Classification of Benign vs Malignant Tumors

Group 5 Project Proposal

Prepared by:

Reese Gillan

Ridwan Hossain

Fisher Latham

Samirrah Nyo

Miguel Soto

Submission Date:

02/25/2024

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

TEAM INFORMATION

Team #: 5

Team Members:

1. Ridwan Hossain; rhossain35

   [Background in Biotechnology:Working in a molecular diagnostics lab, BS in Biology and MS in Biotechnology, worked on bioinformatics projects and lab automation programming in the past]

2. Reese Gillian; rgillan6
   [Background in United States Air Force: Integrate airpower and electromagnetic effects to ground forces. BS in Operations Research.]

3. Miguel Soto; msoto40
   [Background in Simulation/Data Science: Current experience revolves around providing consulting services to NAVSEA (Navy) on simulation model development. Most of my previous work has involved working on process improvement and data analytics for defense manufacturing. BS in Industrial and Systems Engineering]

4. Fisher Latham; flatham3
   [Background in Data Science in FinTech: Analytics Engineer building descriptive dashboards and models for business functions. BS in Mathematics and Economics.]

5. Samirrah Nyo; snyo3
   [Background in Cybersecurity: My current work experience is in Governance, Risk, & Compliance in the healthcare industry. BS & MS in Cybersecurity]

OBJECTIVE/PROBLEM

Background Information

Interactive image processing techniques have been used to create features that can be exploited for classification and diagnosis of breast tumors. 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.

Problem Statement

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

State your Primary Research Question (RQ):

What classification method provides the highest accuracy in classifying whether or not a tumor is benign or malignant?

Supporting Research Questions:

1. What classification method provides the highest precision?
2. What classification method provides the highest sensitivity?
3. What classification method provides the highest specificity?

Business Justification:

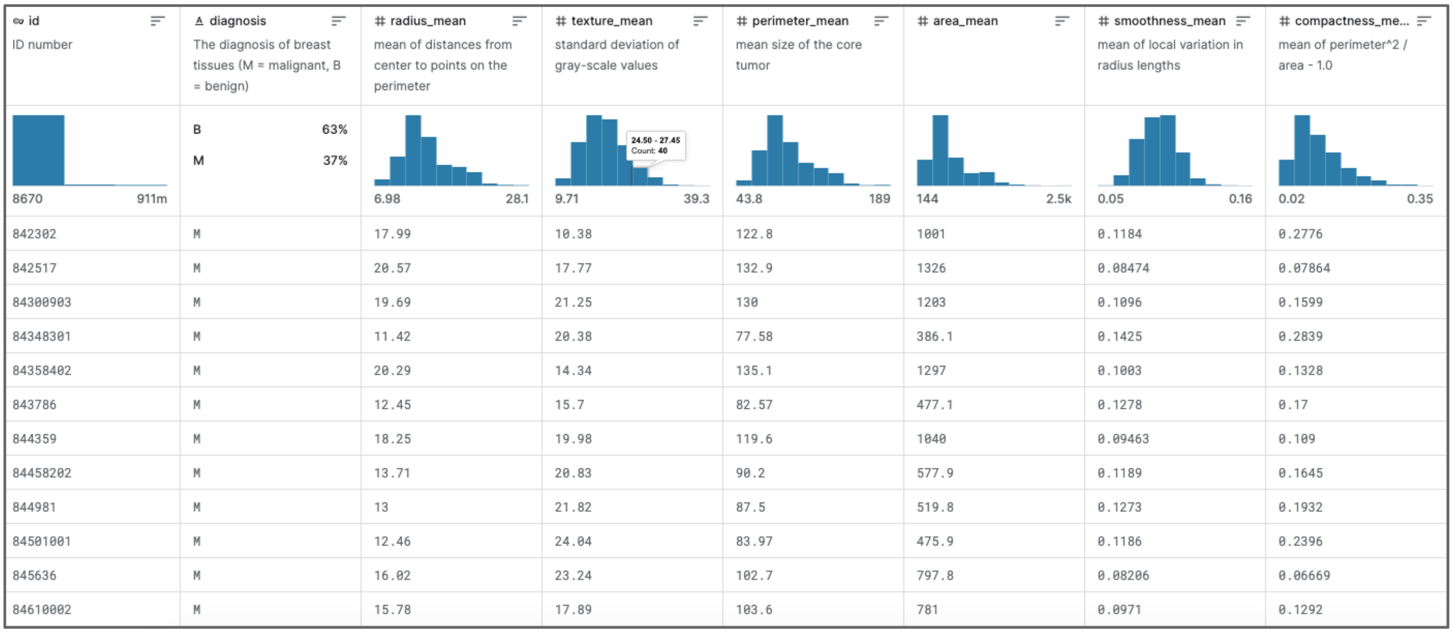
The benefits of solving this problem set are self-evident. If doctors can more accurately and quickly identify whether or not tumors are benign or malignant, they will be able to provide more timely treatment and options to their patients, potentially saving many lives.

