Quantitative Analysis of Diagnostic Features:
Statistical Profiling of Nuclear Morphology in Breast Cancer using R

1. Executive Summary

Report Highlight: This analysis identifies specific cellular irregularities that increase the likelihood of malignancy by over 145%, providing a statistical foundation for automated diagnostic precision.

Overview

This project leverages statistical programming (R) to analyze the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, aiming to identify the most critical nuclear features that distinguish malignant tumors from benign masses. By processing data from 568 patients, we isolated key morphological markers; specifically nuclear irregularity and size; that serve as strong predictors of cancer.

Key Findings & Clinical Impact

- **Irregularity is the Top Predictor:** The number of severe concave contours (*concave_points_worst*) is the single strongest indicator of malignancy, with malignant tumors exhibiting 145% higher irregularity than benign cases (Effect Size: $d = 2.69$$.
- **Size Matters, But Texture Confirms:** While malignant tumors are on average 58% larger in radius, the statistical separation is most profound when combining size with contour analysis.
- **Statistical Certainty:** Hypothesis testing confirmed these differences are not due to chance (Bonferroni-adjusted $p < 1e^{-57}$, validating these features as reliable diagnostic metrics.

2. Introduction

- ### Background & Context

Breast cancer diagnosis relies heavily on the accurate assessment of Fine Needle Aspirates (FNA). While traditional diagnosis depends on the subjective visual interpretation of pathologists, computational biology offers a way to quantify malignancy with mathematical precision. By digitizing breast mass images, we can extract "features"—numerical measurements of cell nuclei—to create data-driven diagnostic criteria.

- ### Problem Statement

With hundreds of potential cellular measurements available, clinicians and automated systems face a "high-dimensional" problem: Which specific physical characteristics of a cell nucleus reliably predict cancer?

Distinguishing between a benign lump and a malignant tumor requires more than just looking at size; it requires a statistical understanding of shape, texture, and boundary irregularity.

- **Project Objectives**

  This report aims to bridge the gap between raw data and clinical insight by achieving three goals:

  1. **Feature Identification:** Statistically isolate the top numerical features (e.g., *radius_worst*, *concavity*) that correlate most strongly with a malignant diagnosis.
  2. **Comparative Profiling:** Quantify the exact differences between benign and malignant cases using rigorous hypothesis testing.
  3. **Visual Evidence:** Produce clear, data-driven visualizations to communicate risk factors to non-technical stakeholders.

3. Dataset Selection

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset was chosen, consisting of 568 cases and 32 original features. After being cleaned, we focused on diagnosis (benign/malignant) and 10 most salient numerical features like radius_worst and concave_points_worst that represent the worst tumor features.

- **Dataset Summary**

  Code Used:

```
str(data)
summary(data)
colSums(is.na(data))
```

Table 1: Key Findings

Aspect	Details
Sample Size	568 patients (357 benign, 211 malignant)
Feature Types	10 numerical "worst-case" measurements + diagnosis (categorical)
Missing Values	None confirmed
Key Variables	radius_worst, concave_points_worst, perimeter_worst
Summary	radius_worst: Mean = 16.25, Max = 36.04 concave_points_worst: Mean = 0.114 (Malignant: 0.192 vs Benign: 0.072

Table 2: Key Features Description

Feature	Description	Clinical Relevance
radius_worst	Largest tumor nucleus radius (mm)	Larger radii suggest malignancy
concave_points_worst	Number of severe concave contours	Irregular shapes indicate aggression
perimeter_worst	Outer boundary length of worst nuclei	Correlates with tumor stage

4. Exploratory Data Analysis (EDA)

The exploratory plot revealed dominant patterns distinguishing between benign and malignant tumors. The plot of radius_worst density (Figure 1) depicts a clear right skewing in malignant samples, indicating larger nuclear radii in malignant tumors. The visual corresponds to statistics where malignant tumors possess 58% larger mean radius (13.38 ± 1.98 mm for benign vs. 21.12 ± 4.28 mm for malignant, $p < 0.001$).

