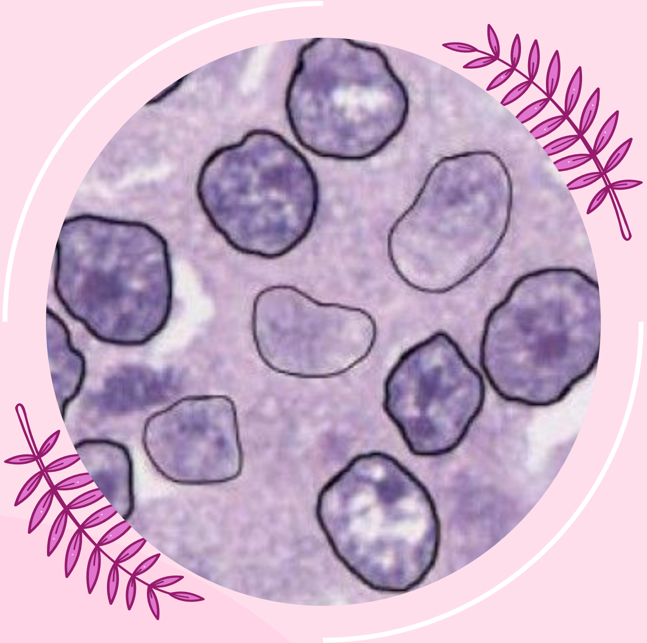
Breast Cancer Wisconsin Data Analysis

CellSight Diagnostics

The Diagnostic Challenge

- Our client is a radiology network exploring computer-aided diagnosis (CAD) systems for breast cancer screening
- We had received cell nuclei imaging data from fine needle aspirate (FNA) samples.
- Our team worked to identify which cellular characteristics best distinguish malignant from benign tumors to create a machine learning model that accurately predicts whether a tumor is cancerous



Why Our Mission Matters

- Breast cancer is the 2nd most common cancer and 2nd leading cause of cancer death among women in the U.S
- It is the leading cause of cancer death for black and hispanic women
- 13% of U.S women will develop invasive breast cancer in their lifetime
- Early detection at the localized stage could yield a >99% 5-year survival rate

1 in 8 women in the United States will develop breast cancer in her lifetime.

Dataset Description

Wisconsin Diagnostic Breast Cancer Dataset (WDBC)

- **Source**: UCI ML Repository / Kaggle
- **File**: data.csv
- **Rows / Columns**: 569 rows, 33 columns

Target Variable:

- **Diagnosis (Categorical)**
  - M = Malignant
    B = Benign
- **Purpose**:
  - Binary classification to predict if a tumor is malignant or benign.

Features:

- 30 numeric tumor measurements: **Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry , fractal dimension.**

- Each of the 10 attributes is recorded in all three groups, giving 30 total numeric features

- Measurement Groups: **Mean, Standard Error, Worst values**

- One empty column: **Unnamed: 32**

Data Dictionary

Column	Description	Feature Type	Valid Values/Range	Notes/Issues
id	Unique patient/sample identifier	Identifier	8670-911320502 integers	Variable length, inconsistent formatting
diagnosis	Tumor classification	Binary Categorical	M (malignant), B (benign)	Target variable
radius_mean	**Mean distance from center to nucleus perimeter**	**Continuous**	**6.98 - 28.11**	**Larger = more likely malignant**
texture_mean	Mean gray-scale variation in nucleus	Continuous	9.71-39.28	Higher = more irregular surface
perimeter_mean	Mean nucleus perimeter	Continuous	43.79 - 188.50	Correlated with radius
area_mean	**Mean area of cell nucleus**	**Continuous**	**143.50 - 2501.00**	**Malignant cells typically larger**
smoothness_mean	Mean variation in radius lengths	Continuous	0.05 - 0.16	Lower values indicate smoother borders, higher values indicate irregular borders
compactness_mean	Mean nucleus compactness	Continuous	0.02 - 0.35	0 = perfect circle; higher = more irregular
concavity_mean	**Mean severity of concave contour portions**	**Continuous**	**0.00 - 0.43**	**Higher = more indentations**
concave points_mean	Mean number of concave contour points	Continuous	0.00 - 0.20	Malignant tumors have more concave points
symmetry_mean	**Mean nucleus symmetry**	**Continuous**	**0.11 - 0.30**	**Lower = more symmetric = likely benign**
fractal_dimension_mean	Mean boundary complexity	Continuous	0.05 - 0.10	Higher = more irregular border

