BREAST CANCER CLASSIFICATION

Josh Ascher and Seyed Amir Mohammad Bagheri

Problem

- **Prevalence of Breast Cancer:**
  - Breast cancer is the most common cancer among women globally, accounting for a significant proportion of cancer diagnoses.
  - Early and accurate detection is crucial for improving survival rates and quality of life.
  - Traditional diagnostic methods like biopsies are invasive and time-consuming.
- **Impact of Accurate Classification:**
  - Reducing false negatives is critical—missing a malignant tumor can have severe consequences.
  - Reducing false positives can minimize unnecessary procedures, reducing patient stress and healthcare costs.
- **Public Health Implications:**
  - Empowering healthcare systems, especially in resource-limited settings, with cost-effective diagnostic tools.
  - Contributing to global efforts in cancer screening and prevention strategies.

Data

- **Diagnostic Breast Cancer Wisconsin Dataset (1995)**
  - Available on UCI Machine Learning Repository
- **569 observations, 30 features**
  - Features are from images of fine needle aspirate of breast mass
    - Tumor characteristics: Area, texture, concavity, etc.
  - Class imbalance
    - (62.7%) benign, (37.3%) malignant
- **Binary classification of masses as benign or malignant**

Prior and Related Work:

- **This dataset has been widely used in academic and research settings for classification tasks and predictive modeling in healthcare**
  - 30+ papers over nearly two decades
- **Extensive studies have applied machine learning algorithms for breast cancer detection using this dataset**
  - Logistic regression, SVM, PCA, ensembles, etc.
- **Challenges Highlighted by Previous Works**
  - High dimensionality relative to the sample size increases the risk of overfitting
  - Non-linear relationships in data require advanced models but increase interpretability challenges

Approach:

The primary objectives of this project : identify the most significant features and classify instances as benign or malignant.

Feature significance will be analyzed using dimensionality reduction techniques like PCA.

Two binary classifiers, Logistic Regression and Support Vector Machines (SVMs), will be implemented and evaluated for performance.

Model performance will be assessed using standard metrics, including accuracy, precision, recall, and F1-score.

Standard machine learning workflows, including data splitting (training and testing) and cross-validation, will be employed to ensure reliable results.

References

- Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.
- *Looveren, Arnaud Van and Janis Klaise. "Interpretable Counterfactual Explanations Guided by Prototypes." ArXiv abs/1907.02584 (2019): n. pag.*