Enhancing Breast Tumor Diagnosis: Insights & Implications

Data Science II

COSC 4337

Submitted to

Dr. Ricardo Vilalta

Ly Ha 1920058

Erica Hay 2049545

Khuong Ngo 1857210

In the medical field, traditional tumor diagnosis has heavily relied on full biopsies. While laboratory tests and imaging techniques serve as complementary diagnostic tools, the definitive confirmation for many medical practitioners often requires a biopsy, despite its invasiveness. Nonetheless, the invasive nature and inherent complexities of biopsies emphasize why exploring alternative diagnostic modalities is desirable. Due to the need for advancement in this area, a study was done at the University of Wisconsin in 1992, spearheaded by collaborative efforts between the Departments of Computer Science and Surgery. This study aimed to discern correlations among tumor features and malignancy by using interactive image processing techniques and linear programming-based inductive classification. Building on the insights gained from this research, our application of logistic regression, a supervised learning technique, seeks to enhance tumor diagnostics by identifying correlations among tumor features and malignancy. This approach shows promises to advance the diagnostic process by elucidating malignant features and improving diagnostic accuracy.

Using the dataset from this study, we analyzed the data from 569 images of fine needle aspirations of breast masses in order to examine the characteristics of the cells and contextual features. When taking a look into what the dataset consists of, we are given twelve main features. In the table below, we are able to see each feature, its data type, and its description. The dataset holds three different data types:

1. ‘Int’ types: data that is presented as an integer (number without decimals)
2. ‘Str’ types: data that is presented as a string (used to represent text)
3. ‘Float’ types: data that is presented as a float number (number with decimal points)

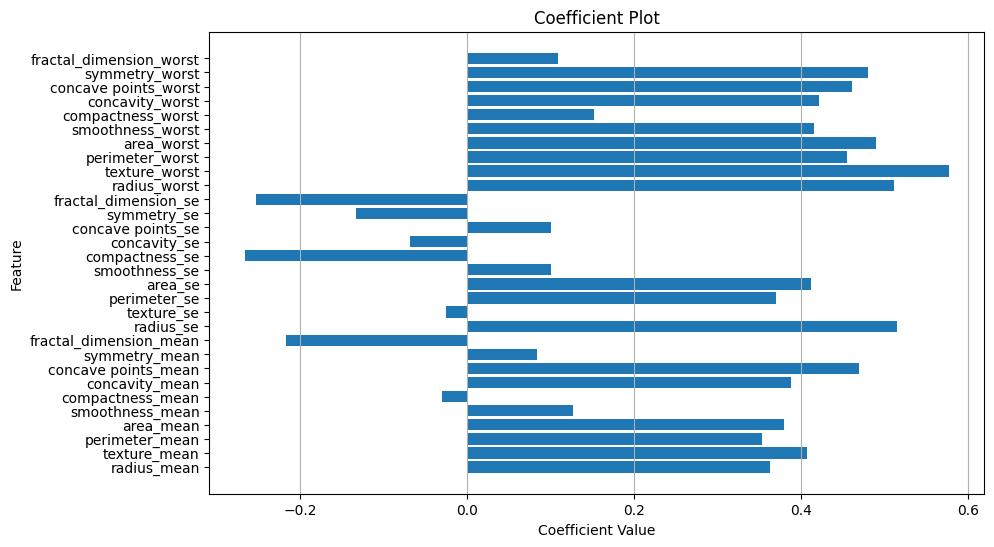
| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | int | Unique identification number for each cell nucleus |
| Diagnosis | str | Malignant (M) or Benign (B) tumor diagnosis |
| Radius | float | Mean distance from the center to points on the perimeter |
| Texture | float | Standard deviation of gray-scale values |
| Perimeter | float | Perimeter of the cell nucleus |
| Area | float | Area of the cell nucleus |
| Smoothness | float | Local variation in radius lengths |
| Compactness | float |  |
| Concavity | float | Severity of concave portions of the contour |
| Symmetry | float | Symmetry of the cell nucleus |
| Fractal Dimension | float |  |

