Léonard Seidlitz
Romain Requena
Adrien Servas
Raphaël Roux

Machine Learning Pre-Project
Breast Cancer Classification: State-of-the-Art

1. Business Challenge and State-of-the-Art

Breast cancer diagnosis remains a critical task in healthcare, where early and accurate detection of malignant tumors can significantly improve patient outcomes. The main challenge in this project is to leverage machine learning to classify breast tumors as benign or malignant based on specific visual characteristics.

Recent advances in machine learning have enhanced cancer diagnosis, utilizing classification algorithms like **Logistic Regression**, **K-Nearest Neighbors (KNN)**, and **Support Vector Machines (SVM)**, all of which can be suitable for binary classification problems. These algorithms have shown effectiveness in scenarios where clear distinctions between classes (malignant and benign tumors) are observable through measurable features. In this project, we aim to assess and implement one or more of these algorithms to develop a robust predictive model that identifies cancer types based on patients' tumor characteristics.

2. Data Description and Data Sources

The dataset under analysis is the **Breast Cancer Wisconsin (Diagnostic) Dataset**, publicly accessible through the UCI Machine Learning Repository. Key characteristics include:

- **Features**:
  - *Unique ID*: Each patient has a unique identifier.
  - *Diagnosis*: Binary label indicating tumor type—malignant ("M") or benign ("B").
  - *Visual characteristics*: A set of mean values for tumor features, such as `radius_mean`, `texture_mean`, `perimeter_mean`, `area_mean`, `smoothness_mean`, `compactness_mean`, `concavity_mean`, and `concave points_mean`.

These features collectively offer a quantitative overview of the tumor's shape, size, and other textural properties. These characteristics serve as inputs to classify the tumor type.

- **Source**: The dataset is available from multiple repositories, including:
    - Kaggle
    - UCI Machine Learning Repository

3. Business Objectives and Project Scope

The primary objective is to construct a machine learning model that can reliably classify a breast tumor as benign or malignant based on patient data. Achieving this can facilitate earlier diagnosis, guide medical intervention, and potentially reduce healthcare costs by reducing the need for invasive diagnostic procedures.

Secondary objectives include:

- Identifying the most relevant visual features that contribute to accurate classification.
- Experimenting with and comparing multiple machine learning models to identify the best-performing algorithm in terms of accuracy and computational efficiency.
- Exploring feature engineering and selection techniques to improve model interpretability and performance.

Scope:

- **Data Preprocessing**: Cleaning, normalizing, and preparing the data for modeling.
- **Model Training and Testing**: Experimenting with classification algorithms (Logistic Regression, KNN, SVM).
- **Evaluation Metrics**: Accuracy, precision, recall, and F1-score for model evaluation.

4. Work Plan

Phase 1: Data Exploration and Preprocessing

- Conduct initial exploratory data analysis (EDA) to understand feature distributions and identify any missing or anomalous values.
- Normalize the features to ensure consistent scaling across all input variables.
- Split the dataset into training and test sets (80/20 split) to facilitate model training and evaluation.

Phase 2: Model Selection and Training

- Implement baseline models: Logistic Regression, K-Nearest Neighbors, and Support Vector Machines.
- Optimize each model using techniques such as grid search for hyperparameter tuning.
- Compare models based on evaluation metrics such as accuracy and recall, with particular emphasis on recall (to minimize false negatives).

Phase 3: Evaluation and Optimization

- Analyze model performance on the test dataset and identify any areas for improvement.
- Consider additional techniques, such as feature selection or dimensionality reduction (PCA), to improve model interpretability and accuracy.
- Fine-tune the best-performing model for deployment readiness.

Phase 4: Reporting and Documentation

- Document the methodology, results, and challenges encountered during model development.
- Prepare visualizations to communicate model results effectively (ROC curve, confusion matrix).
- Summarize findings in a final report, detailing the model's potential applications and any recommendations for future work.

5. Conclusion

This project aims to leverage machine learning techniques to improve the diagnostic process for breast cancer, focusing on distinguishing benign from malignant tumors with high accuracy. By training on comprehensive visual characteristics, we expect to enhance prediction accuracy, thus supporting healthcare providers with a reliable diagnostic tool. Through model experimentation and evaluation, we anticipate gaining insights into the most influential features for tumor classification, guiding future work in medical diagnostics.

6. References

1. UCI Machine Learning Repository. (n.d.). Breast Cancer Wisconsin (Diagnostic) Dataset. Retrieved from

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

2. Kaggle. (n.d.). Breast Cancer Data Set. Retrieved from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
3. Sokullu, E. T. (n.d.). Cancer Prediction Notebook Series. Kaggle. Retrieved from:
    a. K-Nearest Neighbors: https://www.kaggle.com/code/erdemtaha/prediction-cancer-data-with-k-nn-95
    b. Logistic Regression: https://www.kaggle.com/code/erdemtaha/cancer-prediction-96-5-with-logistic-regression
4. Creative Commons License. Retrieved from https://creativecommons.org/licenses/by-nc-sa/4.0/