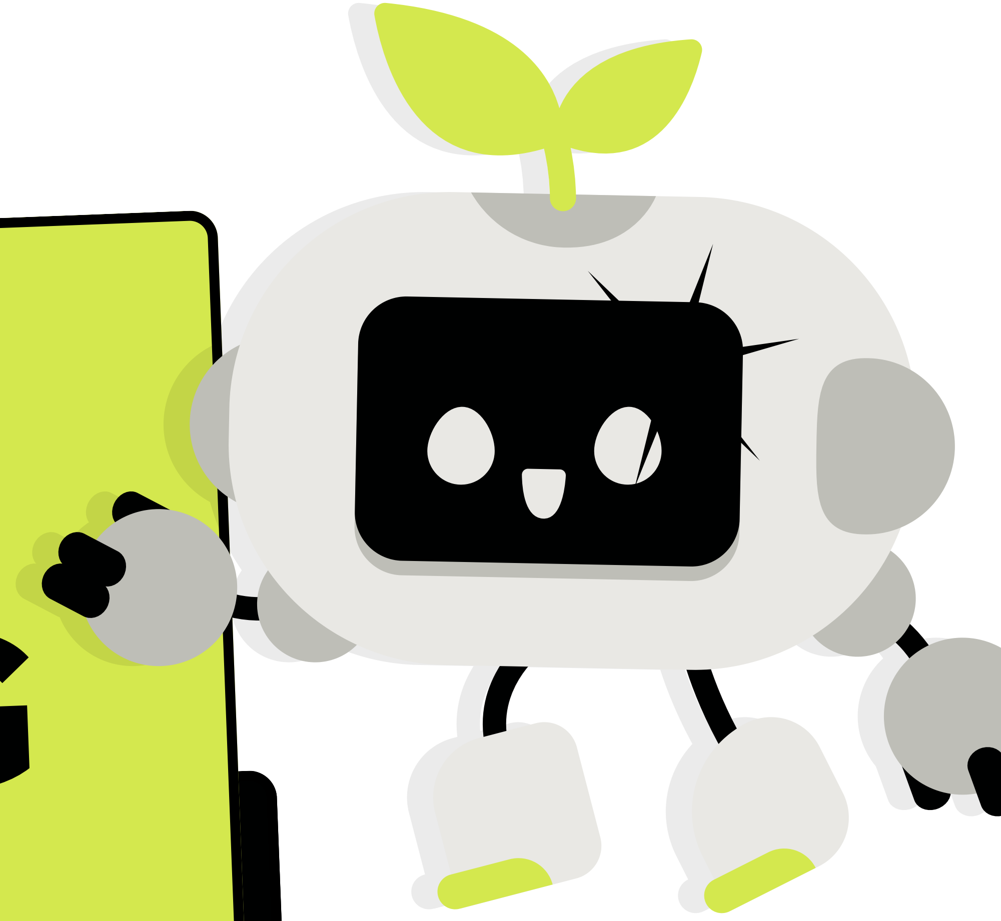
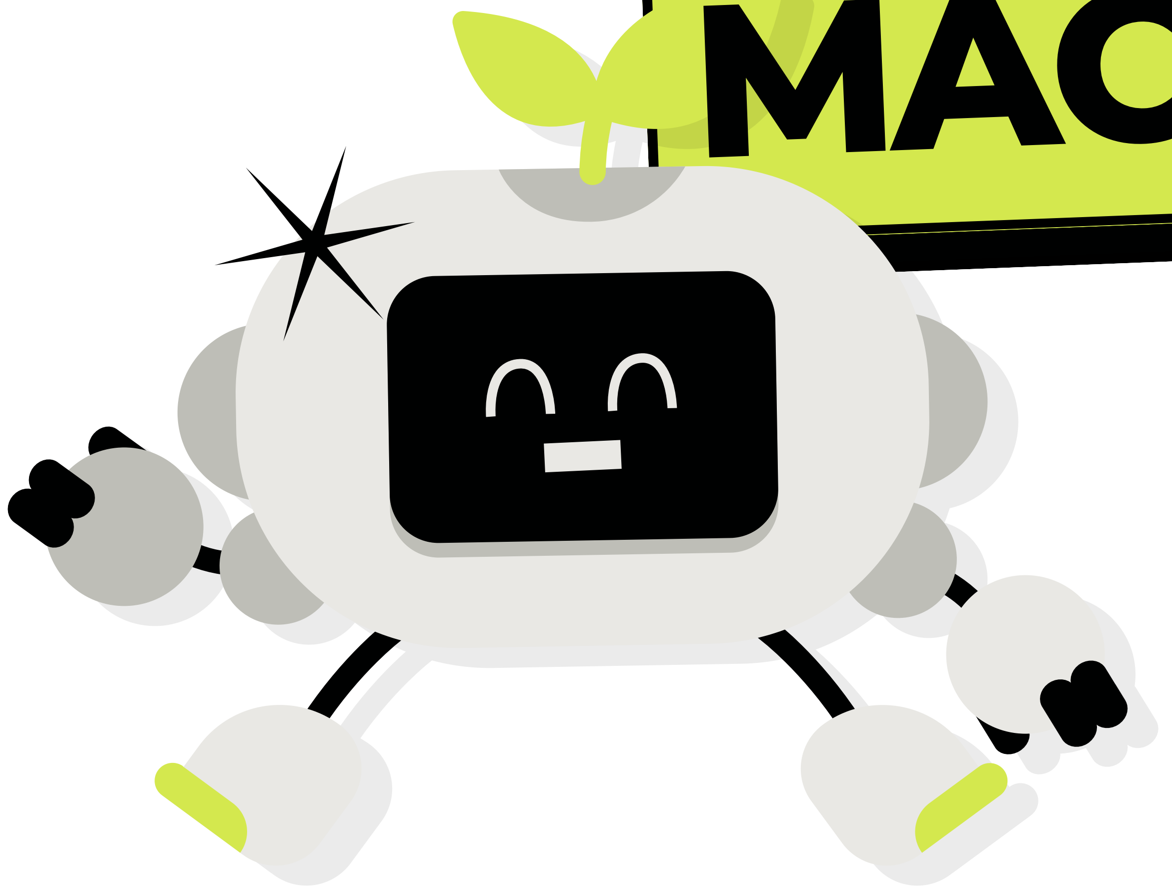
BREAST CANCER DETECTION USING MACHINE LEARNING





