Breast Cancer Cell Classification

ML2 Final Presentation

Rachel Seo, Elaine Liu, Minu Choi
December 10, 2025

MEET THE TEAM

Rachel Seo
Favorite ML Model: Decision Tree

Elaine Liu
Favorite ML Model: Random Forest

Minu Choi
Favorite ML Model: XGBoost

ROADMAP

01 Objective

02 Dataset

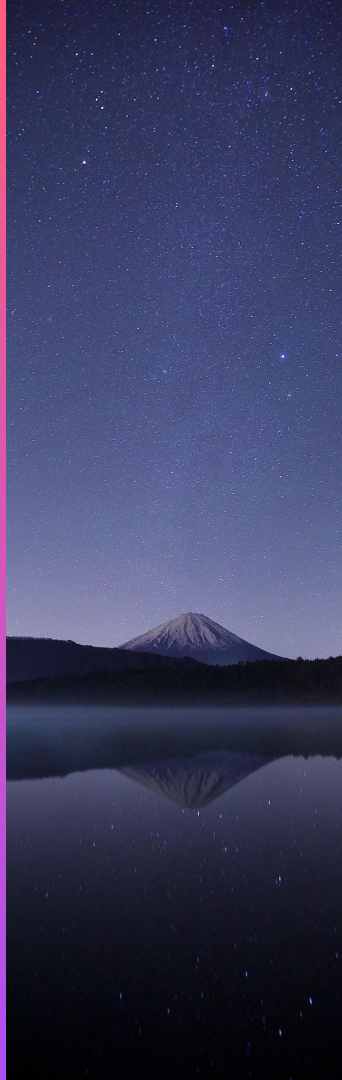
03 Methodology

04 Results

05 Conclusion

OBJECTIVE

Predict the presence of malignant or benign breast cancer given information from tumorous cells



DATASET

Example features from dataset (not comprehensive)

Feature	Description
radius_mean	Mean radius of the cell nuclei
texture_mean	Standard deviation of gray scale values
perimeter_mean	Mean of the cell nuclei perimeter
area_mean	Mean area of the cell nuclei
diagnosis	0 (benign) or 1 (malignant)

Methodology: Logistic Regression

Preprocessing: StandardScaler applied to normalize features within a Pipeline to prevent data leakage during cross-validation.

Hyperparameters tuned: Regularization strength (C: 0.001-100), penalty type (L2), and solver (lbfgs, liblinear).

Method: GridSearchCV with 5-fold StratifiedKFold cross-validation on 455 training samples.

Evaluation metric: Recall (to minimize false negatives/missed cancer cases).

Results:

Best CV accuracy: 0.9529 (95.29% recall)
Best param: C=1, penalty='l2', solver='lbfgs'

Methodology: SVM

Preprocessing: StandardScaler fitted per fold to prevent data leakage

Hyperparameters tuned: svm__C, svm__kernel, svm__gamma, svm__degree

Method: GridSearchCV with 5-fold cross-validation

Evaluation metric: Recall (minimize false negatives / missed cancers)

Results:

Best CV accuracy: 0.9588
Best param: svm__C: 1, svm__kernel: rbf, svm__gamma: scale, svm__degree: 2

Methodology: Extreme Gradient Boosting (XGBoost) & Neural Network

XGBoost

Gradient Boosting Ensemble

Hyperparameters tuned: n_estimators, max_depth, learning_rate, subsample, colsample_bytree, min_child_weight, gamma

Method: GridSearchCV with 5-fold StratifiedKFold cross-validation

Evaluation metric: Recall (minimize false negatives / missed cancers)

Neural Network

Model: Feedforward Neural Network (Multilayer Perceptron)

Hyperparameters tuned: learning_rate, batch_size, hidden layer sizes, dropout rate

Method: Manual 5-fold StratifiedKFold CV with early stopping (patience=10)

Preprocessing: StandardScaler fitted per fold to prevent data leakage
Evaluation metric: Recall (minimize false negatives / missed cancers)

RESULT 1: XGBoost

Key Metrics:

