Revolutionizing Breast Cancer Diagnosis via *Machine Learning*

PREPARED BY

DERICK CAZARES, MELISSA MORALES,

RENEE PEREZ, AND FEVEN SURAFEL

Breakthrough AI Predicts Future Cancer Risk with Advanced Tools

Project Overview and Inspiration:

MIT researchers improved a machine learning system for cancer risk prediction from mammograms, validated internationally, including test sets from Sweden and Taiwan.

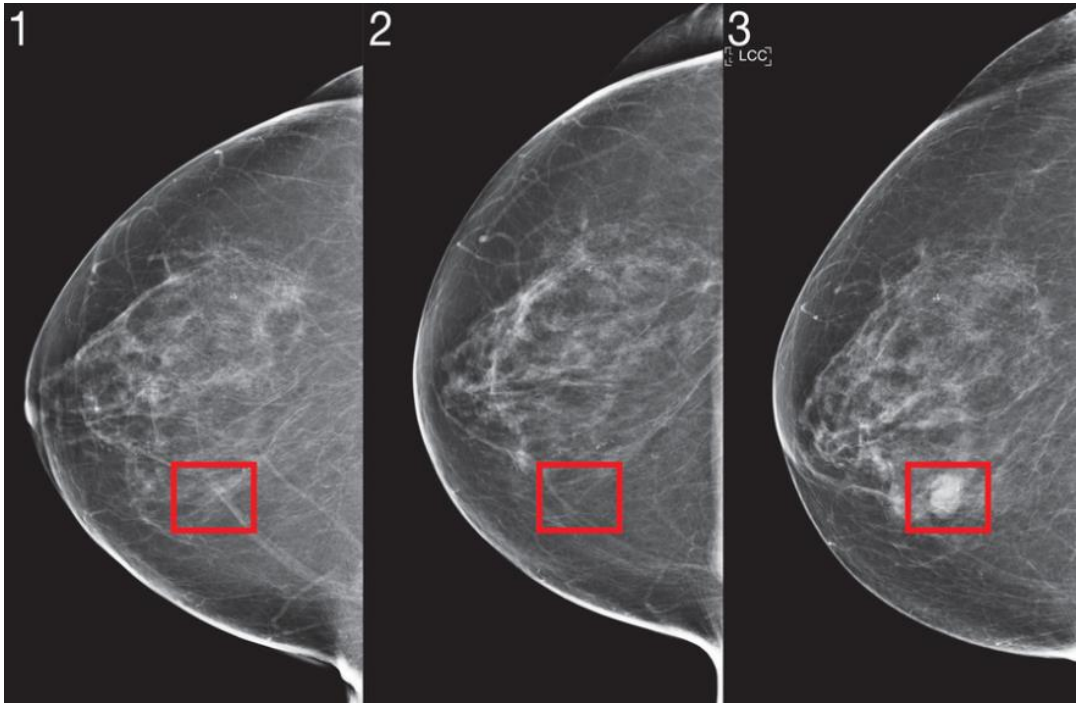
Objective:

Use the Wisconsin Breast Cancer Diagnostic Dataset to develop a machine learning model that classifies tumor samples as benign or malignant, enhancing early breast cancer detection.

Key Deliverable:

Create a model to classify breast cancer biopsy cell nuclei as benign or malignant.

Source: MIT News: Robust AI Tools for Cancer Prediction



Project Summary

Objective

Create a machine learning model to classify breast cancer biopsy cell nuclei as benign or malignant.

Dataset

Wisconsin Breast Cancer Diagnostic Dataset.

ML Models

Logistic Regression and Random Forest

Goal

Develop a model that demonstrates meaningful predictive power at least 75% classification accuracy

Model Development
from Initialization to Evaluation

STEP 1

Retrieve the breast cancer dataset from Spark and launch notebook in Google Colab

STEP 2

Split the data into training and testing sets

STEP 3

Train the models on the training data (X_train and y_train).