presented by UG group 16 -
Aadi, Andre, Gaurav, Kevin, Rhea

PROJECT OVERVIEW - DATASETS, RESEARCH QUESTIONS, AND REQUIREMENTS

01

DATASETS

- UCI Breast Cancer Wisconsin (Diagnostic)
- CBIS-DDSM Mammogram Images

02

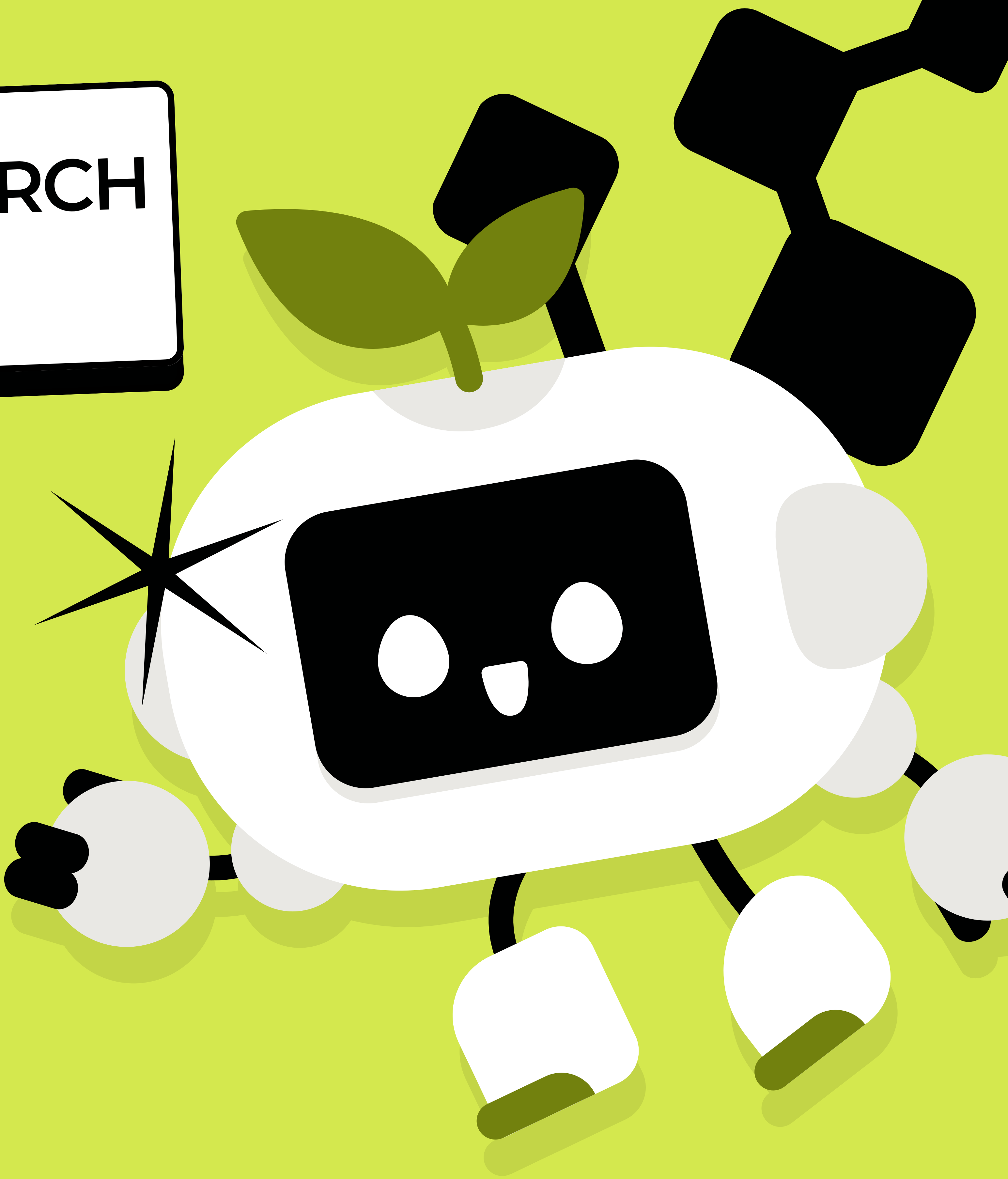
RESEARCH QUESTIONS

- Performance of classical ML models on UCI dataset
- Deep learning model performance on CBIS-DDSM

03

COURSEWORK REQUIREMENTS (R1-R4)

- R1: Define topic and objectives using datasets
- R2: Data cleaning, EDA, clustering
- R3: Baseline ML training and evaluation
- R4: Neural Networks (MLP, CNN)



UCI DATASET (DATA EXPLORATION & PREPROCESSING)

01

WHY THIS DATASET?

- Clean, structured medical dataset
- 569 samples, 30 numerical diagnostic features
- No missing values → reliable baseline
- Ideal for classical ML evaluation → helps answer RQ1

02

EXPLORATORY DATA ANALYSIS (EDA)

- Class imbalance: 63% benign / 37% malignant
- Correlation heatmap: strong multicollinearity (radius, perimeter, area)
- Feature distributions consistent with known tumour characteristics

03

PREPROCESSING STEPS

- StandardScaler applied (different feature scales)
- Label encoding (M = 1, B = 0)
- Stratified 80/20 train-test split