DATASET/PLAN FOR DATA

Data Sources:

Breast Cancer Wisconsin (Diagnostic) Data Set

Data Description:



id	diagnosis	# radius_mean	# texture_mean	# perimeter_mean	# area_mean	# smoothness_mean	# compactness_me...
ID number	The diagnosis of breast tissues (M = malignant, B = benign)	mean of distances from center to points on the perimeter	standard deviation of gray-scale values	mean size of the core tumor		mean of local variation in radius lengths	mean of perimeter^2 / area - 1.0
8670 911m	B 63% M 37%	6.98 28.1	24.50 - 27.45 Count: 40 9.71 39.3	43.8 189	144 2.5k	0.05 0.16	0.02 0.35
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864
84300903	M	19.69	21.25	130	1203	0.1096	0.1599
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17
844359	M	18.25	19.98	119.6	1040	0.09463	0.109
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645
844981	M	13	21.82	87.5	519.8	0.1273	0.1932
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292

The dataset we had chosen includes details on Breast cancer from a digitized image. The features describe breast mass geometry specifically for cell nuclei additionally, a diagnosis of malignant or benign is provided describing the disease condition. The various features provided in the dataset can be used as independent or dependent variables that can help lead to insightful analysis for breast cancer.

Key Variables:

Diagnosis could be considered a dependent variable in this case.

Independent variables:

- Texture (standard deviation of gray-scale values) of cell nucleus
- Smoothness (local variation in radius lengths) of cell nucleus
- Compactness (perimeter^2 / area - 1.0) of cell nucleus
- Concavity (severity of concave portions of the contour) of cell nucleus
- Concave points (number of concave portions of the contour) of cell nucleus
- Symmetry of cell nucleus
- Fractal dimension ("coastline approximation" - 1) of cell nucleus

Most important Independent variables:

- Perimeter of cell nucleus: The perimeter could indicate whether cells are malignant or benign.
- Area of cell nucleus: The area could indicate whether cells are malignant or benign.
- Radius (mean of distances from center to points on the perimeter of the cell nucleus): Radius could be higher for malignant cells.

Examples with similar important variables:

1. Breast Cancer Wisconsin (Diagnostic)
2. breast-cancer-wisconsin f6b4d9
3. knn-breast-cancer

APPROACH/METHODOLOGY

Planned Approach

Before diving into the planned approach, it is important to bring attention to the cons of this data set. Since these are anonymous patients, we cannot link external data to learn more about the demographics of this study, which would provide valuable insight into how this model could be used and potentially lead to other significant predictors. The data set itself is also fairly small, with 569 rows, so we do not have a large sample size to work with; this eliminates some of the “data hungry” classification methods like Neural Nets/ Deep Learning. This is because it would be difficult to ensure that our models aren’t overfitting to the data, and since the sample size is small, predicting new patients would probably have low results (bias vs variance trade-off). Recall that the variable we are trying to predict, diagnosis, is binary (either M or B); we can use a variety of classification methods: Logistic Regression (offers the most explainability), Support Vector Machine Classification (could try different kernels), and Random Forest with gradient boosting (might be too data hungry and difficult to explain) are likely the best candidates. Since there are a lot of features, we could try using a high dimension reduction technique like PCA analysis to create new predictors and reduce the number of predictors we use in the model, again this comes with the cost of explainability to stakeholders since this could literally be a life-or-death model.

We would use accuracy (general model performance) and recall since we would want to minimize the amount of false negatives/people who had cancer but were not diagnosed to evaluate and compare our models. Each model would have different hyperparameters, and logistic regression could probably be left alone (outside of C potentially), but SVM and Random Forest definitely require more tinkering to get the best model. A grid search or random search approach would probably be the best way to find the optimal hyperparameters for these more complex models.

Also note that prior to modeling, we would need to transform/scale the data before doing any modeling (except for logistic regression). This is because the range of our numerical data is quite substantial. For example, the smoothness mean variable ranges from 0.05 to .16, whereas area mean ranges from 144 to 2,500. Also, note that all of our data is numeric, so we don't have to create dummy variables for categorical predictors. To transform our data, we could try standardization, normalization, or taking the log of the mean ranges. As of right now, I would lean towards normalization since the distributions of our features are not consistent (some skewed, some normal). Outside of transforming the data, we would also have to consider how to deal with the class imbalance (more negatives than positives) in the data, which will affect how we sample and split our data.

Anticipated Conclusions/Hypothesis:

We will be able 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.

