Operation C.P.E.C.T.R.U.M.

A Comparative Analysis of Tumor Attributes and Malignancy

Talha Jaffer (23140003)

Syed Muhammad Waris Shah (23140020)

Muhammad Rayyan (23140018)

Mirza Ibrahim Ahmad (24100171)

Department of Biological Sciences, Lahore University of Management Sciences

BIO5333: Advanced Biostatistics

Dr. Safee Ullah Chaudhary

April 28, 2024

Abstract

Tumor malignancy is a hallmark of cancer and is critical in its diagnosis and subsequent treatment planning. Tumor size and irregularity have been implicated in tumor metastasis, prompting investigation into their potential relationship. This study employs a data-driven approach to explore the hypothesis of a significant association between these factors. The study employs a statistical analysis to test this hypothesis and subsequently train a logistic regression model to assess the predictive power of the features on tumor malignancy. The findings provide insights into the potential associations between tumor characteristics and malignancy.

A Comparative Analysis of Tumor Attributes and Malignancy

Introduction

Tumor malignancy is a hallmark of cancer where the cells display uncontrolled growth, invasion into surrounding tissues, and the potential for metastasis, making it a significant factor in cancer prognosis. Researchers over the years have worked on identifying the factors that contribute to enable the malignancy of these tumors and subsequently making them invasive. It is amongst these that the size of the tumor cells along with their irregularity have been found to have an impact on tumor's invasiveness as well; while several studies have proven the impact each of these factors have on tumor malignancy individually, there is a lack of studies exploring the cumulative relationship that tumor size and tumor irregularity have on its malignancy and prepare a subsequent prediction model.

This paper dives into this very phenomenon with the belief that there is a significant relationship between tumor size and tumor irregularity and their impact on its malignancy. This study uses a systematic examination of these features to establish a data-driven analysis relationship amongst these features in breast cancer.

Hypothesis

- There is a significant relationship between tumour size (mean radius, perimeter, and area) and tumour irregularity (mean concavity, number of concave portions, and fractal dimension) on tumour malignancy.

Methodology

1. **Dataset**

   The dataset for this analysis has been acquired from 'Kaggle' as an excel file where it has measurements from the biopsy features of 569 benign and malignant breast masses.

These features were extracted using computation means from the digital images of the biopsies where the properties of cell nuclei, such as size, shape and regularity were measured in terms of their mean, standard error, worst value of each of 10 nuclear parameters is noted for a total of 30 features. These features have been displayed in Table 1.

Column	Description
x.radius_mean	Mean radius of the tumor cells
x.radius_mean	Mean radius of the tumor cells
x.texture_mean	Mean texture of the tumor cells
x.perimeter_mean	Mean perimeter of the tumor cells
x.area_mean	Mean area of the tumor cells
x.smoothness_mean	Mean smoothness of the tumor cells
x.compactness_mean	Mean compactness of the tumor cells
x.concavity_mean	Mean concavity of the tumor cells
x.concave_points_mean	Mean number of concave portions of the contour of the tumor cells
x.symmetry_mean	Mean symmetry of the tumor cells
x.fractal_dimension_mean	Mean "coastline approximation" of the tumor cells
x.radius_se	Standard error of the radius of the tumor cells

x.texture_se	Standard error of the texture of the tumor cells
x.perimeter_se	Standard error of the perimeter of the tumor cells
x.area_se	Standard error of the area of the tumor cells
x.smoothness_se	Standard error of the smoothness of the tumor cells
x.compactness_se	Standard error of the compactness of the tumor cells
x.concavity_se	Standard error of the concavity of the tumor cells
x.concave_points_se	Standard error of the number of concave portions of the contour of the tumor cells
x.symmetry_se	Standard error of the symmetry of the tumor cells
x.fractal_dimension_se	Standard error of the "coastline approximation" of the tumor cells
x.radius_worst	Worst (largest) radius of the tumor cells
x.texture_worst	Worst (most severe) texture of the tumor cells
x.perimeter_worst	Worst (largest) perimeter of the tumor cells
x.area_worst	Worst (largest) area of the tumor cells
x.smoothness_worst	Worst (most severe) smoothness of the tumor cells
x.compactness_worst	Worst (most severe) compactness of the tumor cells
x.concavity_worst	Worst (most severe) concavity of the tumor cells

x.concave_points_worst	Worst (most severe) number of concave portions of the contour of the tumor cells
x.symmetry_worst	Worst (most severe) symmetry of the tumor cells
x.fractal_dimension_worst	Worst (most severe) "coastline approximation" of the tumor cells
y	target

