CS 7641 Machine Learning (ML)

Predicting Breast Cancer Tumor Malignancy Using Machine Learning Algorithms

Haocheng Yu
Tianyi Yu
Yuxin Sun
Zihan Chen
Ziyi Zhou

June 14, 2024

Summer 2024
Instructor: Dr. Max Mahdi Roozbahani

Introduction/Background

Literature Review

Breast cancer is one of the most prevalent cancers affecting women worldwide. Accurate prediction of tumor malignancy can significantly enhance diagnosis and treatment planning. Previous studies have leveraged machine learning (ML) algorithms to improve malignancy prediction, demonstrating the potential of ML to surpass traditional statistical methods in accuracy and reliability. For instance, a study by Lu and Liu [1] explored various ML models, such as K-Nearest Neighbors and Gradient Boosting Models, to predict breast cancer survivability. Their results highlighted the importance of integrating clinical features with molecular data to improve prediction accuracy.

Additionally, research has shown that molecular diagnostics can complement traditional clinical tools, potentially enhancing the accuracy of breast cancer prognosis [2]. However, there is still no consensus on the most effective computational methods for predicting breast cancer outcomes, emphasizing the need for further exploration in this area [3].

Dataset Description

The dataset used in this project is obtained from Kaggle and is originally from the SEER Program of the National Cancer Institute (NCI). It includes records of female patients diagnosed with infiltrating duct and lobular carcinoma breast cancer between 2006 and 2010. The dataset contains 4024 patients after excluding those with incomplete records or survival times of less than one month. Key features include age, race, marital status, T stage, N stage, differentiation, grade, A stage, and tumor size.

Dataset Link

The dataset can be accessed on Kaggle: `https://www.kaggle.com/datasets/reihanenamdari/breast-cancer/data`Breast Cancer Dataset. The file used in this project is named `Breast_Cancer.csv`.

Problem Definition

Problem

The primary challenge is to accurately classify breast tumors as malignant or benign using patient medical records and tumor characteristics. This prediction is crucial for early diagnosis and effective treatment planning.

Motivation

Early and precise identification of malignant tumors can significantly improve patient outcomes, guiding timely and appropriate medical interventions. The development of an accurate predictive model can assist healthcare professionals in making informed decisions, ultimately enhancing patient care and survival rates.

Methods

Data Preprocessing Methods

- **Feature Selection:** Identify and retain relevant features for predicting tumor malignancy, such as age, race, marital status, T stage, N stage, differentiation, grade, A stage, and tumor size.
- **Handling Missing Values:** Use imputation techniques to handle any missing data.
- **Feature Scaling:** Apply MinMaxScaler to normalize feature values for consistent model training.

ML Algorithms/Models

- **CatBoost Classifier:** Known for handling categorical data effectively and providing robust performance.
- **Gradient Boosting Classifier:** An ensemble method that builds models sequentially to minimize errors.
- **AdaBoost Classifier:** Focuses on difficult cases by adjusting weights, enhancing recall for malignant tumors.
- **Random Forest Classifier:** An ensemble method that builds multiple decision trees to improve accuracy and control overfitting.
- **Support Vector Machine (SVM):** Effective in high-dimensional spaces and for binary classification.

CS 7641 Learning Methods

- **Supervised Learning:** Primary method, focusing on classification of tumor malignancy.
- **Unsupervised Learning:** Exploratory analysis to identify patterns and relationships within the data.

(Potential) Results and Discussion

Quantitative Metrics

- **Accuracy:** Measure of overall correctness of the model.
- **Precision:** Measure of the model's ability to correctly identify malignant tumors.
- **Recall:** Measure of the model's ability to detect all malignant tumors.
- **F1-Score:** Harmonic mean of precision and recall, providing a balance between the two.
- **ROC-AUC:** Area under the receiver operating characteristic curve, evaluating model performance across all classification thresholds.

Project Goals

- Achieve high accuracy, precision, recall, and F1-score.
- Identify the most influential features for predicting tumor malignancy.
- Develop a robust and reliable predictive model for clinical use.

Expected Results

- A machine learning model capable of accurately predicting the malignancy of breast cancer tumors.
- Insights into key features influencing malignancy, aiding healthcare professionals in diagnosis and treatment planning.
- Potential for early detection of malignant tumors, improving patient outcomes and survival rates.

References

References

[1] C. Lu and J. Liu, "Breast Cancer Prognosis," Stanford University, 2012. Available: `https://cs229.stanford.edu/projects2012.html`.

[2] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," Nature, 2012.

[3] S. Paik et al., "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," N. Engl. J. Med., vol. 351, no. 27, pp. 2817-2826, Dec. 2004.

Gantt Chart

TASK TITLE	TASK OWNER	START DATE	DUE DATE
Project Team Composition	All	5/13/2024	5/24/2024
Project Proposal			
Introduction & Background	Zihan Chen	5/24/2024	6/14/2024
Problem Definition	Zihan Chen	5/24/2024	6/14/2024
Methods	Tianyi Yu & Yuxin Sun	5/24/2024	6/14/2024
Potential Dataset	Tianyi Yu & Yuxin Sun	5/24/2024	6/14/2024
Potential Results & Discussion	Tianyi Yu & Yuxin Sun	5/24/2024	6/14/2024
Video Creation & Recording	Haochen Yu	5/24/2024	6/14/2024
GitHub Page	Ziyi Zhou	5/24/2024	6/14/2024

Contribution Table

Name	Proposal Contributions
Haocheng Yu	Video Presentation
Tianyi Yu	Report Writing
Yuxin Sun	Report Writing
Zihan Chen	Report Writing
Ziyi Zhou	Website Establishment