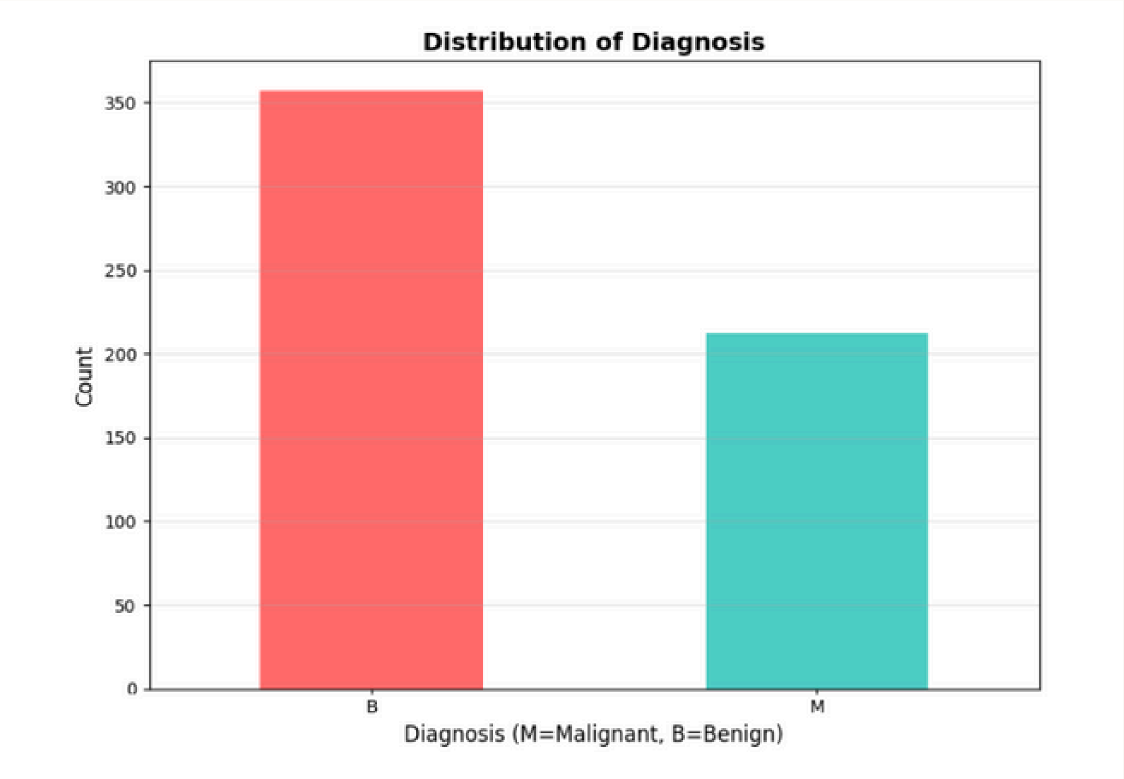