Best Cross-Validation Recall: 0.9588

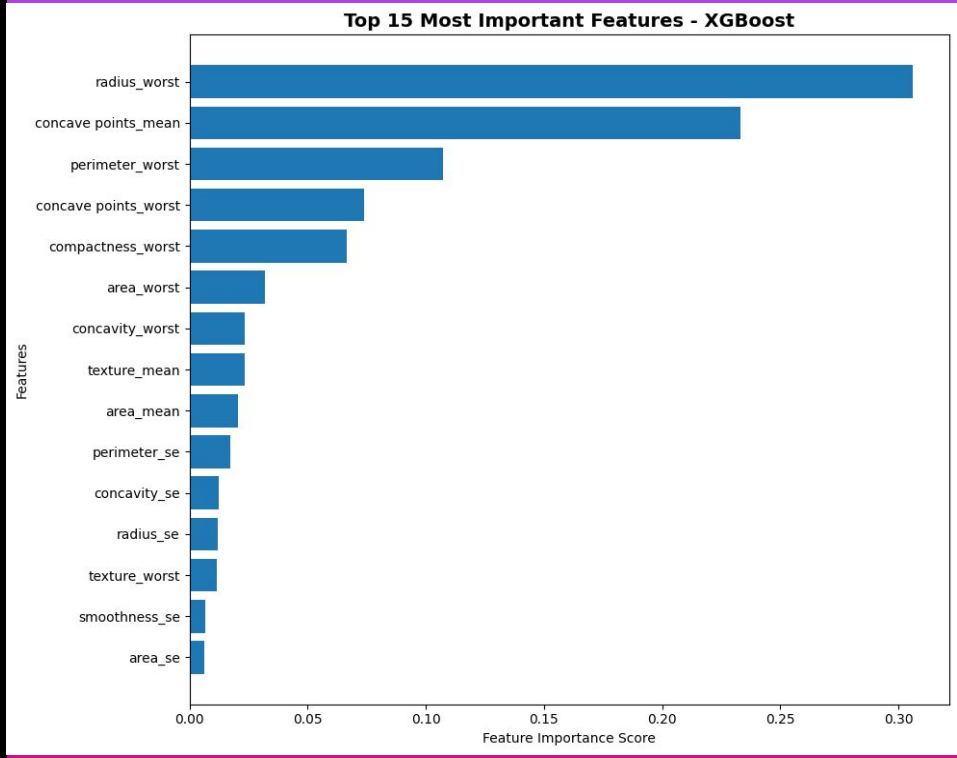
Hyperparameter combinations tested: 972 (4860 total training runs)

Best Hyperparameters Found:

learning_rate=0.3, max_depth=5, n_estimators=200 subsample=1.0, colsample_bytree=0.8, min_child_weight=1, gamma=0

Key Insight:

Top predictors: radius_worst, concave points_mean, perimeter_worst — aligns with clinical knowledge that tumor size and morphology indicate malignancy



RESULT 2: Neural Network

Key Metrics:

Best Cross-Validation Recall: 0.9882

Hyperparameter combinations tested: 243 (1,215 training runs with early stopping)

Best Hyperparameters Found:

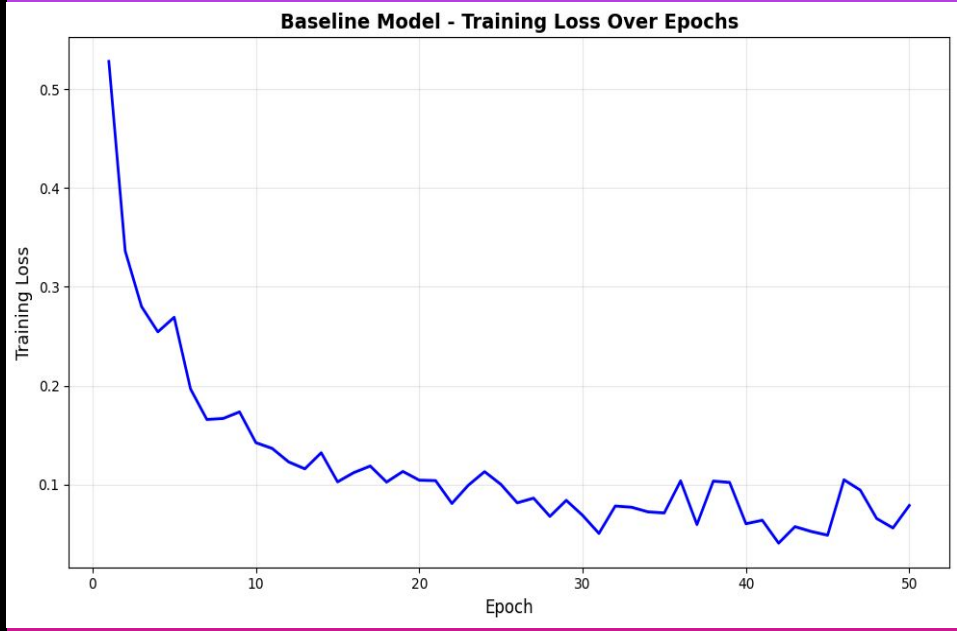
learning_rate=0.1, batch_size=64

hidden layers=[128, 64], dropout=0.3

Converged in avg 1 epoch (early stopping)

Key Insight:

Early stopping significantly reduced training time while achieving highest recall among tested configurations