Analysis of 10-Point Inspection

Core Findings

- **Dimensions:** 569 biopsy samples across 30 clinical features
- **Data Integrity: 100% complete** features, no missing values or duplicates
- **Target Balance:** Split is **62.7% Benign** vs. **37.3% Malignant** – clinically representative and sufficient for model training

Data Quality Surprises

- **Empty Feature:** ‘**Unnamed: 32**’ column with 100% null values (likely CSV trailing comma error)
- **Class Distribution:** Malignant class is more frequent (~37%) than general population averages (~20%), providing a stronger signal for analysis

Overall: The dataset is high-quality, verified for internal logic, and required minimal cleaning

Tumor Size Categories

```
##Step 1, create the tumor size category column

df['tumor size category'] = pd.cut(
    df['radius mean'],
    bins=[0, 10, 12, 15, 20, float('inf')],
    labels=['Very Small', 'Small', 'Medium',
'Large', 'Very Large'],)

print("Tumor Size Categ
    include lowest=True
ory Counts:")
print(df['tumor_size_category'].value_counts().sort
_index())
```

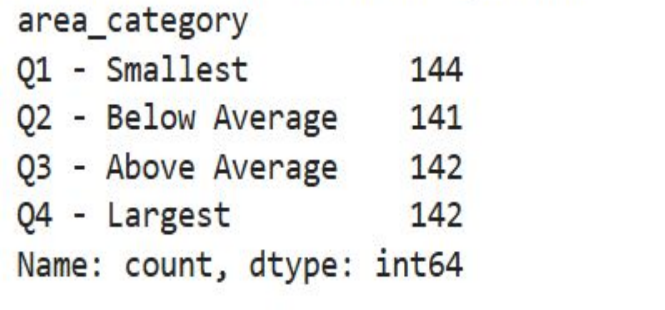
```
Tumor Size Category Counts:
tumor_size_category
Very Small      47
Small          124
Medium         225
Large          128
Very Large      45
Name: count, dtype: int64
```

```
 size_analysis =
df.groupby('tumor_size_category')['diagnosis'].value_counts(
).unstack(fill_value=0)
df.groupby('tumor_size_category')['diagnosis']

size_analysis['Total'] = size_analysis.sum(axis=1)
size_analysis['Malignant_Rate_%'] = (size_analysis['M'] /
size_analysis['Total'] * 100).round(2)

print(size_analysis)
```

```
diagnosis                  B    M  Total  Malignant_Rate_%
tumor_size_category
Very Small                47    0     47              0.00
Small                    118    6    124              4.84
Medium                   180   45    225             20.00
Large                     12  116    128             90.62
  Very Large                0   45     45            100.00
```

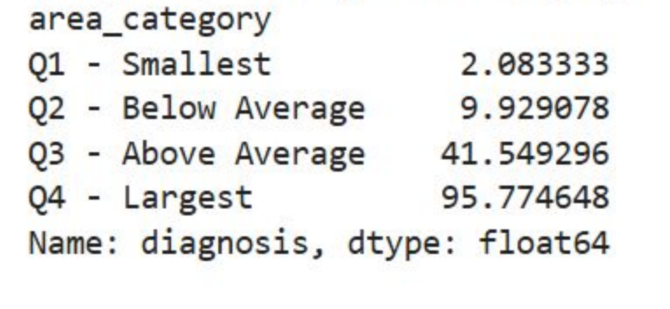
Tumor Area Categories

- Counts per Area Category:



```
area_category
Q1 - Smallest         144
Q2 - Below Average    141
Q3 - Above Average    142
Q4 - Largest          142
Name: count, dtype: int64
```

