Predicting Breast Cancer

Breast Tumor Biopsy by Irish FNA Group

Danielle McDowell
Jack Brenner
Brian Witarsa
April 2025

Irish
FNA
Group

Join us in exploring a data-lead approach using machine learning to identify:

Can we accurately classify breast cancer tumors as malignant or benign using machine learning models?

Irish
FNA
Group

Content Overview

- Our Team
- Our Product
- The Data
- Exploring The Data
- Creating Features
- The Models
- Comparing Models
- Improving Sensitivity
- Key Insights & Conclusions
- Implementation Recommendations

Irish
FNA
Group

MEET THE TEAM

Danielle McDowell

Danielle is a graduate student in the University of Notre Dame's M.S. in Data Science program, combining her background in entrepreneurship with a passion for AI and analytics. She brings a strategic lens to data-driven projects with a focus on real-world impact.

Jack Bremer

Jack is a student in Notre Dame's M.S. in Data Science program, with a professional background in engineering and operations. He is passionate about applying data science tools to optimize decision-making and solve practical problems.

Brian Witarsa

Brian is currently pursuing his Master's in Data Science at the University of Notre Dame, with a strong interest in statistical modeling and machine learning. He leverages his analytical mindset to turn complex data into actionable insights.

Irish
FNA
Group

Our Product

What we've built and why

- **The Challenge:** How to interpret microscopic cell images with precision to diagnose breast cancer

- **How we can solve this using data:** Use machine learning can enhance diagnosis accuracy and reliability

- **Our Goal:** Create models to classify tumors with high sensitivity and specificity to predict the presence of breast cancer

Irish
FNA
Group

Our Product

What we've built and why

- **How it works:**

  - We use ***Fine Needle Aspiration*** (FNA) accompanied with a sensitive data model that ***utilizes machine learning*** to offer cost-effective, rapid diagnostics.

  - To build our model, we measured several data points within cell nuclei to determine and tune a sensitive model that increases accuracy in breast cancer prediction.

Irish
FNA
Group

Data Overview

Details of our Dataset

- Dataset Source: FNA cancer measurements (569 samples)
- Features: 30 measurements of cell nuclei characteristics
- Includes mean, standard error, and maximum values for 10 properties
- Properties: radius, texture, perimeter, area, smoothness, etc.
- Target Variable (Training for ‘Yes or No’ Diagnosis - Malignant = 1, Benign = 0)
- Data Quality: Ensure no missing values in relevant features

