Classification of Benign vs Malignant Tumors

Group 5 Final Report

Prepared by:

Reese Gillan

Ridwan Hossain

Fisher Latham

Samirrah Nyo

Miguel Soto

Submission Date:

04/21/2024

Georgia Institute of Technology

Georgia Tech's Colleges of Computing, Business, and Engineering.

225 North Ave, Atlanta, GA 30332-0530, United States

Table of Contents

Abstract

Breast cancer is a predominant type of malignant tumor. In our study we hope to classify breast cancer tumors utilizing breast cancer imaging data. Imaging data can reveal the geometry of tumors leading researchers to determine the severity of breast cancer tumors. We hope to classify precisely whether a tumor is benign or malignant using the Wisconsin breast cancer dataset. We have conducted logistic regression,Random Forest and SVM to classify the tumors. Additionally, we are also looking at the Queen's University Belfast cancer real breast cancer data set to predict the likelihood of patient survival.The classification of tumor types that would allow doctors to identify and accurately treat breast cancer tumors.

Github Link: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/MGT6203_Project_Team-5

Introduction

Background

Image processing has been an effective tool in breast cancer detection. This involves discovering potential cancerous regions with breast regions. Interactive image processing techniques have been used to create features that can be exploited for classification and diagnosis of breast tumors. For the data used in this study, a small fraction of a fine needle aspirate slide is selected and digitized. With an interactive interface, the user initializes active contour models, known as snakes, near the boundaries of a set of cell nuclei. The customized snakes are deformed to the exact shape of the nuclei. This allows for precise, automated analysis of nuclear size, shape and texture. Ten such features are computed for each nucleus, and the mean value, largest (or 'worst') value and standard error of each feature are found over the range of isolated cells. We are utilizing logistic regression,Random Forest and SVM to classify the tumors utilizing the Breast Cancer Wisconsin (Diagnostic) Data Set. This dataset contains the features and characteristics of a digitized image from a fine needle aspirate (FNA) of a breast mass. We hope to determine whether or not a tumor is benign or malignant through our analysis.We are also utilizing the Queen's University Belfast cancer research real breast cancer data set to predict whether the patient will survive or not. Comparison of these two analyses should help us determine which is the best method for classifying breast cancer tumors. There are many advantages to addressing this issue with this technique. Enhanced accuracy and an accelerated process of distinguishing between various benign and malignant tumors would allow doctors to administer personalized treatments and choices to patients, therefore leading to the saving of numerous lives.

Problem Statement

Using a form of classification, determine whether or not a tumor is benign or malignant.

Data Overview

Data Sources & Research

Machine learning (ML) is becoming an important tool in the fight against cancer. ML allows the application of different techniques to improve the accuracy of diagnosis, prognosis, and treatment strategies. One of the main methods is the use of diagnostic imaging in which different ML techniques are applied to classify digitized images of breast masses as benign or malignant. Our dataset contains information from diagnostic imaging and we will use different ML techniques to predict cancer cases. We will then evaluate the different methods applied to determine which is the most accurate.

Logistic Regression

Logistic Regression (LR) is widely used in medical research for its simplicity and high level of explainability. It models the probability of the occurrence of an event by fitting data to a logistic curve. Its application in cancer prediction has been substantial due to its interpretability, allowing clinicians to understand the relationship between predictors and the likelihood of cancer presence (Kumar & Gota, 2023). Kumar and Gota (2023) explain the effectiveness of LR in breast cancer risk prediction, using demographic and lifestyle factors as predictors. The model's transparency in showing how each factor contributes to the risk makes it a valuable tool for informing patients and clinicians. However, LR's linear nature limits its ability to handle complex, non-linear relationships without transformation or augmentation of features.

