Predictive Diagnostic Assistant for Breast Cancer Screening

A use case for machine learning tools in the health care setting by:

Anna Remler

Antonette Goroch

Daniel Rodriguez

Grace Yoo

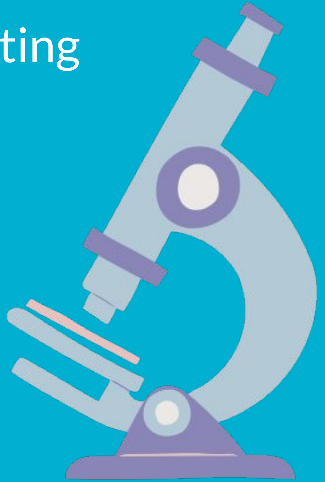
Luis Hernandez

Pooja Rajesekharan

About the Project

What need does this tool address?

- Breast cancer affects about 13% (about **1 in 8)** of U.S. women. 8*
- Misdiagnosis by medical professionals is a serious problem (46% of 2155 case were misdiagnosed in one often quoted study)
- Providing an **easy-to-access diagnostic tool** for health care professionals, which could utilize machine learning tools and existing research data to isolate key diagnostic features and direct to additional resources, could help reduce misdiagnosis.
- Since early detection is one of the best tools in reducing negative outcomes, such a tool could truly help save lives.



***Source: American Cancer Society, based on 2022 data.*

Who & How

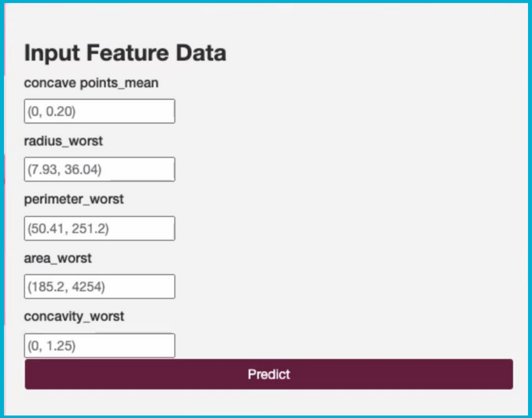
Target Audience:

A wide spectrum of healthcare professionals who represent the first line of diagnostic defense in treating breast cancer. Such as:

- Lab Techs
- Nurses
- Physician Assistants

How the tool works:

- User inputs five cell nuclei features into a web interface, based on a patient's lab data.
- System returns a suggested course of action based on the probability of malignancy based on the machine learning model.



Our Process

Steps Taken + Who Did What

- Dataset Discovery (Pooja)
- Initial scoping & planning
- Data cleaning & pre-processing
  - Daniel converted the initial dataset to an SQL database.
- Exploratory analysis
  - Initial research on the dataset .
  - Histograms using Numpy (Grace)
  - Logistical Regression model (Luis)
  - Random Forest model (Grace)
- Deploy Model & Dashboard
  - Pooja, Grace & Luis worked on the Flask app and HTML to display the dashboard using Bootstraps.
- Final Presentation & Repository
  - Antonette created the final presentation deck for the repository.
  - Anna created the final ReadMe file & supporting image files.
  - Luis maintained our central Github repository.

- Data Cleaning & Preprocessing
  - PANDAS
  - SQL
- Exploratory Data Analysis
  - PANDAS
  - Numpy
  - SciKitLearn
  - Matplotlib
- Dashboard Creation
  - Pickle
  - Flask
  - HTML
  - Bootstrap