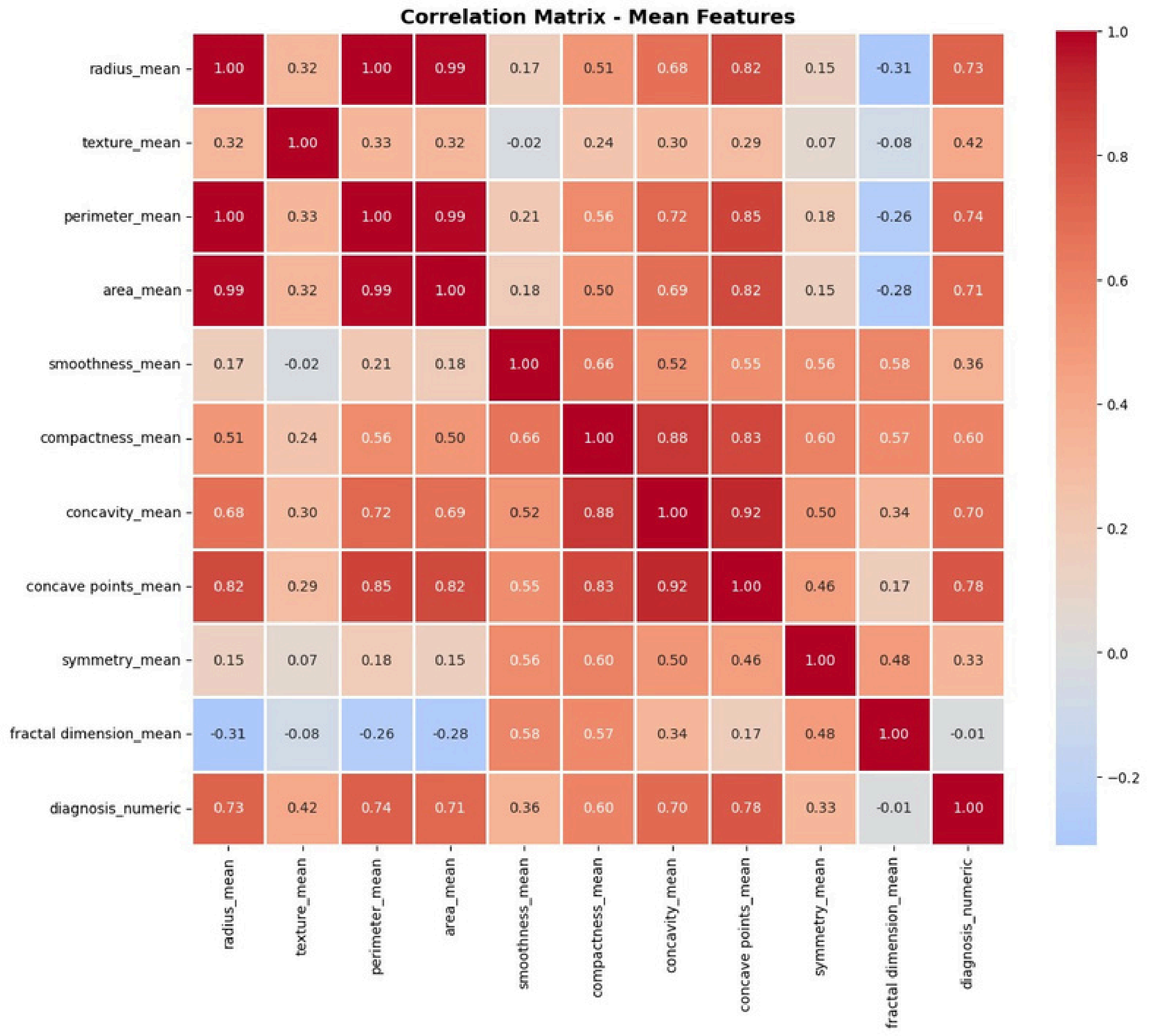