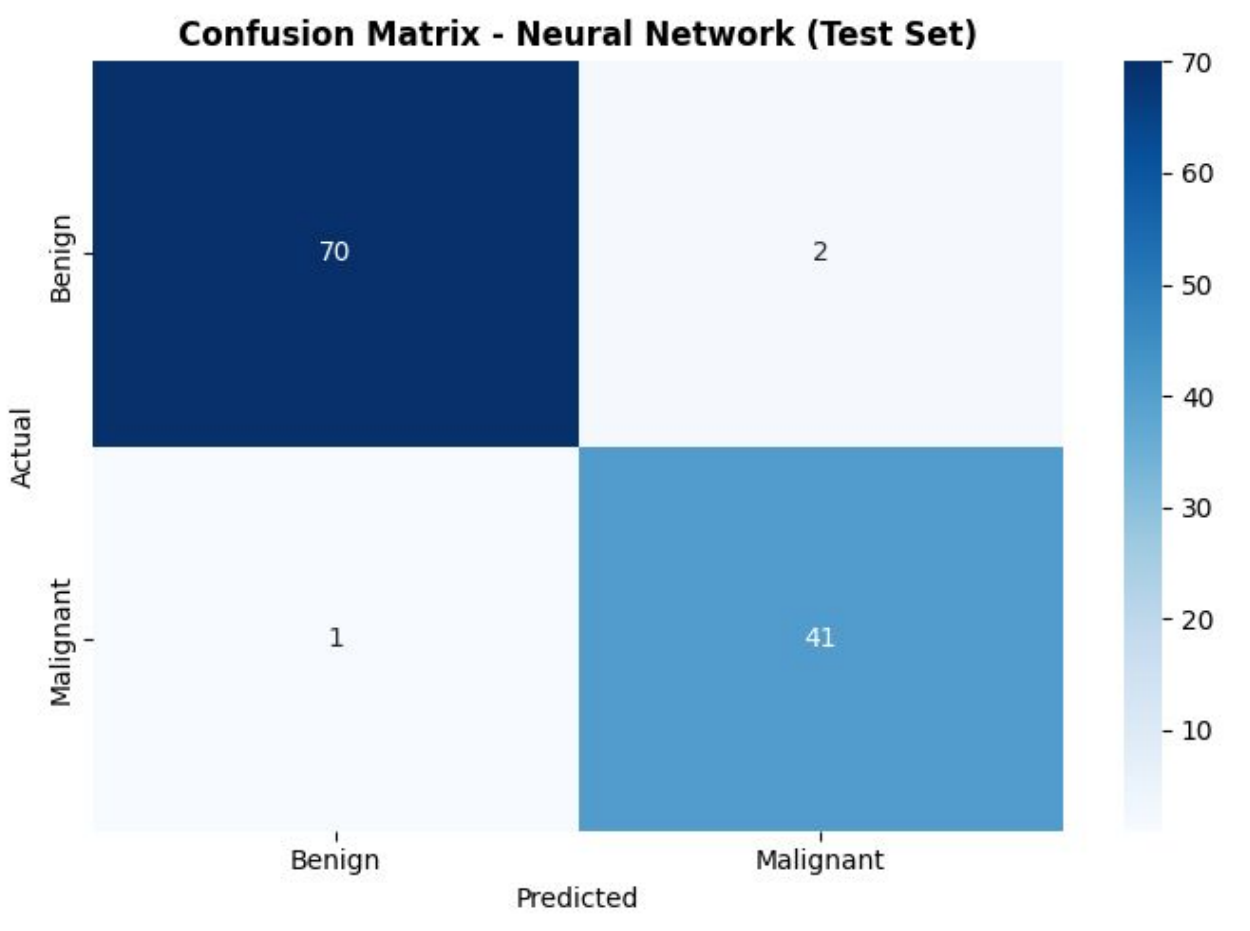
RESULTS: MODEL SELECTION based on CV-RECALL

Best Model: Neural Network (Feedforward MLP)