Data Processing & Pipeline

Load csv to SQLAlchemy db and EDA

Explore different models

Export chosen model and model scaler to pickle files

App

Backend
Model + Scaler

Connect online form to model using **predict_proba**

Frontend
User Form

Prediction

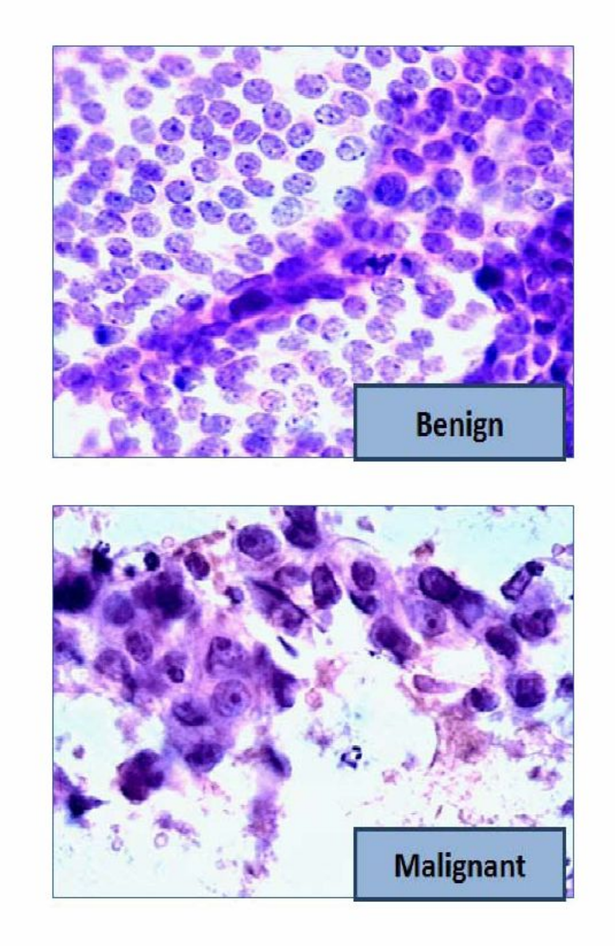
User's cell nuclei data

About the Data

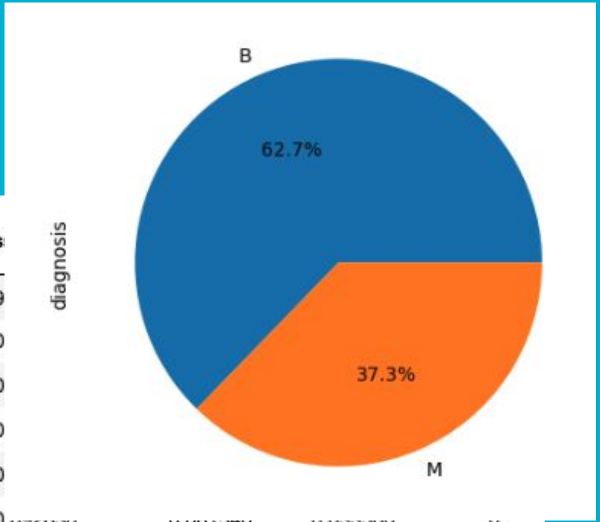
The data was obtained on Kaggle and uploaded by UC Irvine Machine Learning Repository but the results were found by University of Wisconsin.

The dataset was developed in 1995.

The dataset consists of 30 features–derived from 10 features by getting the mean, the standard error and the mean of the three largest values.