Figure 1: Class Distribution Bar Chart



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal dimension_mean	diagnosis_numeric
radius_mean	1.00	0.32	1.00	0.99	0.17	0.51	0.68	0.82	0.15	-0.31	0.73
texture_mean	0.32	1.00	0.33	0.32	-0.02	0.24	0.30	0.29	0.07	-0.08	0.42
perimeter_mean	1.00	0.33	1.00	0.99	0.21	0.56	0.72	0.85	0.18	-0.26	0.74
area_mean	0.99	0.32	0.99	1.00	0.18	0.50	0.69	0.82	0.15	-0.28	0.71
smoothness_mean	0.17	-0.02	0.21	0.18	1.00	0.66	0.52	0.55	0.56	0.58	0.36
compactness_mean	0.51	0.24	0.56	0.50	0.66	1.00	0.88	0.83	0.60	0.57	0.60
concavity_mean	0.68	0.30	0.72	0.69	0.52	0.88	1.00	0.92	0.50	0.34	0.70
concave points_mean	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1.00	0.46	0.17	0.78
symmetry_mean	0.15	0.07	0.18	0.15	0.56	0.60	0.50	0.46	1.00	0.48	0.33
fractal dimension_mean	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1.00	-0.01
diagnosis_numeric	0.73	0.42	0.74	0.71	0.36	0.60	0.70	0.78	0.33	-0.01	1.00

CBIS-DDSM (DATA EXPLORATION & PREPROCESSING)

01

WHY THIS DATASET?

- Real mammogram images used in screening
- Allows testing CNNs on clinical imaging
- Essential for answering RQ2 (image-based classification)