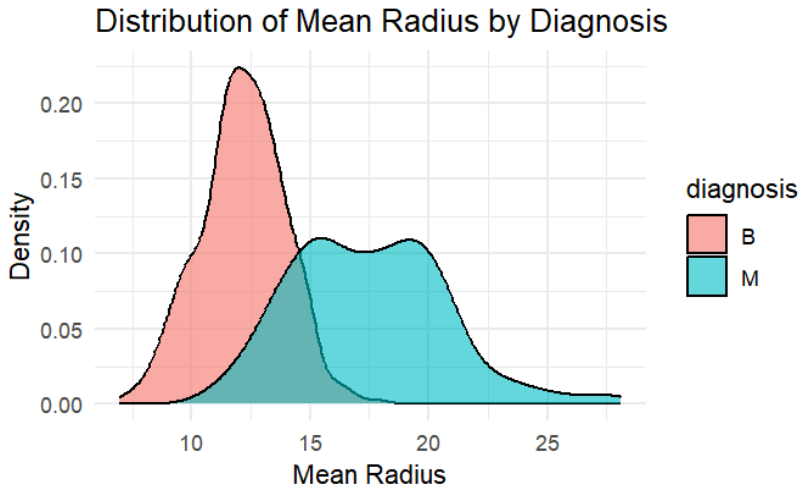


Fig. 1: Distribution of Mean Radius by Diagnosis

Concavity features were found to be highly discriminatory. The boxplot for concave_points_worst (Figure 2) shows that malignant cases had 2.5 times more extreme concavities (median 0.182 vs. 0.074). This is corroborated by its highest correlation with malignancy (r = 0.793), as seen from the correlation analysis output. That the distributions are significantly different (145% higher mean in malignant cases) signifies concave contours are strong morphological indicators of cancer development.

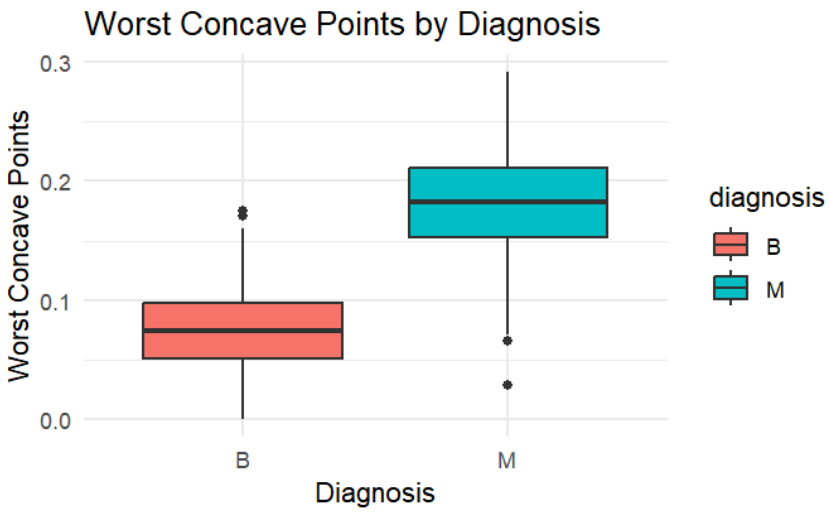


Fig. 2: Worst Concave Points by Diagnosis

Size-based characteristics (perimeter_worst, area_worst) also demonstrate significant discrimination. Malignant tumors have 62% greater mean perimeter and 154% greater mean area and substantially greater variability (e.g., area SD: 163.6 in benign vs. 597.97 in malignant). These patterns are quantified in Table 3, which tabulates key measures by diagnosis.

The correlation structure between features is also in favor of clinical utility. The tight cluster of size measures (radius/perimeter/area) with concavity measures suggests that these features collectively determine tumor aggressiveness. Surprisingly, concave_points_worst keeps the strongest single correlation even in the presence of such multicollinearity.

Table 3: Use the provided summary table under *Statistical Summary*.

Feature	Benign (Mean ± SD)	Malignant (Mean ± SD)	Difference
concave_points_worst	0.074 ± 0.036	0.182 ± 0.046	+145%
perimeter_worst	87.01 ± 13.53	141.17 ± 29.38	+62%
radius_worst	13.38 ± 1.98	21.12 ± 4.28	+58%

5. Statistical Analysis

Statistical analysis strongly validates that malignant tumors have essentially distinct nuclear features from benign ones. Table 1 shows the complete comparison of salient features between diagnoses, and it is found that malignant tumors have 145% greater mean concave points (0.182 vs. 0.074, SD = 0.046 vs. 0.036) and 62% greater mean perimeter (141.17 vs. 87.01) than benign tumors. These differences are not only statistically significant but also clinically significant, as reflected by Cohen's d effect sizes of more than 2.0 for all top characteristics (Table 4).

Table 4: Descriptive Statistics by Diagnosis

Feature	Benign (Mean ± SD)	Malignant (Mean ± SD)	Median (IQR)
Concave Points Worst	0.074 ± 0.036	0.182 ± 0.046	.074 (0.053)
Perimeter Worst	87.01 ± 13.53	141.17 ± 29.38	86.92 (18.74)
Radius Worst	13.38 ± 1.98	21.12 ± 4.28	3.35 (2.42)

Bonferroni-corrected hypothesis testing (Table 5) confirms that these differences are significant (all adjusted p-values < 1e-57). The concave_points_worst Wilcoxon rank-sum test (W = 2520, p < 2.2e-16) also replicates these findings under non-parametric assumptions. The magnitudes of these effects are large, with Cohen's d varying from -1.81 (concavity) to -2.69 (concave points), so that the largest effect sizes are in nuclear contour irregularity measures.