Support Vector Machine Classification

SVM is powerful and is capable of performing linear and non-linear classification by finding the optimal hyperplane that separates different classes in the feature space. Its application in cancer prediction is often cited for its ability to handle high-dimensional data with a relatively small number of samples (Kamel et al., 2019). SVM can capture complex patterns and relationships in the data, enhancing its predictive performance. According to Kamel et al. (2019), SVM is proficient in classifying cancer types based on gene expression data, achieving remarkable accuracy. Despite its advantages, SVM's "black-box" nature, particularly with non-linear kernels, poses challenges in interpretability, making it difficult for practitioners to grasp how decisions are made.

Random Forest and Gradient Boosting

Random Forest (RF) and Gradient Boosting are ensemble learning methods that combine multiple decision trees to improve prediction accuracy and control overfitting. Hassan et al. (2023) note that RF builds numerous decision trees and merges their results to get a more accurate and stable prediction. On the other hand, Gradient Boosting builds trees sequentially, with each new tree correcting errors made by previously trained trees (Hassan et al., 2023). Both techniques have been effectively used in cancer prediction, offering advantages in handling various types of data, including imbalanced datasets common in cancer diagnosis.

Challenges and Considerations

Data Heterogeneity: One of the primary challenges in ML research on cancer data is the heterogeneity and complexity of the data, including variations in data types (e.g., imaging, genetic, clinical) and in cancer biology itself. Model Interpretability: As ML models, especially deep learning, become more complex, their interpretability diminishes. This poses a significant challenge in clinical applications where understanding the decision-making process is crucial for trust and ethical reasons. Ethical and Privacy Concerns: The use of sensitive patient data in ML research raises ethical and privacy concerns. Ensuring data security and patient privacy while maintaining data utility for research is an ongoing challenge.

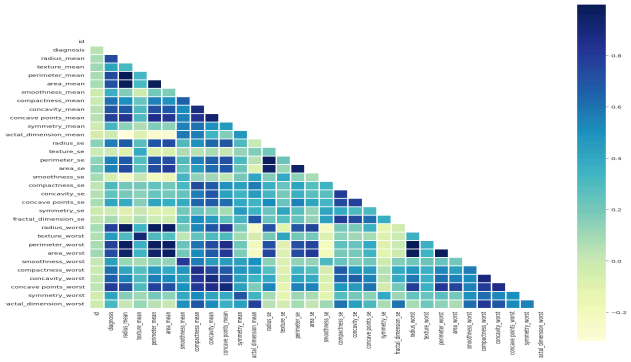
Data Cleaning

Data Clean up link → Team 5 - Github

Exploratory Data Analysis / Feature Generation

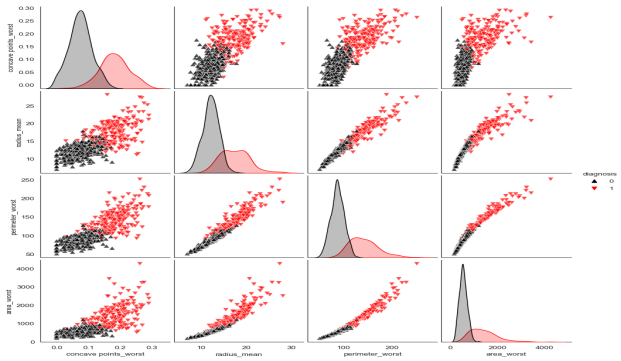
Wisconsin Breast Cancer Dataset

Exploratory Data Analysis is a critical part prior to model ingestion as this is the basic understanding of the data and which patterns are present. Detection of outliers and anomalies is essential for ensuring a quality



and reliable analysis. The first graph above is a correlation matrix heatmap based on the Wisconsin Breast Cancer dataset.