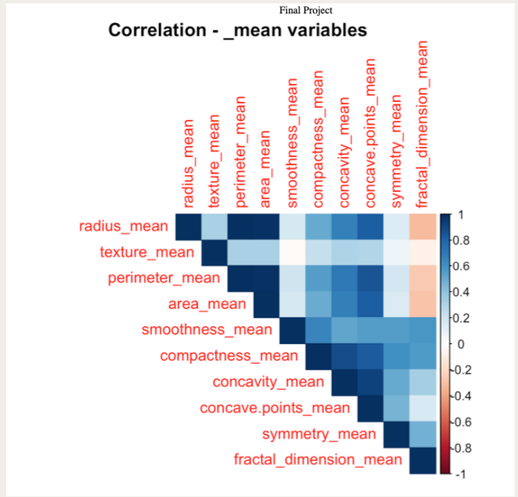
Irish
FNA
Group

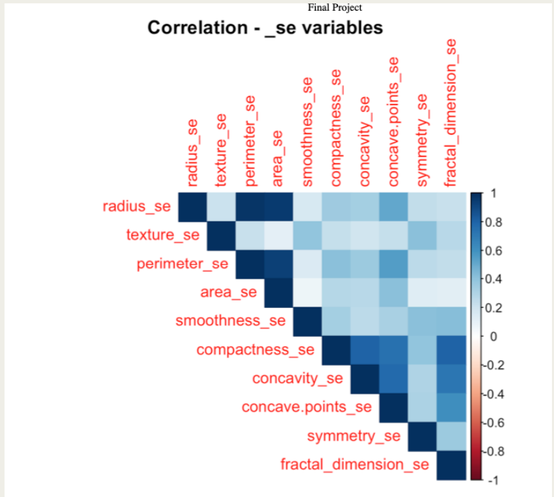
Exploring the Data

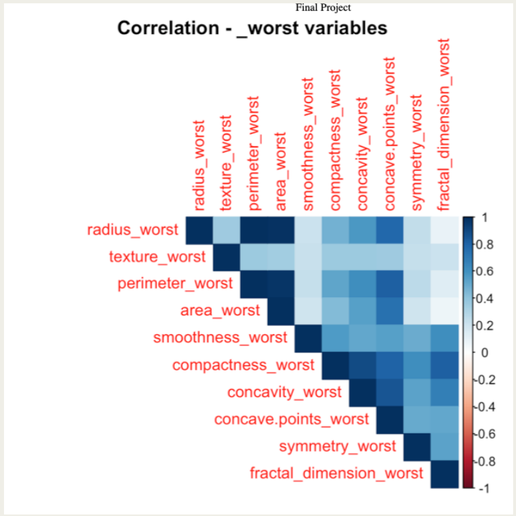
Exploratory Data Analysis

Before applying any models or implementing “Machine Learning”, we reviewed our dataset and implemented an analysis process to drive initial findings of how we would utilize and focus on different data points

Exploring the Data







Exploratory Data Analysis

Key Findings:

- High correlation among size-related variables (radius, area, perimeter) ($r > 0.9$)
- These features likely measure a shared concept: tumor size
- Strong relationship between concavity measures and malignancy
- Just a few features drive most predictive power
- Data points ‘perimeter_worst’, ‘concave.points_worst’, and ‘area_worst’ ranked highest

Creating Features

Honing in on key data

Data can be ‘noisy’. We used our exploratory data process to hone key datapoints that are critical to driving a diagnosis. Fewer features means a simpler, less expensive, and faster product

We used 3 key measurements to drive our models which:

- Reduces complexity of FNA image analysis
- Increases speed of diagnosis
- Lowers implementation costs
- Greater chance of real-world clinical adoption

Irish
FNA
Group

What is Machine Learning

Machine learning enables computer systems to recognize patterns in data—much like medical students learning through examples—by identifying features that distinguish malignant from benign cells, and improving their accuracy with more data and refined feature selection.

Irish
FNA
Group

The Models

Decision Tree

Interpretable model with clear decision paths

Random Forests

Ensemble approach to improve accuracy

K-Nearest Neighbors

Classification based on similar cases

Logistic Regression

Probability-based classification

Irish
FNA
Group

Comparing Models

Next, we review the key points of the models

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	94.7%	92.5%	95.8%
Random Forest	93.0%	94.1%	92.1%
Decision Tree	91.2%	94.1%	89.5%
KNN	88.6%	94.1%	85.3%

Irish FNA Group

Improving Sensitivity

Reducing False Positives

The next step was to improve and tune our models to reduce false positives. We needed to ensure the models sensitivity was high.

- Improving the model to achieve a consistent pattern across all methods:
  - All models achieved sensitivity > 92%
  - KNN and Decision Tree both reached 94.1% sensitivity

Irish
FNA
Group

Key Insights & Conclusions

Features

A few key features — like perimeter_worst, concave.points_worst, and area_worst — contribute most of the model's predictive power, allowing for simpler, faster, and more cost-effective implementations.

Sensitivity Tuning

Each model offers unique strengths: logistic regression had the highest overall accuracy, decision trees provided strong sensitivity with interpretability, and KNN delivered high sensitivity with minimal tuning.

Models

All models significantly improved sensitivity compared to traditional FNA, which helps reduce dangerous false negatives in cancer diagnosis.

Irish
FNA
Group

Implementation Recommendations

Take action

- Deploy logistic regression as primary classification model
- Use decision tree as interpretable backup for physician reference
- Focus on measuring the 3-5 most important features
- Implement regular model retraining with new data
- Establish monitoring system for prediction confidence

Irish
FNA
Group

Thank you

Danielle McDowell
Jack Brenner
Brian Witarsa
April 2025

Irish
FNA
Group