STEP 4

Generate and save predictions for the testing data (X_test)

STEP 5

Assess performance with a classification report

STEP 6

Optimize and evaluate the model through iterative changes

Random Forest: First Look at Classification Performance

In this context, the target variable is diagnosis, where a value of B indicates a benign diagnosis and a value of M signifies a malignant diagnosis

- **The model correctly classifies 96% of all samples, indicating high overall performance.**
- **Benign (B)**: The model has high recall (0.99), meaning it identifies most class B samples correctly. Precision is also high (0.98), indicating few false positives.
- **Malignant (M)**: The model shows high precision (0.98) and good recall (0.93), correctly identifying most class M samples with a small number of false positives.

Overall, the random forest model demonstrates strong performance across both classes however we were aiming for a higher accuracy rate. It was time to evaluate the data and review any correlations from the tumor characteristics so we may better train the model..

Random Forest Optimization

Optimization techniques used to try and achieve a higher accuracy rate:

- Visualizations were created to assess the Top 5 and Bottom 6 Features of importance from the tumors
- Top 5 Features were used as predictors. This resulted in a lower overall accuracy rate of 0.95
- Bottom 6 Features were removed to avoid skewing performance. This resulted in same 0.96 accuracy score
- Adjusted the number of estimators in the model and ended up with same accuracy score of 0.96 for 10, 50, 100, 200 and 500 estimators

In summary, the first version of the Random Forest model had the highest overall performance in each category yet we still were aiming for a higher accuracy score, leading us to create a different type of model...

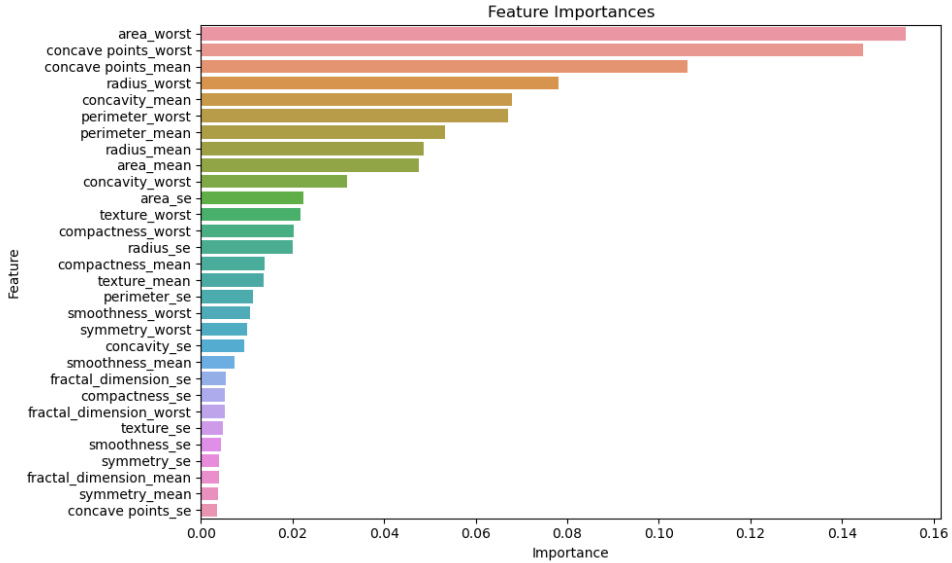
```
feature_importances = rf_model.feature_importances_

#alongside their labels
importance_df = pd.DataFrame({
    'Feature': X.columns,
    'Importance': feature_importances
})

importance_df = importance_df.sort_values(by='Importance', ascending=False)
importance_df
```