02

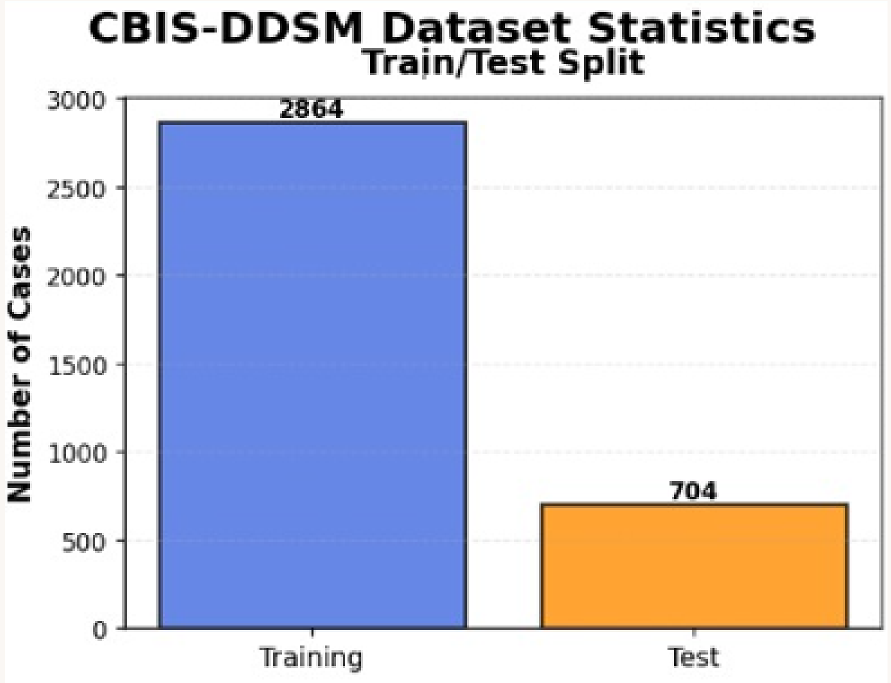
EXPLORATORY DATA ANALYSIS (EDA)

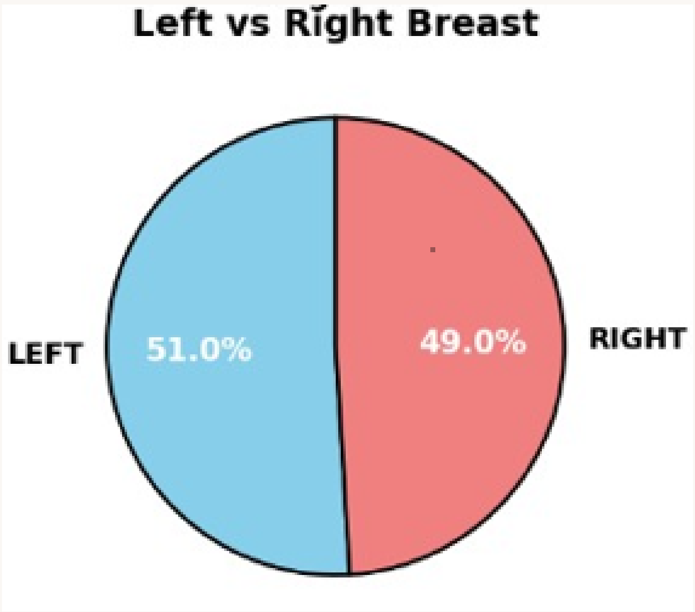
- High variation in image sizes and lighting
- Clear visible tumour regions that need careful resizing
- JPEG version used (lighter than DICOM → easier to process)