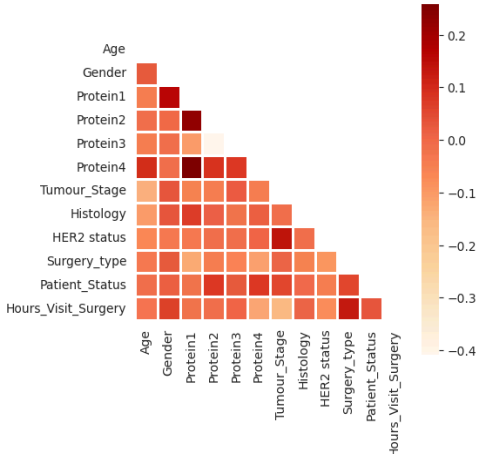
From the correlation matrix, we see a moderate to strong correlation between the diagnosis and several variables. The concave points_worst, radius_mean, and concave_points_mean contain darker shades of blue which indicates a positive correlation whereas the variables texture_se, smoothness_se, symettry_se have lighter shades indicating negative correlation. Most of the features available are already positively correlated with diagnosis and there are not many inverse relationships found. An increase in our feature magnitude indicates a likely diagnosis.

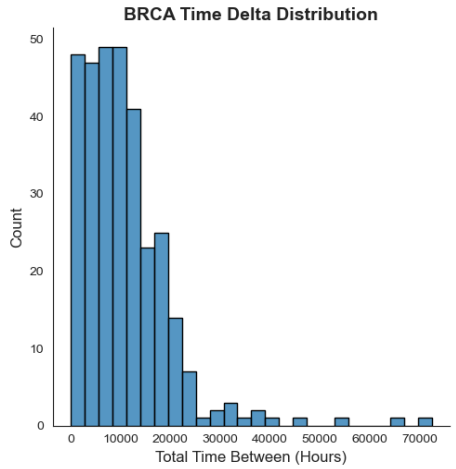


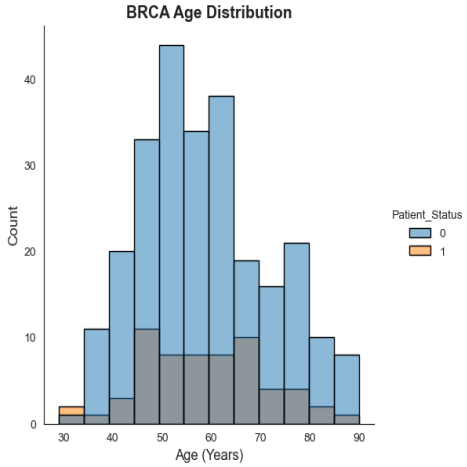
The pairwise plot to the right represents a grid of scatter plots in which variables are plotted against each other. The diagonal charts are histograms that represent the distribution of the individual variables which are crucial for understanding skewness, outliers, and necessary data transformation. Specifically, the chart shows a clear distinction between the malignant tumors (indicate invasive cancer) against those that are benign (indicate non invasive cancer). The malignant tumors show a tighter Gaussian distribution for many of the features where the peak magnitude is higher compared to the benign tumors.

BRCA Cancer Dataset

For this dataset, EDA was performed and presented the findings. This dataset is smaller than the Wisconsin Breast Cancer dataset, but we can still make correleations based on the charts.







The correlation matrix heatmap which shows the variables with the most positive and negative correlation. There are not as many impactful variables to a patient's condition, but we can make the inference that variable protein4, Patient_Status darker shades indicate a more positive correlation than most, meanwhile protein3 would have a slightly more negative correlation opposed to some of the other variables. The histogram represents the time distribution between the patient's visits and surgery. Most patients had a final appointment closely to their scheduled surgery date. As shown, there is a right skewed distribution explaining the higher number of patients having those final appointments. The last Histogram which is representing the distribution of ages where patients with breast cancer survived against those who did not. The percentage of BRCA Cancer patients who died were 19.56% meanwhile Wisconsin Cancer patients who died from having malignant tumors were 37.26%. The minority class has the values which we are aiming to predict within both datasets and the distribution is not a 50/50 split. By employing stratification on the target variable, we can ensure that the machine learning models correctly capture the patterns of those with malignant conditions and those who have deceased.