Furthermore, the data set includes important numbers from measurements of each tumor, giving us a ‘mean’, ‘standard error’, and ‘worst’ (the mean of the three largest values) for each of the features in the table. This gives us a comprehensive representation of the distribution of values within each cell feature. The ‘mean’ provides the measure of central tendency. ‘Standard error’ reflects the variability or uncertainty associated with the mean estimate, while ‘worst’ gives us insight on the rand and potential significance of observed values.

Our analysis begins by leveraging logistic regression, a statistical technique, that models the probability of a breast mass being malignant based on these features. Logistic regression helps us predict the likelihood of an event happening. In medical settings, it is often used to predict the probability of a certain outcome based on different factors. In our case, clinicians may need to figure out whether a patient with certain symptoms has breast cancer. Logistic regression helps us analyze how different factors, such as age, gender, or test results, contribute to the likelihood of the patient having that condition. The model calculates coefficients for each symptom or test result, allowing us to identify which features exert the most influence in predicting malignancy. Positive coefficients are indicative of features associated with an increased likelihood of malignancy, while negative coefficients suggest the opposite effect. Additionally, the results of logistic regression models are often presented as odds ratios. An odds ratio greater than 1 suggests that a factor increases the likelihood of the outcome, while an odds ratio less than 1 suggests it decreases the likelihood. Our predictive analysis yields valuable insights into the early detection of breast cancer. Here are the results generated by our model:

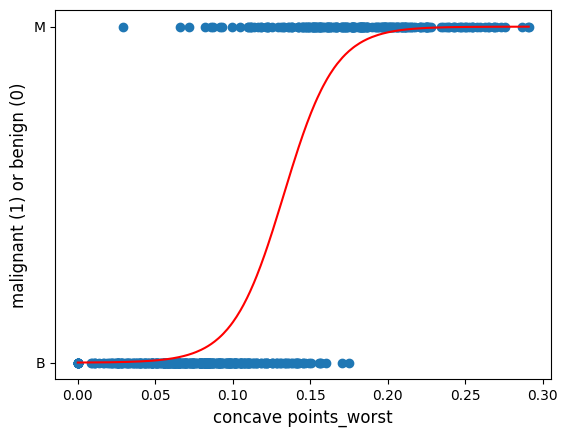
|  | **Coefficient** | **Odds Ratio** |
| --- | --- | --- |
| **Mean Radius** | 0.36365307094544297 | 1.4385750440947966 |
| **Mean Texture** | 0.4074393127528309 | 1.5029642320045415 |
| **Mean Perimeter** | 0.3529376427600434 | 1.423242391199154 |
| **Mean Area** | 0.3794043268232813 | 1.4614138051047514 |
| **Mean Smoothness** | 0.12682831311848605 | 1.1352220983075227 |
| **Mean Compactness** | -0.029350417935999563 | 0.9710761223486043 |
| **Mean Concavity** | 0.38782449793216595 | 1.4737711117124 |
| **Mean Concave Points** | 0.4695575981337264 | 1.5992865093521276 |
| **Mean Symmetry** | 0.08404965184017221 | 1.087682897925519 |
| **Mean Fractal Dimension** | -0.2165399524055209 | 0.8053003605933824 |
| **Standard Error Radius** | 0.515401121391353 | 1.6743099686628509 |
| **Standard Error Texture** | -0.025325117178481602 | 0.9749928735617074 |
| **Standard Error Perimeter** | 0.3702632609240722 | 1.44811579678877 |
| **Standard Error Area** | 0.4127113372849948 | 1.5109088199367307 |
| **Standard Error Smoothness** | 0.10052371499309481 | 1.1057498642435377 |
| **Standard Error Compactness** | -0.26630051330797094 | 0.7662088369478764 |
| **Standard Error Concavity** | -0.0687305907073329 | 0.9335781608330143 |
| **Standard Error Concave Points** | 0.10021742829318461 | 1.1054112396274478 |
| **Standard Error Symmetry** | -0.13301222278020353 | 0.8754543915567699 |
| **Standard Error Fractal Dimension** | -0.25309044898746136 | 0.7763976542660842 |
| **Worst Radius** | 0.512019597615853 | 1.6686578115340118 |
| **Worst Texture** | 0.5777597694624887 | 1.7820417713070849 |
| **Worst Perimeter** | 0.4553482454062391 | 1.57672237376027 |
| **Worst Area** | 0.4897350121871681 | 1.6318837333547729 |
| **Worst Smoothness** | 0.4155369982041489 | 1.5151841734466325 |
| **Worst Compactness** | 0.1528319885011118 | 1.1651292073927302 |
| **Worst Concavity** | 0.4216434214270429 | 1.5244648362593824 |
| **Worst Concave Points** | 0.4608214440780806 | 1.5853757477781627 |
| **Worst Symmetry** | 0.48030003103106744 | 1.6165593474077335 |
| **Worst Fractal Dimension** | 0.10958302194253271 | 1.115812704027996 |