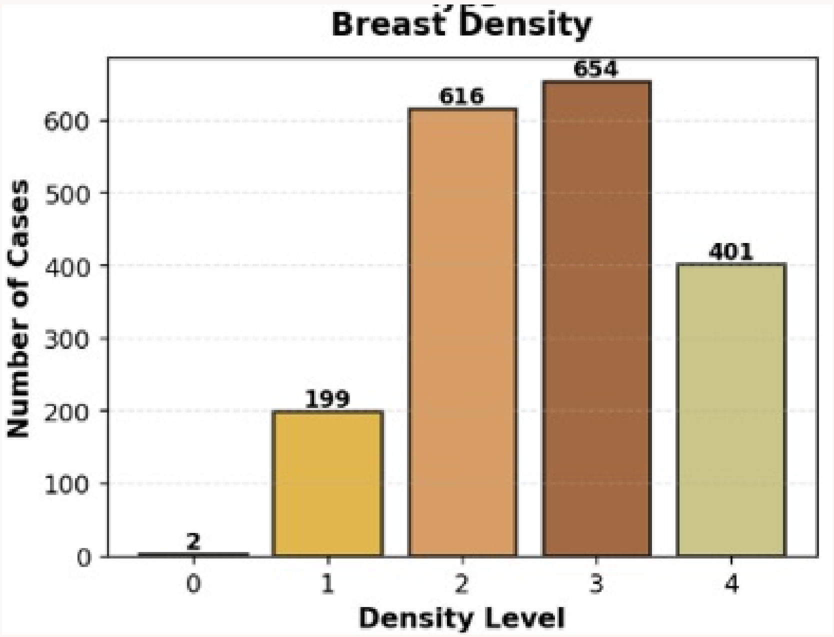
03

PREPROCESSING STEPS

- Convert to RGB (for ImageNet weights)
- Resize all images to 224×224 (ResNet input size)
- Label consolidation: malignant = 1, benign = 0
- Save as NumPy arrays for faster training
- Stratified 80/20 split to maintain balance







TABULAR MODELS TRAINING

MODELS IMPLEMENTED

- Logistic Regression
- Support Vector Machine
- Random Forest (100 trees)
- Decision Tree
- Naïve Bayes
- K-Means Clustering (unsupervised)

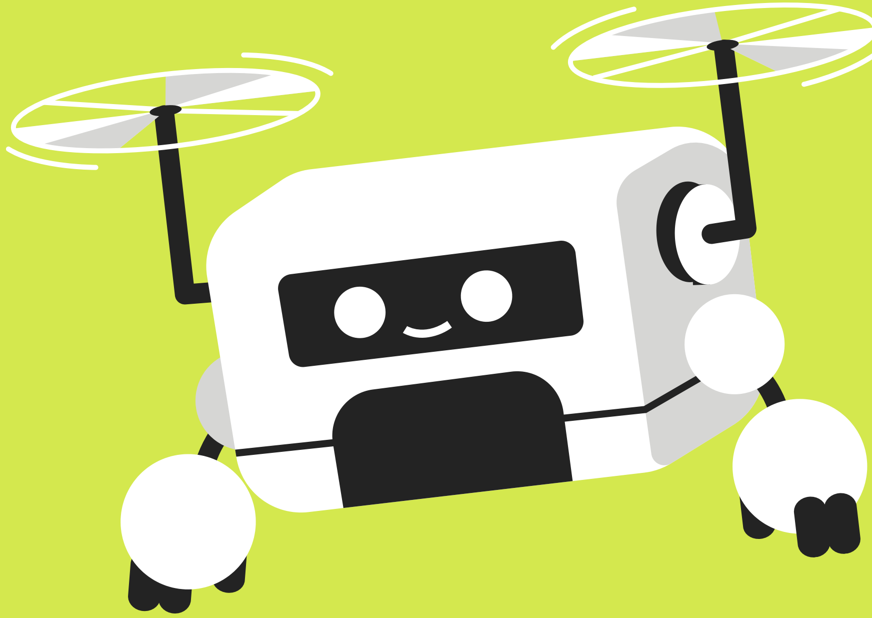
TRAINING METHODOLOGY

- StandardScaler normalization
- Train-test split (80-20)
- Hyperparameter tuning via Grid Search
- 5-fold cross-validation
- Optimization metric: F1-score

KEY DESIGN CHOICES

- SVM: RBF kernel for non-linear boundaries
- Random Forest with 100 trees
- Decision Tree: Max depth=5 to prevent overfitting
- Feature standardization critical for SVM & Logistic Regression

IMAGE MODELS IMPLEMENTED



1. MLP (MULTI-LAYER PERCEPTRON)
   - ARCHITECTURE: 16,384 → 512 → 256 → 2
   - FLATTENED 128×128 GRAYSCALE IMAGES
   - BASELINE MODEL

2. CNN
   - PRE-TRAINED ON IMAGENET
   - FINE-TUNED FINAL CLASSIFICATION LAYERS
   - ARCHITECTURE: 50 LAYERS WITH SKIP CONNECTIONS

RESULT & KEY FINDINGS

01

CLASSICAL MODELS (UCI TABULAR)