2. Data Preprocessing

The dataset was processed and adjusted for further analysis in Python using a Jupyter notebook. The data preprocessing was done on Pandas library to make sure that there are no missing values and anomalies in the dataset. All the values that exceeded the 3 standard deviations on either side of the data were considered as anomaly as a norm and therefore dropped from the dataset.

Moreover, the column for malignancy was renamed from ‘y’ to ‘malignant’ for a better understanding of the data. One-hot encoding was also done on this column for easier processing of the data where the benign tumors were labelled as false and malignant ones as true.

3. Exploratory Data Analysis (EDA)

It was then to further understand the distribution of the data that an EDA was performed. The EDA revealed onto us the descriptive values of each of the features which were then visualized for further analysis to check its distribution against the target variable, i.e. the malignancy.

4. Shapiro Wilk Test for Normality

The data was then for its normality to help decide which statistical model to apply. Shapiro Wilk test was used to check for the normality of the data.

5. Principal Component Analysis (PCA)

Considering the multi-dimensionality of each of the target variable, PCA was done to define two major classes; Tumor Size and Tumor Irregularity. Tumor Size is defined by the values of the mean, radius and perimeter of the tumor cells whereas Tumor Irregularity is defined by the values recorded for tumor texture, smoothness, compactness, concavity and symmetry of these cells.

6. Mann Whitney U Test

Since the normality tests suggested that none of the recorded data follows a Gaussian distribution, Mann Whitney U Test was performed to establish the degree of relationship and impact that these two classes have on tumor malignancy.

7. Correlation Analysis

A correlation analysis was also performed to see the impact of each of the feature on the target variable to further validate the results of the Mann Whitney U test.

8. Feature Importance using Random Forests

In addition to the correlation analysis, the feature importance of Random Forests was also used to match the results of the Mann Whitney U test.

9. Machine Learning Model

A logistic regression model was then trained on the available dataset using sci-kit library in python and then evaluated for its performance.

Results

1. **Data Visualization suggests that there is a difference in the distribution of features against the target variable.**

The visualization of the data was done using histograms to assess the distribution of the features across their malignancy. A visual inspection of these histograms suggests that there is a visible difference in the said distribution for almost all of the features.