Data Exploration

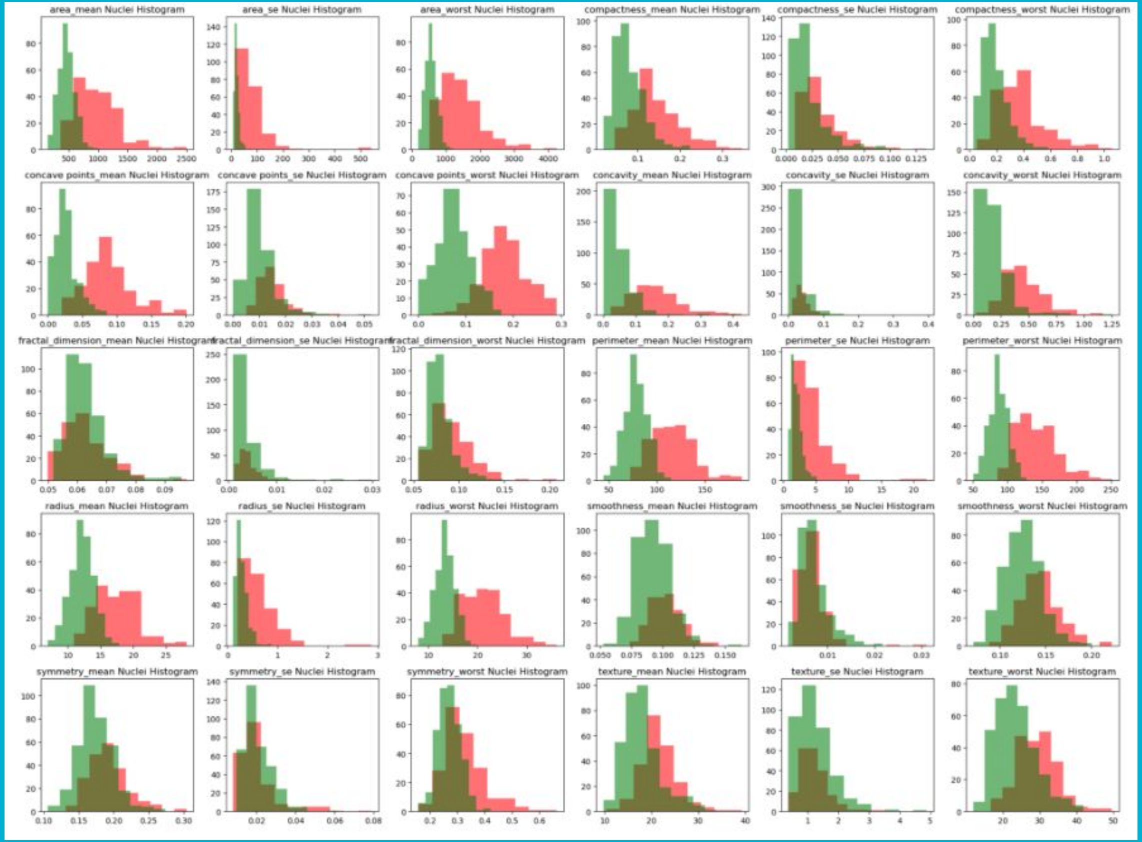


	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactnes			
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569			
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0			
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0			
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0			
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0			
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	[illegible]	[illegible]	[illegible]	[illegible]
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.

8 rows × 31 columns

Pre-Processing & EDA

- From initial CSV, created a SQL DB
- Used numpy to run histogram analysis, showing initial favorable variables.
- Reducing to most responsive variables.



Model Training, Tuning & Evaluation

- Goal: Find a model that reduces the number of user inputs with an acceptable level of accuracy.
- When including all 30 features, the Logistic Regression performed 1% better than Random Forests.