*Figure 1: Coefficients and Odds Ratios of Each Feature*



*Figure 2: Coefficient Plot*

As seen in **Figure 2**, features such as concave points, radius, perimeter, and area emerge as significant predictors of malignancy, with higher values associated with an increased likelihood of a tumor being malignant since they have a coefficient value closer to 1. These findings align with established medical knowledge, where irregularities in cell morphology, size, and shape are indicative of cancerous growth. Conversely, features like smoothness, compactness, and fractal dimension demonstrate a negative correlation with malignancy, suggesting that smoother, more uniform cell characteristics are often indicative of benign tumors. We can also take a closer look into each feature's logistic regression curve, in **Figure 3**, we are able to observe that as the S curve gets closer to 1, or M in this case, the observed cells that exhibit more concave points have a higher correlation to malignancy.

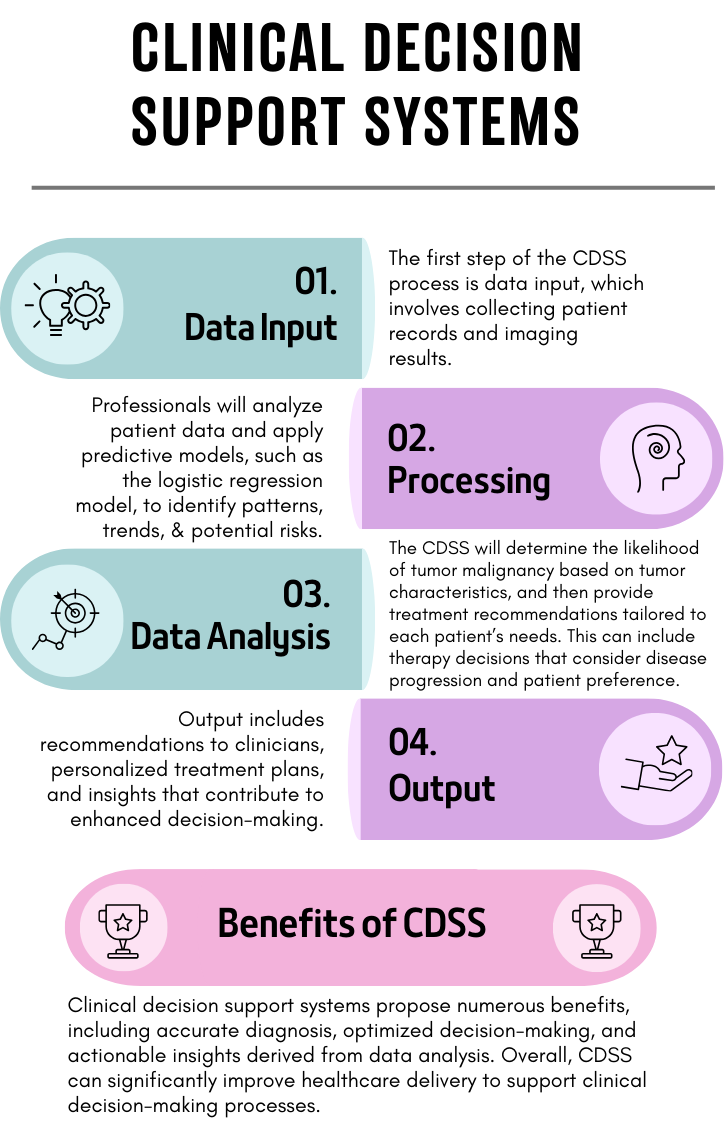


*Figure 3: Logistic Regression curve of Concave Points*

The implications of these predictions for decision-making in clinical practice are profound. Medical professionals can leverage the predictive model’s insights to prioritize diagnostic evaluations for patients with breast masses exhibiting characteristics associated with malignancy. For instance, patients with higher values of concave points and larger radii may warrant further diagnostic procedures, such as biopsies or additional imaging tests, to confirm the presence of cancerous growths. Whereas patients with smoother, more regular cell characteristics may be reassured of the likelihood of a benign tumor, potentially avoiding unnecessary invasive procedures.

The insights gained from our analysis of breast tumor classification has immense potential for transforming clinical practices and healthcare management. Healthcare professionals and organizations can enhance diagnostic accuracy and treatment strategies by leveraging the machine learning techniques and data-driven insights.

A primary application of the findings of our analysis involves the development of clinical decision support systems (CDSS), a type of software that supports the decision-making of clinicians or healthcare professionals using analytical data. Clinical decision support systems can utilize the predictive models used for the breast cancer data set to assist professionals in diagnosing tumors more accurately and efficiently. CDSS can analyze patient data, including imaging results and clinical parameters, to