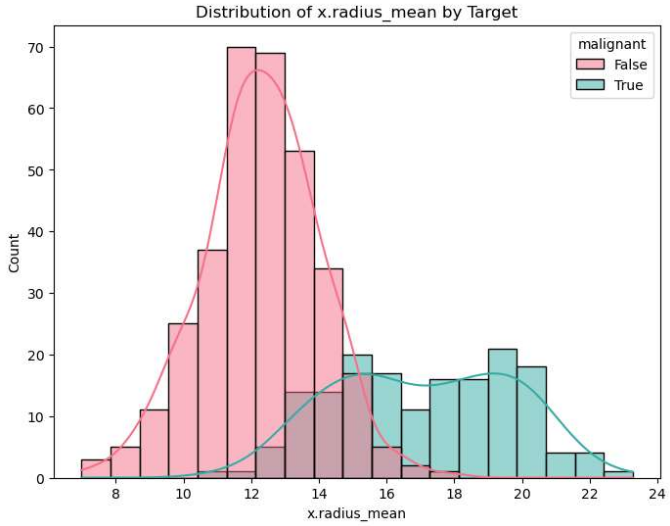

Figure 1. Distribution of Mean Radius Against Malignancy

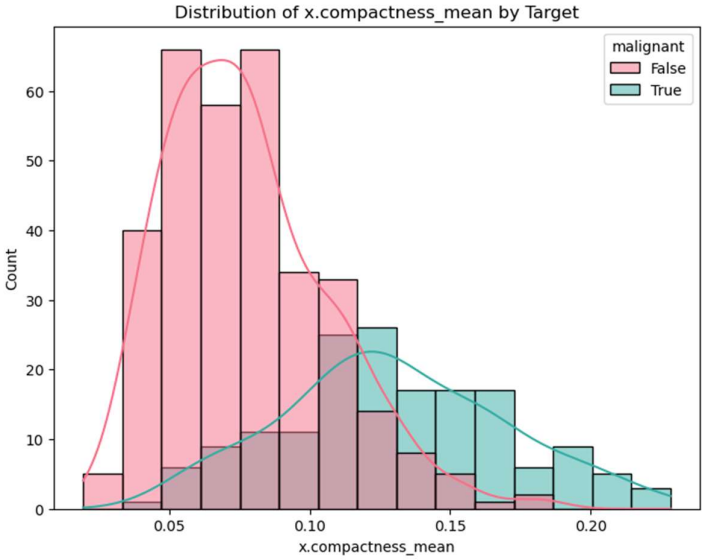

Figure 2 Distribution of Mean Compactness Against Malignancy

2. **The data is not parametric and therefore not Gaussian in nature**

The visual inspection of the histogram hints that the data might be normally distributed. Shapiro Wilk test was therefore done to check the distribution of the data. The test concluded that none of the features in the dataset follow a Gaussian distribution.

3. **There is a significant impact of tumor size and tumor irregularity on its malignancy**

Since the features are not normally distributed, Mann Whitney U test was performed on the PCA classes to examine the significance of this relationship. The test returned a p-value of 9.72e-31 and which suggests that there is a strong impact of these features on the malignancy of the tumor.

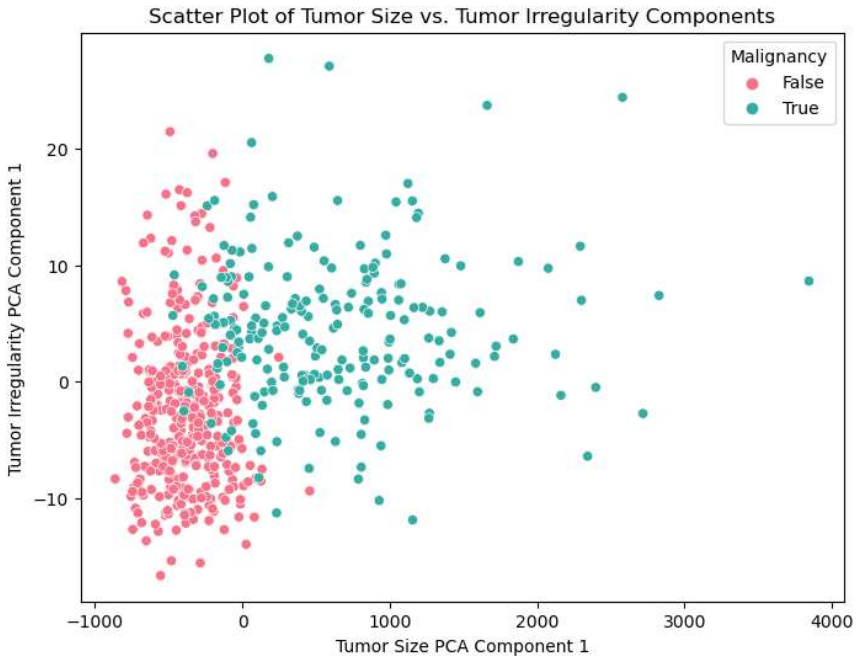

Figure 3 PCA Scatter Plot for Tumor Size vs Tumor Irregularity

Random Forests and Correlation Analysis were also performed on this data to examine this relationship and further validate the results of the Mann Whitney U test. The results of

these tests also provide a similar result where the features of perimeter, concavity, radius and area have been found to have a strong relationship with the malignancy of the tumor.