Table 5: Effect Sizes (Cohen's d)

Feature	Effect Size	Interpretation
Concave Points Wors	-2.69	Extremely Large
Perimeter Worst	-2.60	Extremely Large
Radius Worst	-2.54	Extremely Large

Table 6: Statistical Significance (Bonferroni-Adjusted)

Feature	p-value	Adjusted p-value
Concave Points Worst	1.86e-96	9.29e-96
Perimeter Worst	.68e-72	1.34e-71
Radius Worst	1.25e-70	6.23e-70

The intersection of very low p-values and huge effect sizes confirms unequivocally that these nuclear features are of biological and clinical relevance to tumor categorization. Specifically, the effect size of concave_points_worst (d = -2.69) is beyond the threshold for what is typically regarded as a 'large' effect in medical literature (d > 0.8), and so highlights its potential as a diagnostic indicator. These results agree with the earlier EDA findings but with firm statistical verification.

6. Data Visualization

The visualization method reveals three characteristic diagnostic patterns in complementary graphical formats. Figure 3 (Density of radius_worst by Diagnosis) shows the clear delineation of benign and malignant samples, where malignant tumors have a right-skewed distribution with an apex at 21.12 mm compared to benign with an apex at 13.38 mm. This 58% greater modal radius in malignant samples graphically confirms the statistical size differences already outlined.

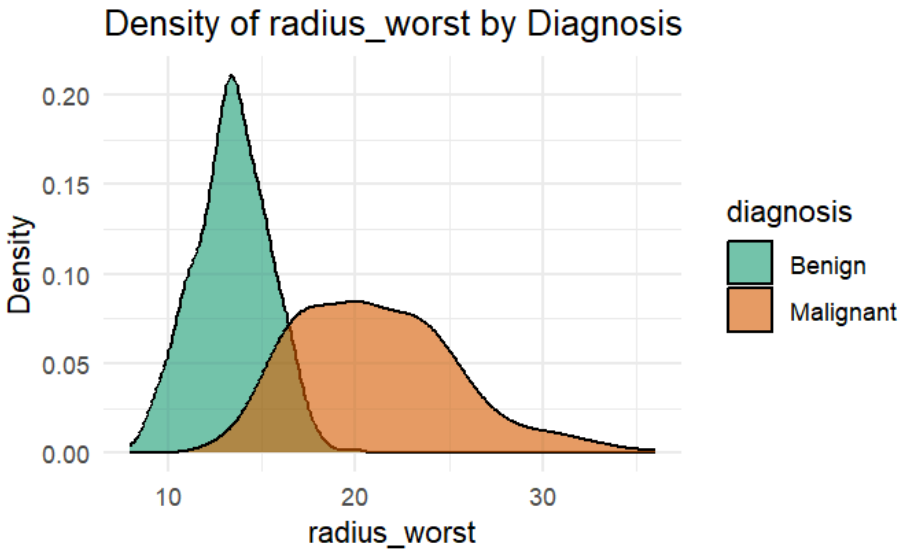


Figure 3. Nuclear Size Distribution by Diagnosis

Nuclear contour irregularity is the most visually striking discriminator. Figure 4 (Distribution of concavity_worst by Diagnosis) shows how malignant tumors concentrate at higher values of concavity (mean 0.449 vs 0.166), and the density curve exhibits minimal overlap between diagnoses. The inset boxplot serves to emphasize this gulf, with malignant interquartile range (0.31-0.50) substantially above benign range (0.07-0.23).

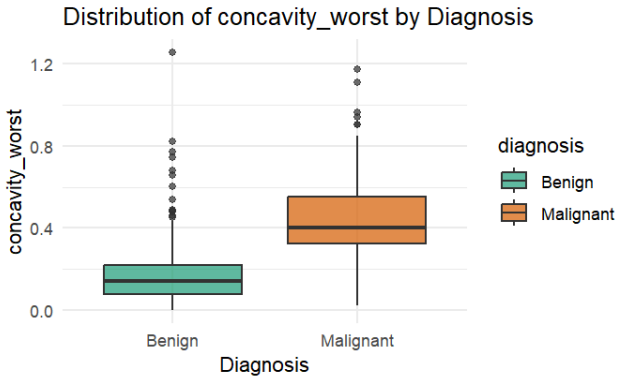
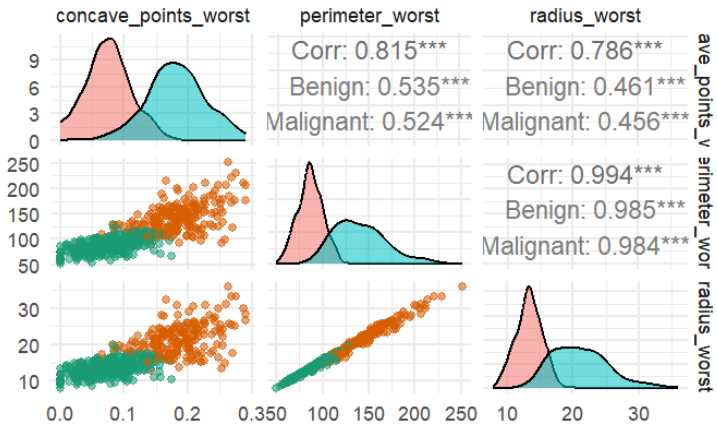