- Malignancy Rates per Quartile

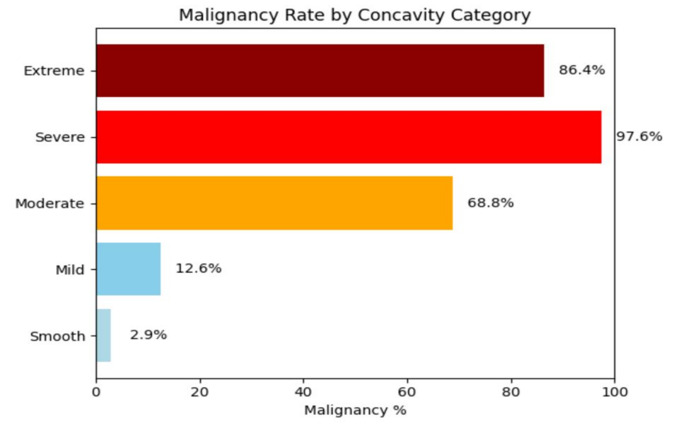


```
area_category
Q1 - Smallest          2.083333
Q2 - Below Average     9.929078
Q3 - Above Average    41.549296
Q4 - Largest          95.774648
Name: diagnosis, dtype: float64
```

- Quartiles for area_mean: 25th Percentile (Q1) - 420.3, 50th Percentile (Median) - 551.1, and 75th Percentile (Q3) - 782.7
- The distribution is nearly equal across the four categories
- There is a dramatic, non-linear increase in the malignancy rate as the tumor area increases
    - Suggests a strong positive correlation between tumor size and malignancy, confirming area_mean as significant predictor for diagnosis

Cell Irregularity Categories

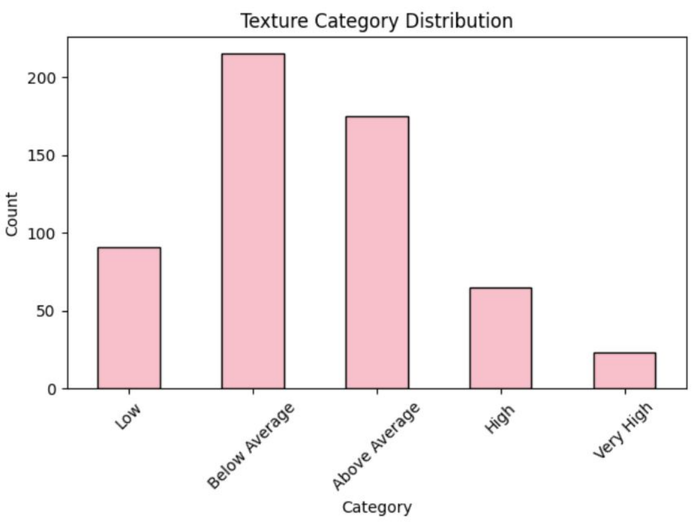
Category *(Intensity of cell indentations)*	**Concavity Mean**
Smooth	< 0.035
Mild	< 0.085
Moderate	< 0.155
Severe	< 0.255
Extreme	0.255 +

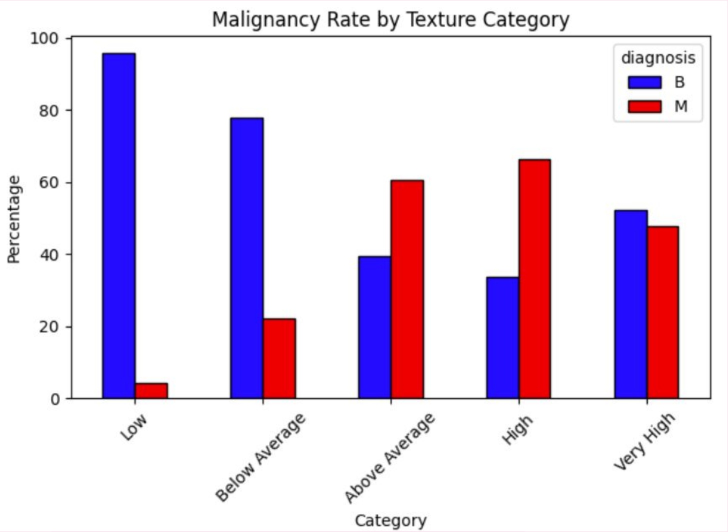


Smooth: 172 cases *(2.91% Malignant)*
Mild: 167 cases *(12.57% Malignant)*
Moderate: 125 cases *(68.80% Malignant)*
Severe: 83 cases *(97.59% Malignant)*
Extreme: 22 cases *(86.36% Malignant)*