Modeling

Wisconsin Breast Cancer Dataset

Model A (Logistic Regression)

The preliminary logistic regression model achieved an accuracy rate of 84% and more importantly, due to the danger of missing a true positive, a sensitivity rate of 97%. This current model uses: concave.points_worst, radius_mean, texture_se, and symmetry_se as predictor variables. The AIC of this model is 126.39. This model uses a threshold value of 0.2 to account for the fact that a false negative is more dangerous than a false negative. The confusion matrix for this model is below:

```
y_thresh
     0   1
0  104  26
1    1  38
```

Model B (SVM)

When exploring different values of "C" or margin value for the SVM model, we found that a C value of 1 gives the highest sensitivity (99% on the test data) and the second highest accuracy (97%). Additionally, this low value for the C hyperparameter will reduce the potential for overfitting in the model due to the wider margin. Due to the lower potential for overfitting, all predictor variables were used in the model. The confusion matrix for this SVM model is shown below.

```
cancer_svm_pred
     0   1
0  126   4
1    1  38
```

Model C (Random Forest)

The third model employed for classification is a random forest model. All predictors were included initially because the random forest algorithm, while not exactly a feature selection algorithm, does assess what predictors are useful in establishing a model and will utilize only the most useful. In this regard, a sample of 5 predictor variables were tried at each node. The confusion matrix for the Random Forest model regarding the test data. We can see from the confusion matrix that the accuracy of this model is 97% and the sensitivity for this model is 100% due to it predicting no false negatives.

```
    0  1 class.error
0 124  6  0.04615385
1   0 39  0.00000000
```

BRCA Dataset

Stratification

Through the use of the Synthetic Minority Oversampling Technique, we were able to synthetically add to the BRCA dataset through the employment of the KNN algorithm in order to achieve near equal

representation of the response variable in the dataset. After using SMOTE, positive cases (those who died from cancer) increased from 20% to 55% of the dataset.

Model A (Logistic Regression)

The current iteration of the logistic regression model for the BRCA dataset has an accuracy rate of 56%. The sensitivity rate however is 98%, a very important statistic for predicting a disease. A significant challenge with this model is the potential for overfitting due to the small size of the training set (429 rows). Additionally, as seen in the EDA, most of the predictor variables do not do well at explaining the variability in the dataset. To account for the greater danger of a false negative versus a false negative a threshold value of 0.35 was used. The confusion table for this model is shown below:

```
y_thresh
    0  1
0  12 47
1   1 48
```

Model B (SVM)

When exploring different hyperparameters, a C value of 1 gave the highest accuracy and sensitivity for the model. With this value, we were able to achieve an accuracy rate of 57% and a sensitivity rate of 82%. Once again, due to the low value of C, we reduce the potential for overfitting. Therefore all 11 predictor variables were utilized. The confusion table is shown below:

```
brca_cancer_svm_pred
    0  1
0  22 37
1   9 40
```

Model C (Random Forest)

The random forest model for the BRCA dataset was employed identically to the Wisconsin Cancer dataset. Because of the smaller number of predictors in this dataset, only a sample of 3 predictors was tried at each node. The confusion matrix shows an accuracy rate of 82% and a sensitivity rate of 85%.

```
   0  1 class.error
0 47 12   0.2033898
1  7 42   0.1428571
```

Conclusion/Results

Interactive image processing techniques are a pivotal advancement in the breast cancer detection field. It offers a precise and efficient method to identify tumors. Our goal is to achieve 80+ accuracy and recall if we choose the important variables. We would expect logistic regression to be the final model (if it performs similarly to the other models, it would edge out any small performance gap since it has more explainability to business stakeholders). We are also concerned that Random Forest and SVM will overfit the training data. Our overall goal for the Wisconsin cancer data analysis is to predict whether a patient has cancer, while our goal for the Queen's University Belfast Cancer Research real cancer set is to determine their odds of survival. For the modeling section, correlation and logistic regression for exploratory data analysis was conducted.The correlation matrix illustrates important relationships between the diagnosis as well as several features. The