Logistic Regression with All Features

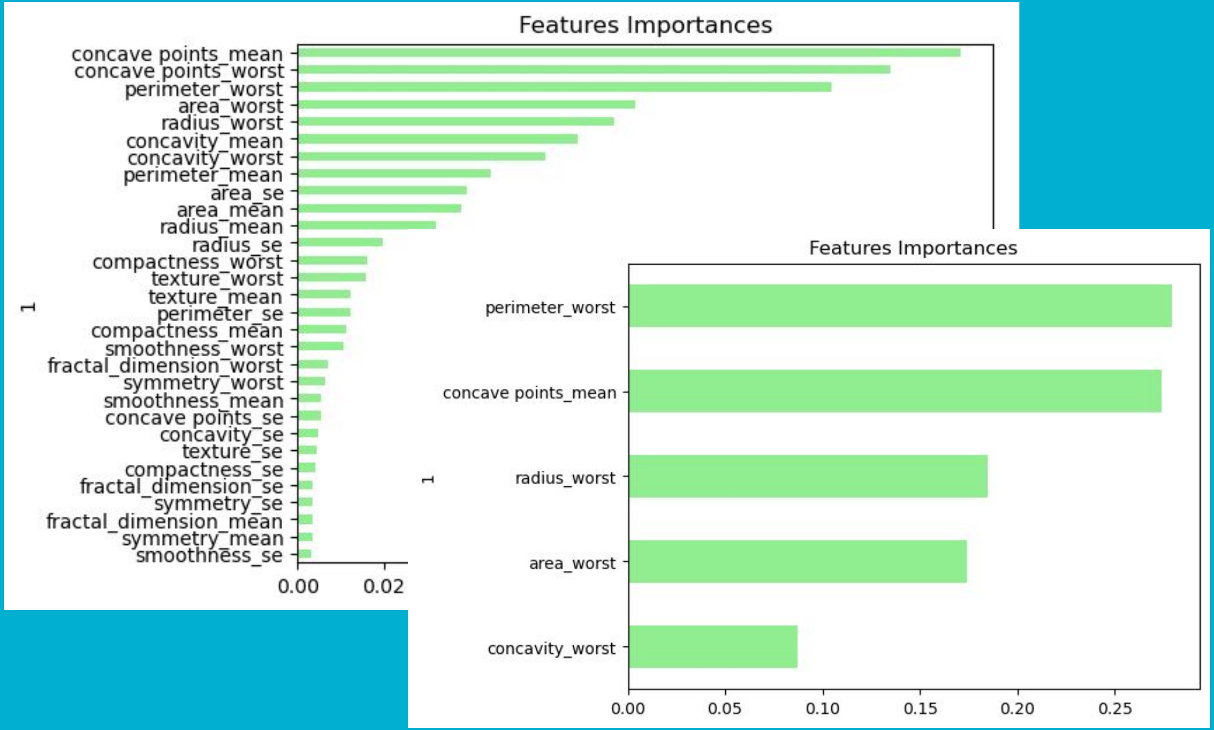
	precision	recall	f1-score	support
benign	0.98	0.99	0.98	88
malignant	0.98	0.96	0.97	55
accuracy			0.98	143
macro avg	0.98	0.98	0.98	143
weighted avg	0.98	0.98	0.98	143

Random Forests with All Features

	precision	recall	f1-score	support
B	0.98	0.96	0.97	84
M	0.95	0.97	0.96	59
accuracy			0.97	143
macro avg	0.96	0.97	0.96	143
weighted avg	0.97	0.97	0.97	143

Reducing the Number of Features

- By analyzing the feature importance through sklearns and the histograms of the initial data analysis, we found the top five most important features.
- We experimented with removing features from the models without sacrificing too much accuracy.



Model Training with Fewer Features

- Reducing to the top five features reduced the overall accuracy by 2-3% on both models.
- The models had equal accuracy at 94%.
- However, Random Forests performed 4% better than Logistic Regression at predicting malignant diagnosis.

Logistic Regression with Fewer Features

	precision	recall	f1-score	support
benign	0.94	0.97	0.96	88
malignant	0.94	0.91	0.93	55
accuracy			0.94	143
macro avg	0.94	0.94	0.94	143
weighted avg	0.94	0.94	0.94	143