provide recommendations and insights to doctors. The recommendations may include the likelihood of malignancy based on various tumor characteristics, treatment options tailored to each individual patient, and assessments to guide therapy decisions. Integrating machine learning techniques and predictive analytics into clinical workflows can help CDSS enhance decision-making, leading to more accurate diagnosis and personalized treatment plans, improving patient outcomes.



*Figure 4: Example of how CDSS could be applied to tumor classification*

The predictive models used in our analysis can play a crucial role in the development of personalized treatment plans for breast cancer patients. Considering multiple variables that may indicate tumor malignancy, the models can provide professionals with valuable insights into the specifics of each tumor, allowing them to tailor to the specific needs of the patient. Professionals will be able to analyze various tumor features such as size, shape, and cellular characteristics to assess how aggressive and malignant the tumor is. In addition, these models can factor in additional patient information such as age, health status, and genetic conditions. This information can help doctors in determining the best treatment plan for the patient, such as surgery, chemotherapy, or a combination of treatments, while considering medical history and also minimizing potential side effects or complications. The models can also assist in monitoring the effectiveness of treatment plans by analyzing the changes in tumors and patient health, allowing professionals to make adjustments to the treatment plan. Overall, the integration of predictive modeling into treatment planning for breast cancer patients holds promise for improving clinical outcomes while enhancing patient care.

Our analysis of breast tumor diagnosis using machine learning techniques provides valuable insights into the classification of tumors and its implications for clinical decision-making and healthcare management. Through the application of logistic regression modeling on the comprehensive breast cancer dataset, we have achieved strong performance in accurately classifying tumors based on various factors. This predictive capability has promise for improving patient care and treatment strategies, underscoring the potential of machine learning in diagnostic accuracy and decision-making for a wider range of cancer. In summary, our analysis represents a significant step forward in utilizing machine learning for medical diagnoses and improving healthcare practices.