TEST SET EVALUATION

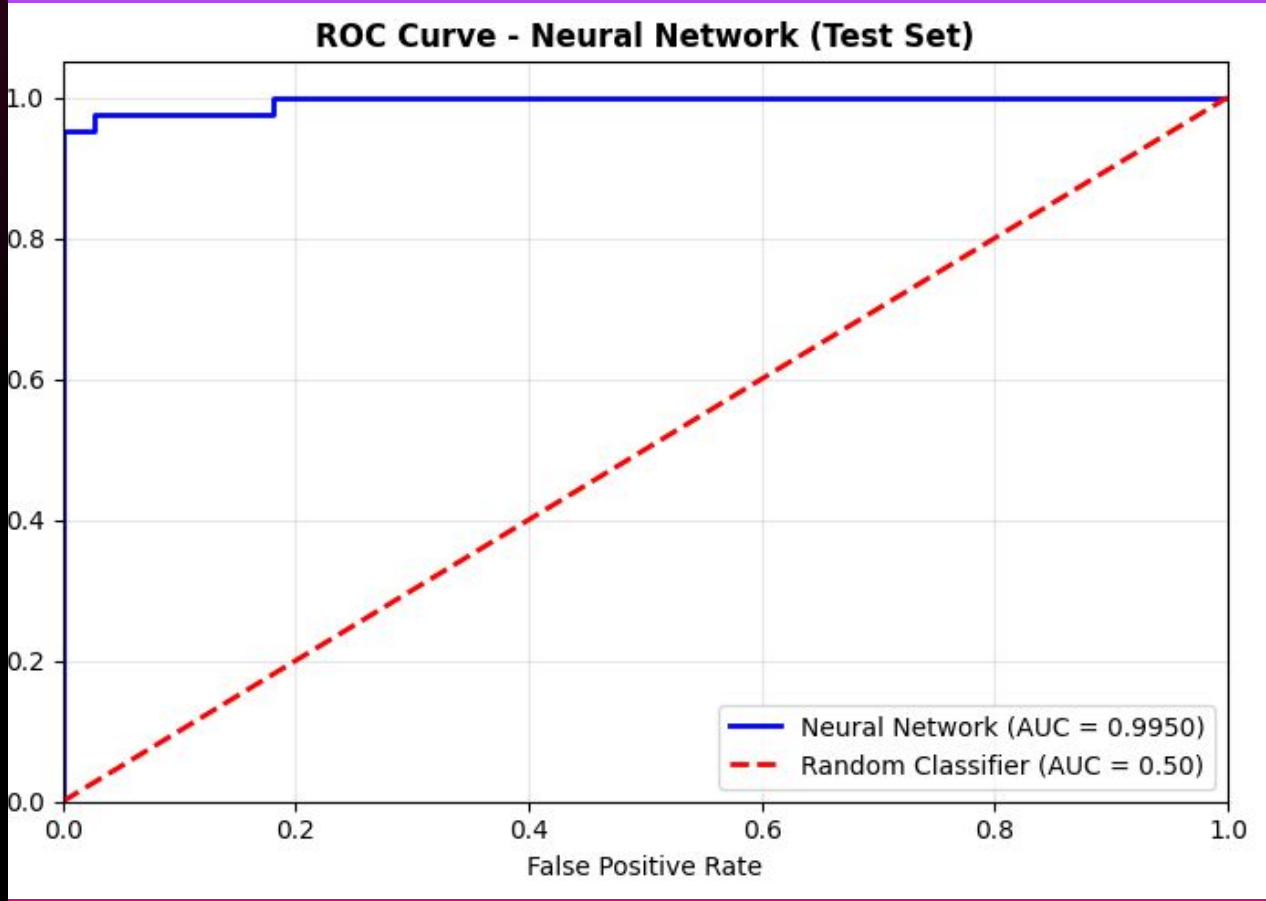
Feedforward Neural Network
(Multilayer Perceptron)

Metric	Score
Accuracy	97.37%
Precision	95.35%
Recall	97.62%
F1-Score	96.47%
ROC-AUC	99.50%



TEST SET EVALUATION

Feedforward Neural Network
(Multilayer Perceptron)



CONCLUSION

Neural Network Achieved Superior Performance

Minimized false negatives the most effectively.

XGBoost and SVM Tied for Second Place

Both effectively captured complex non-linear patterns in the data.

Logistic Regression Provided Strong Baseline with Interpretability

Advantages of simplicity, fast training time, and transparency.

Proper Cross-Validation and Recall Optimization Were Essential

Methodological consistency was imperative.

LIMITATIONS

01

Limited Data and Single-Source Dataset

- Small sample size: Only 455 training samples may not capture the full diversity of breast cancer presentations
- Single dataset source: Wisconsin Breast Cancer Dataset from one institution limits generalization to other hospitals, populations, and geographic regions

02

Methodological and Validation Gaps

- No external validation: Models only validated on splits from the same dataset, not tested on external/independent cohorts
- Overfitting risk: Neural Network's 98.82% recall seems suspiciously high; may indicate overfitting despite cross-validation

FUTURE STEPS

If we were to move forward with this dataset, we would:

- Compare performance metrics (recall, precision, F1-score, AUC-ROC) across models
- Analyze confusion matrices to understand error patterns (false positives vs. false negatives)
- Conduct statistical significance tests to determine if performance differences are meaningful

In order to deliver a **comprehensive test results table** and **recommendation** for possible production deployment.