Random Forests with Fewer Features

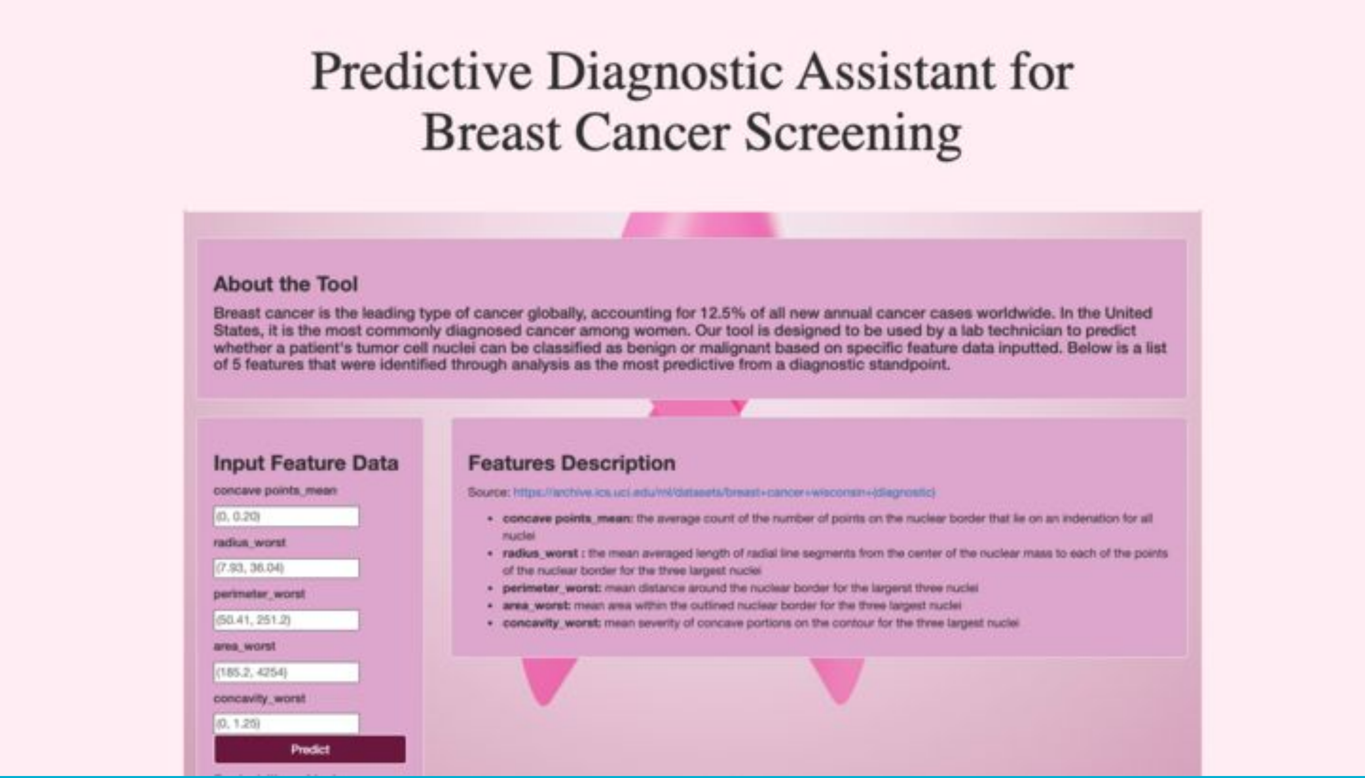
	precision	recall	f1-score	support
B	0.96	0.94	0.95	84
M	0.92	0.95	0.93	59
accuracy			0.94	143
macro avg	0.94	0.94	0.94	143
weighted avg	0.94	0.94	0.94	143

Creating the Dashboard

Pickle

Flask

Java



Challenges

- The main limitation we encountered was in the dataset itself, since it was somewhat limited in both size and scope (limited to one state, Wisconsin).

Next Steps

Several other areas of development would add significant value to the app:

- Refine the tool with additional/better data (i.e. data from more states, more current data).
- Further tools which build on the given outcomes (i.e. 30% chance of malignancy leads to a prompt for follow up with a specialist, or 3% might suggest retesting)
- Further research about the best type of health personnel to utilize the tool.

Thank you!

Special Thanks to Kevin and Mounika for the help!
Congratulations to everyone!