	Feature	Importance
23	area_worst	0.153892
27	concave points_worst	0.144663
7	concave points_mean	0.106210
20	radius_worst	0.077987
6	concavity_mean	0.068001
22	perimeter_worst	0.067115
2	perimeter_mean	0.053270
0	radius_mean	0.048703
3	area_mean	0.047555
26	concavity_worst	0.031802
13	area_se	0.022407
21	texture_worst	0.021749
25	compactness_worst	0.020266
10	radius_se	0.020139
5	compactness_mean	0.013944
1	texture_mean	0.013591
12	perimeter_se	0.011303
24	smoothness_worst	0.010644
28	symmetry_worst	0.010120
16	concavity_se	0.009386
4	smoothness_mean	0.007285
19	fractal_dimension_se	0.005321
15	compactness_se	0.005253
29	fractal_dimension_worst	0.005210
11	texture_se	0.004724
14	smoothness_se	0.004271
18	symmetry_se	0.004018
9	fractal_dimension_mean	0.003886
8	symmetry_mean	0.003770
17	concave points_se	0.003513

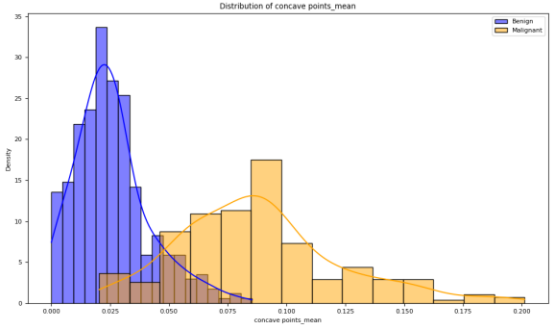
Features Based on Importance

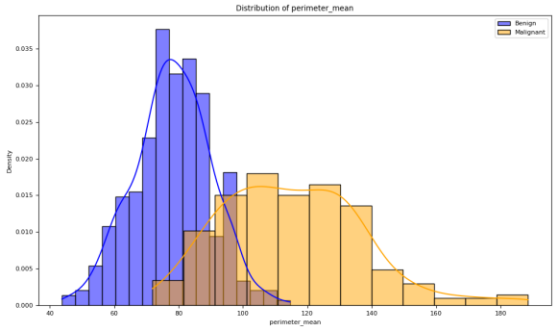


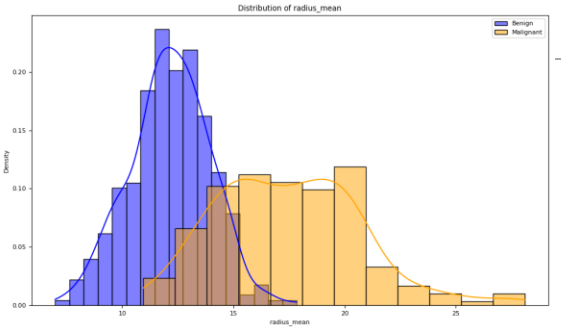
We decided to focus on the top 5 “mean” features to ensure upmost accuracy and avoid the “worst” cases to skew the results

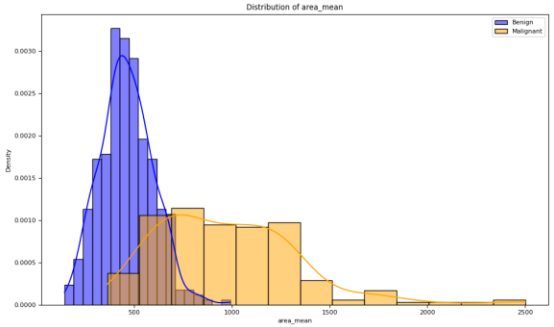
- Concave points_mean
- Perimeter_mean
- Radius_mean
- Area_mean
- Compactness_mean

