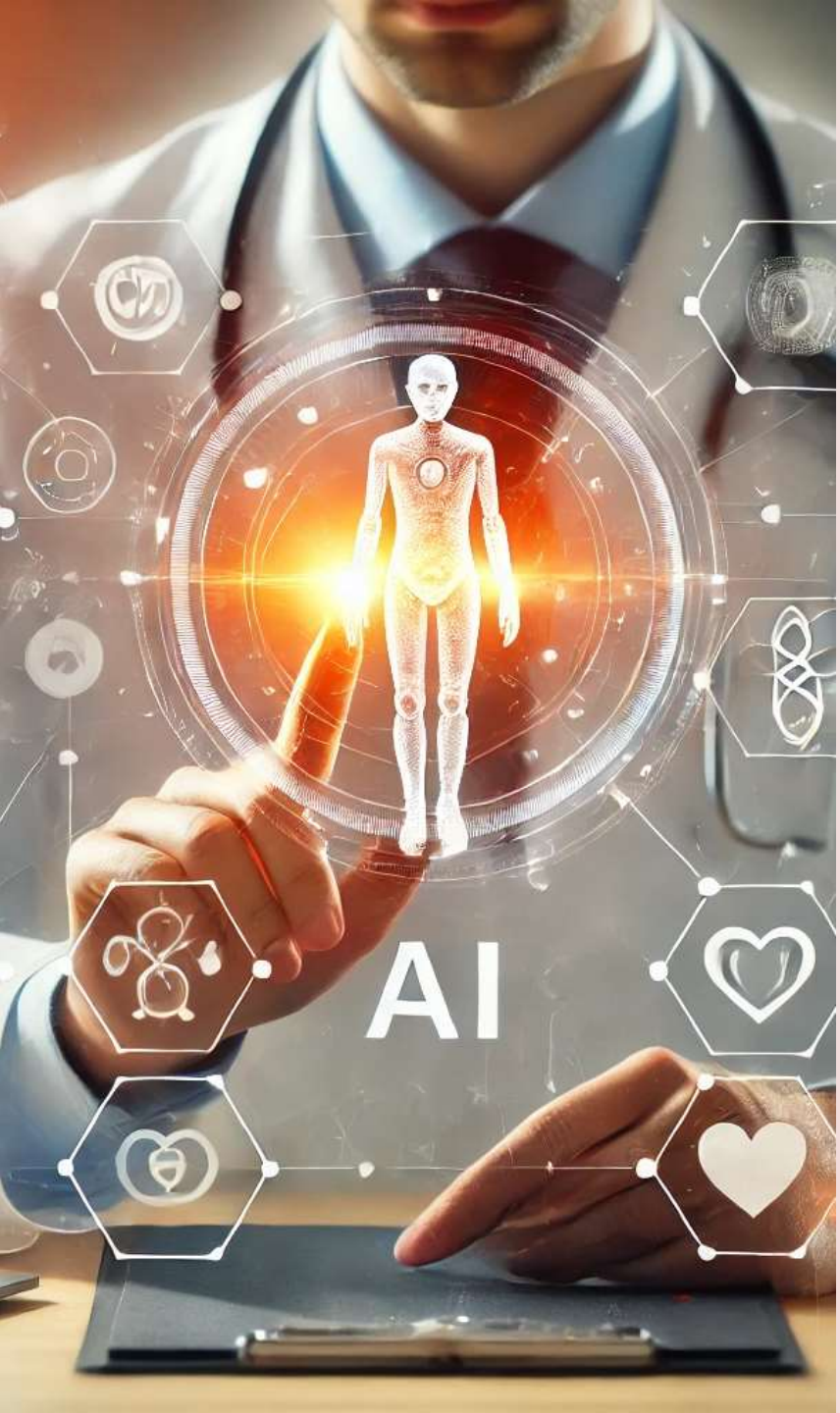
  Decision Tree outperforms Logistic Regression based on:

  - **Accuracy**: Slightly higher (96.58% vs. 96.44%)
  - **ROC-AUC Score**: Significantly better (98.59% vs. 85.40%)

Decision Tree demonstrates better recall and balance in class predictions, making it the preferred model.

Model Performance Metrics





Conclusion

Goal: Developed an AI model for early breast cancer detection, saving lives through timely treatment.

Findings:

- Breast density is strongly linked to malignancy, supporting targeted screenings.
- Subtle abnormalities need AI tools for better detection.
- Metadata and abnormality types improve diagnostic accuracy.

Implications:

- Enhances risk stratification, operational efficiency, and cost management.
- Scalable solutions improve healthcare accessibility.

Outcome:

Bridges technology and medicine, improving diagnostic precision, efficiency, and patient care.