- Top models: SVM (97.37%, 100% precision), Random Forest (~97%), Logistic Regression (~96%, best recall)
- Others: Naïve Bayes (~94%), Decision Tree (~93%, interpretable)
- K-Means: Silhouette 0.343; clear natural split (9% vs 90% malignant clusters)

02

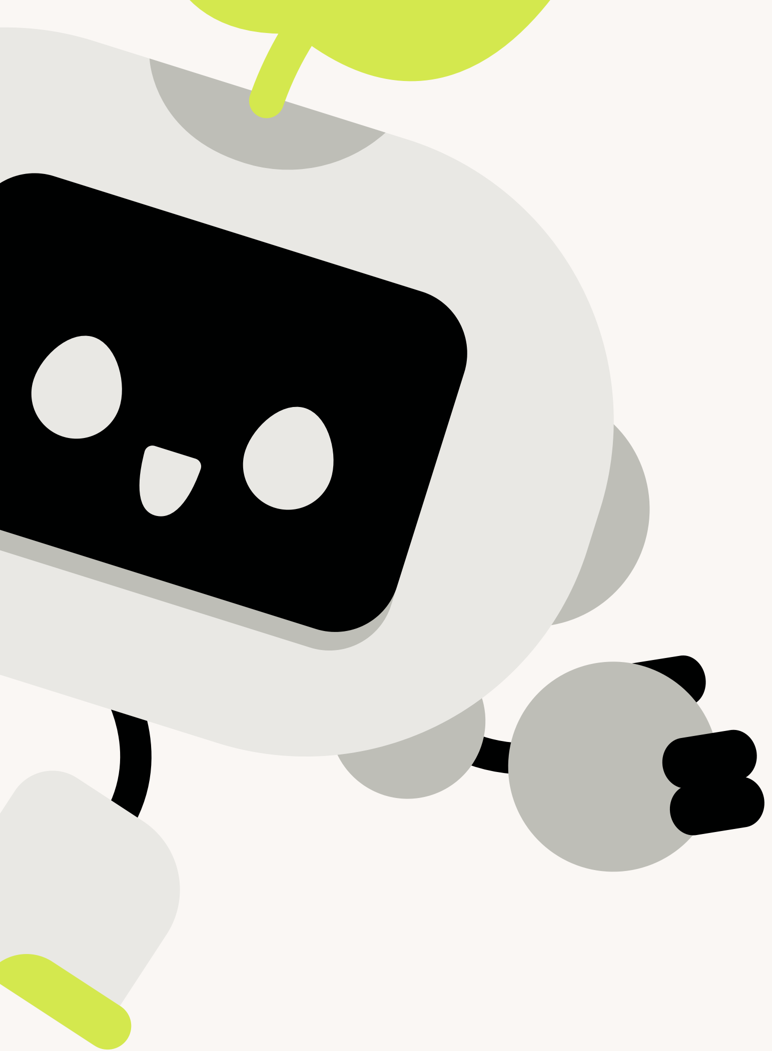
IMAGE MODELS (CBIS-DDSM)

- CNN: 74.6% accuracy, balanced precision/recall
- MLP: 52.7% accuracy, very high recall (92%) but many false positives
- CNN outperforms MLP by +22%

03

KEY INSIGHTS

- Tabular models outperform image models (97% vs 75%)
- Precision vs recall trade-off:
- SVM → no false positives
- Logistic regression → highest recall
- MLP → catches most cancers but over-flags benign cases



CONCLUSIONS

- Classical ML models achieved 97%+ accuracy on UCI diagnostic features
- Logistic Regression showed highest malignant recall, important clinically
- CNN (ResNet-18) achieved best image performance: 0.746 accuracy
- MLP achieved high recall but poor precision: unsuitable for clinical use
- Structured data = easy to model; images = require deeper architechtures.

LIMITATIONS

- CBIS-DDSM is large, high-resolution, and computationally demanding
- Limited compute / time-restricted hyperparameter tuning and epochs
- MLP and baseline CNN attempts showed overfitting without augmentation

MEMBER CONTRIBUTIONS

- Each member implemented different ML models + report sections
- Weekly syncs to merge code, review plots, ensure consistent methods
- Shared debugging for CNN/MLP training issues & dataset organisation
- Final report and slides compiled collaboratively for coherence