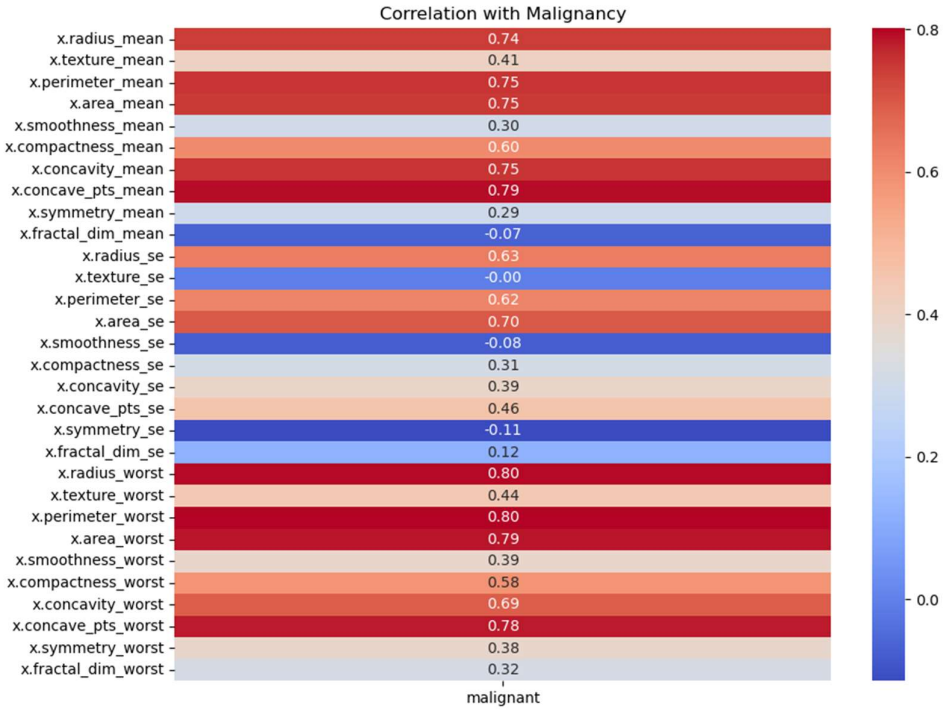

Figure 4 Heatmap for Correlation of each feature against Malignancy

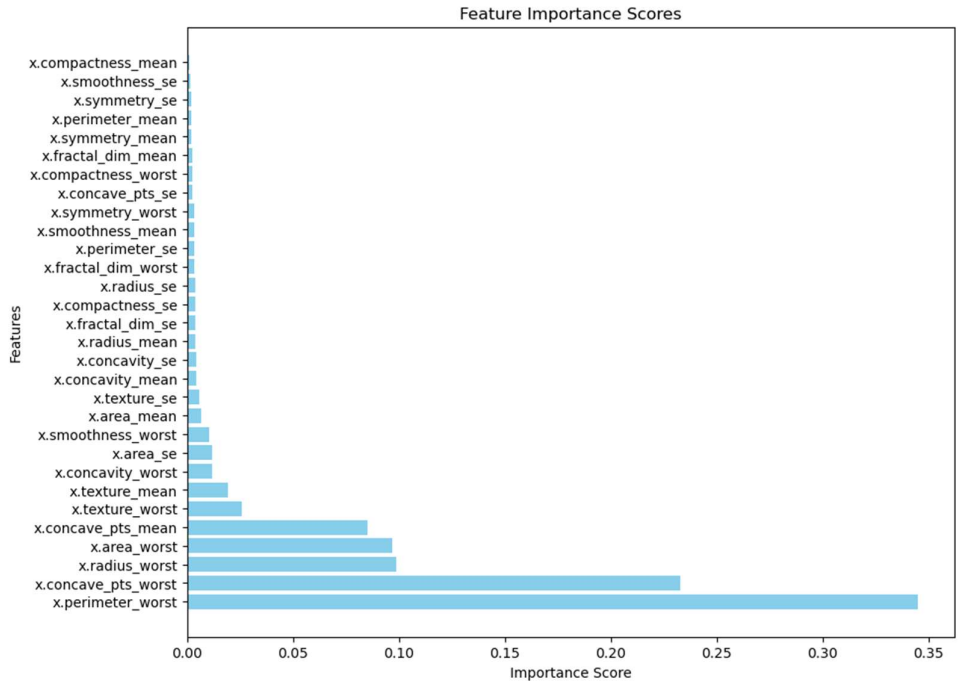

Figure 5 Feature Importance Score of Each Feature against Tumor Malignancy

4. Logistic Regression Model classifies the tumor malignancy with accuracy.

The logistic regression model was trained on the subsequent dataset and used to predict the behaviour of the tumor based on its features. The model returned an F1 score of 0.937 which shows that this model is capable of classifying the data using the given variables with great precision and accuracy.