Total malignant: 212 *(37.26%)*
Total benign: 357 *(62.74%)*
Total cases: 569

Texture Variability Categories

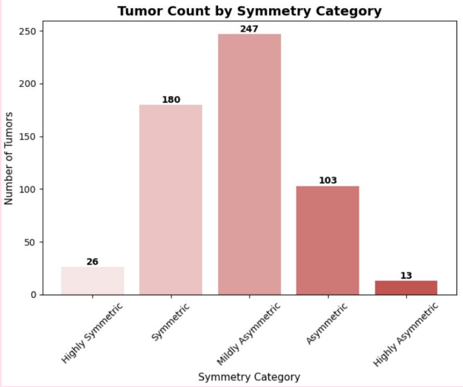


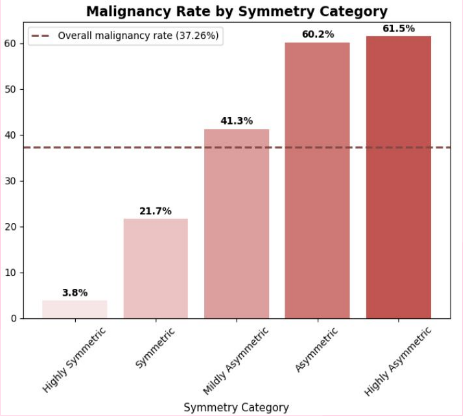


- Texture categories were made based on mean (19.29) and standard deviation (4.30)
- Malignant tumors had a higher average texture_mean (21.60) than benign tumors (17.91)
- Direct relationship between higher texture variability and malignancy

Symmetry-Based Categories

- Asymmetric Cells = Higher Malignancy Risk:
  - Cell division, chromosomal instability, unequal cell division, gene mutations all affect nuclear shape
- Most tumors (43.4%) fall in Mildly Asymmetric
- Malignancy rates rise as asymmetry increases
  - 3.8% -> 61.5%
- Symmetry is a useful diagnostic feature but cannot be used alone





Observations

- Summary of feature analysis results:
  - Very large tumor size = 100% malignancy rate
  - High texture variability = 66.62% malignancy rate
    - Malignant tumors had a higher average texture mean
  - Highly asymmetric = 61.5% malignancy rate
  - Large area category = 95.77% malignancy rate

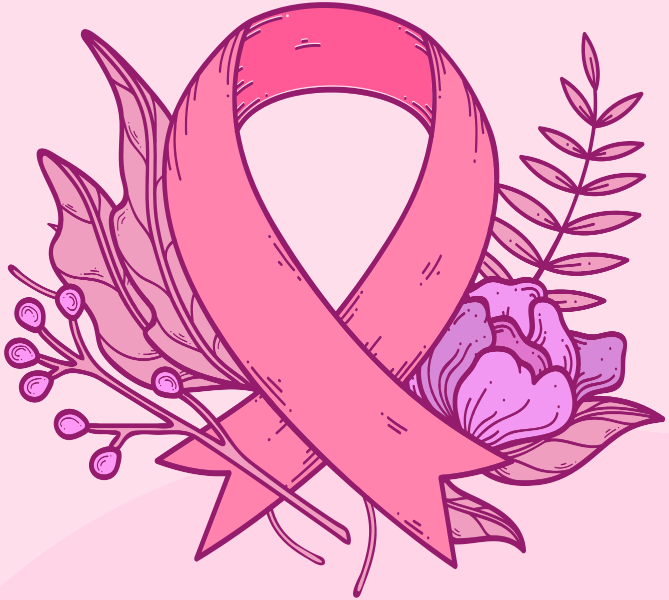
Recommendations & Conclusion

Key Finding:

- Tumor **size** and **area** are the most accurate and strongest predictors of malignancy.
    - Tumors in the **very large** category were **always** malignant
    - Tumors in the **large** category were malignant in more than 95% of cases

What features should the CAD system prioritize? What are the limitations?

- CAD system should prioritize tumor size and area
    - The sample data indicated that those features were the best indicators of tumor malignancy
- This dataset is limited to non-temporal data
    - More data on the progression of tumors may be helpful in predicting the malignancy of cells



Thank You!

ANY QUESTIONS?