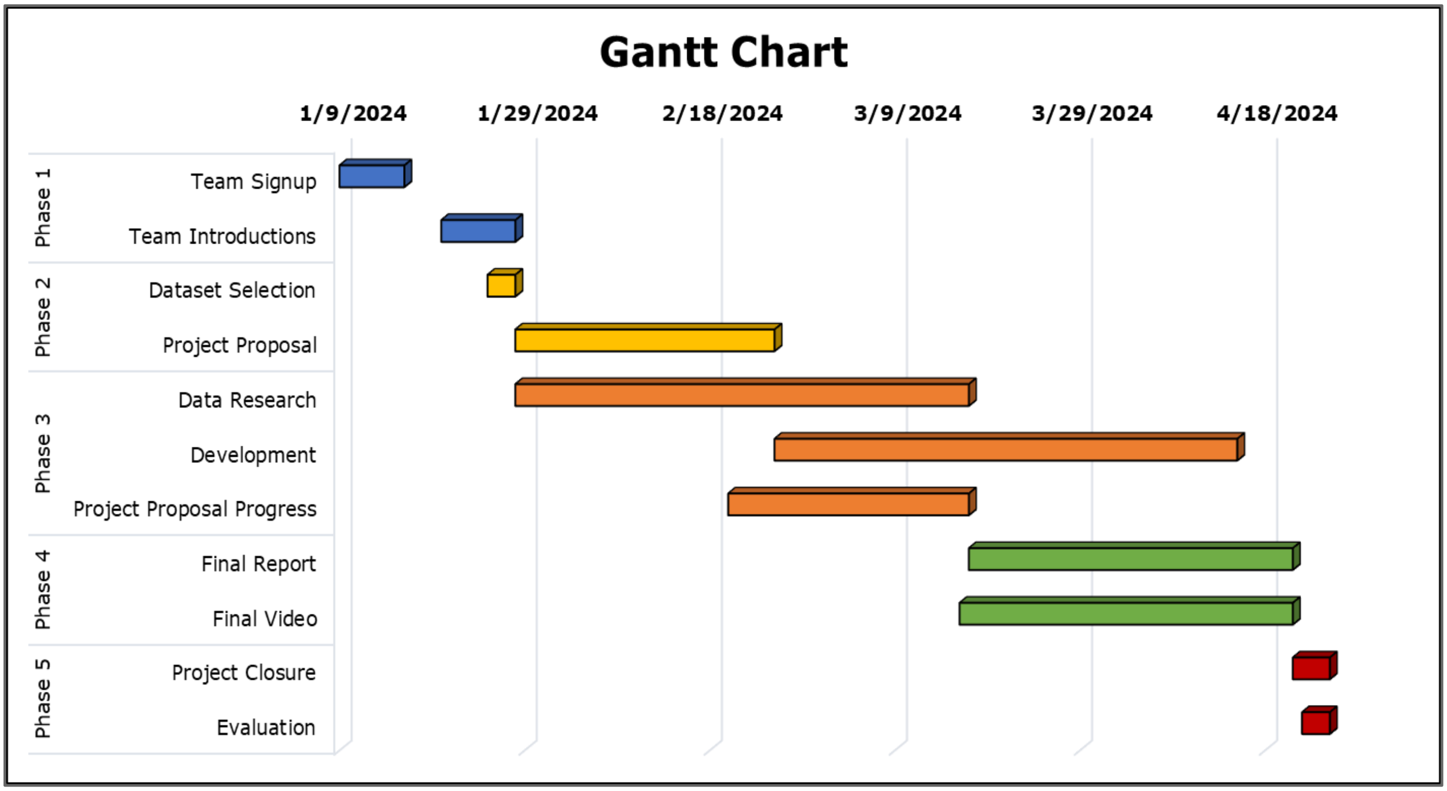
What business decisions will be impacted by the results of your analysis? What could be some benefits?

If we could create a model that could sufficiently predict if a patient had breast cancer or not, then it could be used by doctors to quickly identify if a patient has breast cancer. This could save the patient time (waiting for a diagnosis) and money (perhaps running the model costs less than running machinery). Additionally, since the doctor would be able to identify the disease quicker, the patient could receive proper treatment and care quicker, or if they are negative, peace of mind quicker. However, if they were negative, even with a high-accuracy test, I think it would be worth pursuing a more traditional test to ensure that it's not a false negative (no model is perfect).

PROJECT TIMELINE/PLANNING

Project Timeline/Milestones/Goals:

Gantt charts are widely used in project management for project timelines and milestone achievement. This chart provides a visual representation of which tasks are required for the project and their expected durations. Milestones have been split into phases depending on the project structure and timing.



As shown in the Gantt chart above, Phase 3 will be the phase with the most contributions and added value to the project. The Data Research section will require a sufficient time investment as this task will require investigating the dataset and requiring various analytical approaches to understanding the data. The Development phase will prove to provide the most value added to this project as the group members will leverage programming tools to complete the data models that will provide the optimal solution.

As the project advances, tasks will be distributed based on project requirements and team members' experience, knowledge, and interest. As tasks are created, our goals will adjust depending on what needs to be accomplished. By performing status meetings throughout the project, goals can be identified, measured, and fulfilled through constant communication.

SOURCES

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