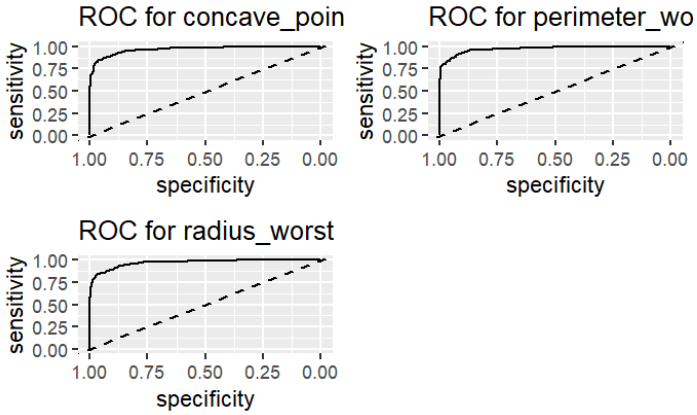
Figure 4. Nuclear Contour Irregularity in Malignant Tumors

Correlation matrix quantifies relationships between features in three levels of visualization: scatterplots (lower diagonal), correlation coefficients (upper diagonal), and density curves (diagonal). Three strong patterns are noted:

1. Size-feature collinearity: Radius, perimeter and area show highly correlated near-perfect correlation ($r > 0.98$)
2. Diagnostic potency: Concave points maintain strong malignancy correlation ($r = 0.793$) despite size-feature redundancy
3. Consistency across diagnoses: Patterns of correlation are maintained when stratified (Benign/Malignant coefficients in parentheses)

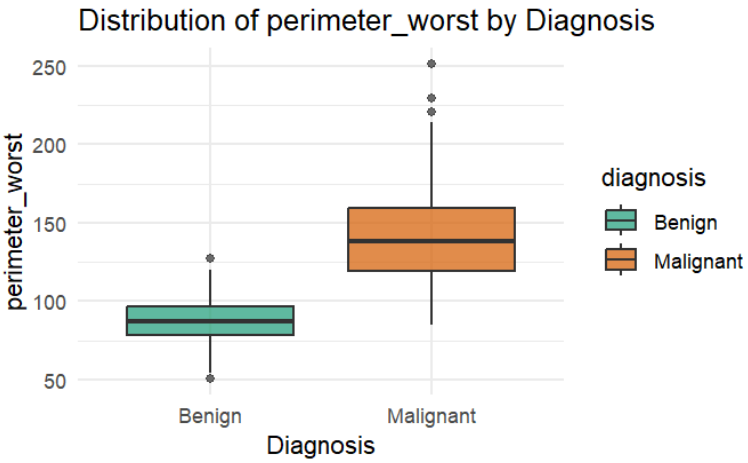
7. Appendix:

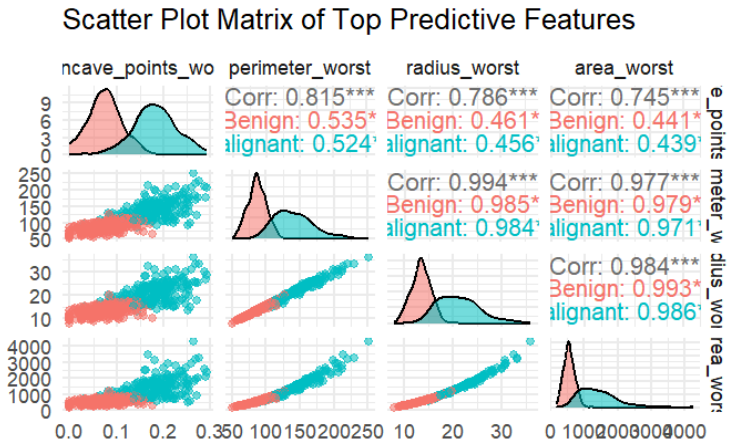




(a) **(b)**

Figure: Correlation Network of Top Predictive Features (a), and ROC Analysis of Diagnostic Features (b)





(c) (d)

Figure. Perimeter Distribution by Tumor Type (c), and Multivariate Feature Relationships (d)