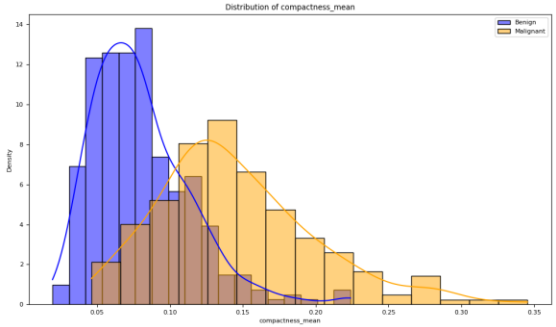
Top Feature Plots



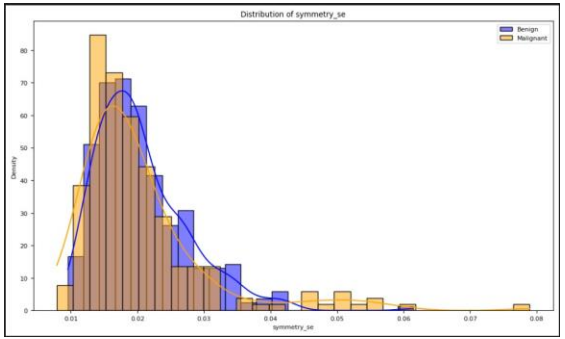








Low importance example:



Boosting Accuracy: Optimized Results with Interaction Terms in Logistic Regression

Interaction terms in logistic regression capture how pairs of features interact by creating new features from their combinations, enhancing the model's ability to understand complex relationships.

- **The model's accuracy improved by 1%, now correctly classifying 97% of cases**
- **Benign (0):** The model achieves a high precision of 0.96 and a perfect recall of 1.00, resulting in an excellent F1-score of 0.98. This means the model is very effective at correctly identifying benign cases and is particularly good at avoiding false positives.
- **Malignant (1):** The model has perfect precision of 1.00 and a recall of 0.92, leading to a strong F1-score of 0.96. This indicates that the model is very effective at identifying malignant cases and has a low rate of false negatives.

Random Forest
Classification Report:

	precision	recall	f1-score
B	0.96	0.99	0.97
M	0.98	0.93	0.95
accuracy			0.96
macro avg	0.97	0.96	0.96
weighted avg	0.97	0.96	0.96

Logistic Regression
Classification Report:

	precision	recall	f1-score
0	0.96	1.00	0.98
1	1.00	0.92	0.96
accuracy			0.97
macro avg	0.98	0.96	0.97
weighted avg	0.97	0.97	0.97

Next Steps in Advancing Breast Cancer Classification with Machine Learning

This project developed a machine learning model for classifying breast cancer biopsy samples as benign or malignant, using the Wisconsin Breast Cancer Diagnostic Dataset.

The Random Forest model outperformed the logistic regression model in recall for malignant cases, a crucial factor when addressing high-risk patients

Our team recommends the Logistic Regression model, enhanced with interaction terms, as it achieved a 97% accuracy, reflecting a +1% improvement in performance.

Moving forward, next steps include further optimizing the model by exploring additional feature interactions and employing other advanced machine learning techniques.

Additionally, validating the model with external datasets and integrating it into clinical workflows could enhance its practical utility for early breast cancer detection.

Questions?

Insights from the Breast Cancer Wisconsin Diagnostic Dataset

DATASET INFO

Includes features computed from digitized fine needle aspirate (FNA) images of breast masses, describing various attributes of the cell nuclei present in the images

CLASS DISTRIBUTION

357 benign, 212 malignant

SOURCE

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download

Attribute Information:

1. **ID number**
2. **Diagnosis**: M = malignant, B = benign

Features Computed for Each Cell Nucleus (30 in total):

For each cell nucleus, ten real-valued features are calculated:

a) **Radius**: Mean distance from the center to the perimeter points
b) **Texture**: Standard deviation of gray-scale values
c) **Perimeter**
d) **Area**
e) **Smoothness**: Local variation in radius lengths
f) **Compactness**: $\frac{\text{Perimeter}^2}{\text{Area}} - 1.0$
g) **Concavity**: Severity of concave portions of the contour
h) **Concave Points**: Number of concave portions of the contour
i) **Symmetry**
j) **Fractal Dimension**: "Coastline approximation" - 1

For each feature, the mean, standard error, and "worst" (mean of the three largest values) are computed, resulting in a total of 30 features.