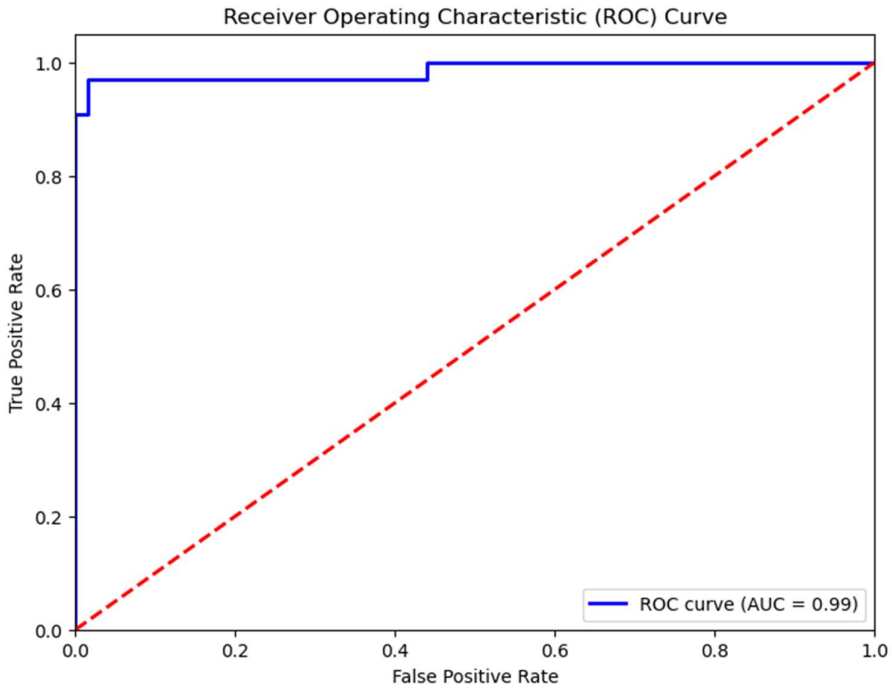

Figure 6 ROC Curve for the Accuracy of ML Model

Conclusion

The findings of this study support the hypothesis that tumor size and irregularity have a significant impact on tumor malignancy. Our analysis revealed notable differences in the distribution of features between benign and malignant tumors, indicating potential associations between these features and malignancy.

These findings underscore the importance of considering these attributes in cancer diagnosis and treatment planning, potentially enabling more accurate and personalized therapeutic strategies. However, it is essential to acknowledge the limitations of this study, including the retrospective nature of the dataset and potential confounding variables not accounted for in the analysis.

References

Baba AI, Câtoi C. Comparative Oncology. Bucharest (RO): The Publishing House of the Romanian Academy; 2007. Chapter 3, TUMOR CELL MORPHOLOGY. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9553/

Donald M. McDonald, Peter Baluk; Significance of Blood Vessel Leakiness in Cancer1. Cancer Res 15 September 2002; 62 (18): 5381–5385.

Fabian Spill, Daniel S Reynolds, Roger D Kamm, Muhammad H Zaman, Impact of the physical microenvironment on tumor progression and metastasis, Current Opinion in Biotechnology, Volume 40, 2016, Pages 41-48, ISSN 0958-1669, https://doi.org/10.1016/j.copbio.2016.02.007.

Ralph M. Böhmer, George Morstyn; Uptake of Hematoporphyrin Derivative by Normal and Malignant Cells: Effect of Serum, pH, Temperature, and Cell Size. Cancer Res 1 November 1985; 45 (11_Part_1): 5328–5334.

Singh, U. (2023) Breast Cancer Wisconsin Diagnostic dataset, Kaggle. Available at: https://www.kaggle.com/datasets/utkarshx27/breast-cancer-wisconsin-diagnostic-dataset (Accessed: 28 April 2024).

Zeng, Y. et al. (2022) ‘Guiding irregular nuclear morphology on Nanopillar arrays for malignancy differentiation in tumor cells’, Nano Letters, 22(18), pp. 7724–7733. doi:10.1021/acs.nanolett.2c01849.