variables such as concave points_worst, radius_mean, and concave_points_mean feature significant positive correlations, which are apparent from their darker shades of blue. Conversely, texture_se, smoothness_se, and symmetry_se demonstrate lighter shades, indicating negative correlations. It suggests strongly that the majority of features are positively correlated with the diagnosis, with few examples of inverse relationships. Therefore, an increase in feature magnitude denotes a higher likelihood of diagnosis. Pairwise plotting was also conducted which illustrates a clear distinction between the malignant tumors against the benign tumors. The malignant tumors have a tighter Gaussian distribution for many of the features which is indicated by the peak magnitude being higher compared to the benign tumors.

Tree types of modeling: Logistic regression, SVM and Random Forest were conducted for each dataset. Logistic regression for the Wisconsin Breast Cancer Dataset showed an accuracy rate of 84% and a sensitivity rate of 97%. The SVM model for this dataset showed that a C value of 1 gives the highest sensitivity (99% on the test data) and the second highest accuracy (97%). The random forest confusion matrix shows that the accuracy of this model is 97% and the sensitivity for this model is 100% due to it predicting no false negatives. These three models are well refined to identify cancer based on concave.points_worst, radius_mean, texture_se, and symmetry_se as predictor variables. Additionally, for the BRCA dataset for logistic regression an accuracy rate of 56%. The sensitivity rate however is 98%. For the SVM, we were able to achieve an accuracy rate of 57% and a sensitivity rate of 82%. The random forest confusion matrix shows an accuracy rate of 82% and a sensitivity rate of 85%. For these models it is evident that there is overfitting and the model could be improved with better variable selection. Moreover, these techniques continue to revolutionize early detection and diagnosis of breast cancer, improving patient outcomes. Overall, research and innovation in this area are crucial to harness the full potential of interactive image processing techniques to classify and diagnose breast cancer. Accurate results are to be expected with the continuous development of the models.

Sources/Citations

Breast Cancer Wisconsin (Diagnostic) data set. (2016, September 25). Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Queen's University Belfast Cancer Research Real Breast Cancer data set.(2021, March 12). Kaggle

https://www.kaggle.com/datasets/amandam1/breastcancerdataset/data

Street, W. (1993). *Nuclear feature extraction for breast tumor diagnosis*. https://www.semanticscholar.org/paper/Nuclear-feature-extraction-for-breast-tumor-Street-Wolberg/53f0fbb425bc14468eb3bf96b2e1d41ba8087f36

Azzayahia. (2024, January 26). *Breast Cancer Wisconsin (Diagnostic)*. Kaggle. https://www.kaggle.com/code/azzayahia/breast-cancer-wisconsin-diagnostic

Rawanhamdy. (2024, January 27). *breast-cancer-wisconsin f6b4d9*. Kaggle. https://www.kaggle.com/code/rawanhamdy/breast-cancer-wisconsin-f6b4d9

Abhasmalguri. (2024, January 14). *knn-breast-cancer*. Kaggle. https://www.kaggle.com/code/abhasmalguri/knn-breast-cancer

Zhang YN, Xia KR, Li CY, Wei BL, Zhang B. Review of Breast Cancer Pathologigcal Image Processing. Biomed Res Int. 2021 Sep 20;2021:1994764. doi: 10.1155/2021/1994764. PMID: 34595234; PMCID: PMC8478535. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8478535/

Hassan, Md. M., Hassan, Md. M., Yasmin, F., Khan, Md. A. R., Zaman, S., Galibuzzaman, Islam, K. K., & Bairagi, A. K. (2023). A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. *Decision Analytics Journal*, *7*, 100245. https://doi.org/10.1016/j.dajour.2023.100245

Kamel, S. R., YaghoubZadeh, R., & Kheirabadi, M. (2019). Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0247-7

Kumar, S., & Gota, V. (2023). Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation. *Cancer Research, Statistics, and Treatment*, *6*(4), 573. https://doi.org/10.4103/crst.crst_293_23