Breast Cancer Classification Using Machine Learning Models

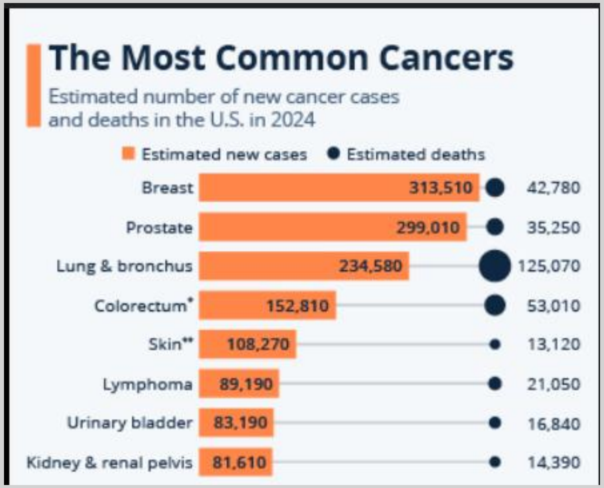
Leena Rajesh Patil

Advanced machine learning

Final project ppt

Why This Matters

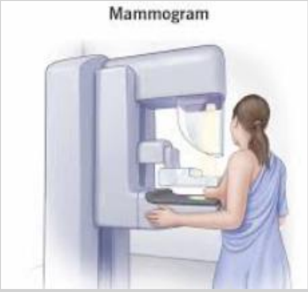
- Breast cancer is the most commonly diagnosed cancer in women worldwide.
- Early diagnosis and treatment significantly improve survival rates.
- Machine learning can assist in providing fast, accurate, and affordable diagnosis.



The Problem

- Traditional methods are like mammography, biopsy etc. They are invasive, expensive, and not universally accessible.
- Mmammograms can miss tumours or set false alarms
- Which may lead to unnecessary biopsies

A better approach is a necessity !




Project Goal

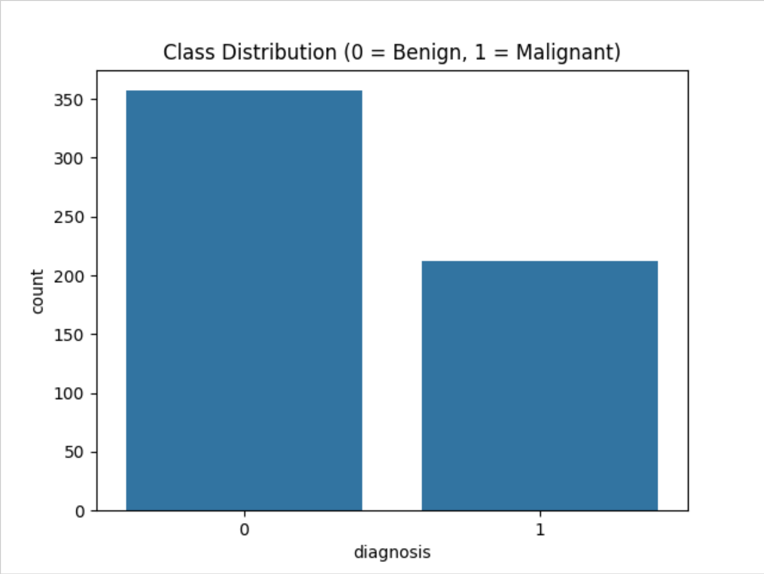
- Build an ML pipeline to classify tumors as benign or malignant using WBCD dataset.
- Preprocessing(Drop ID column, Encode target, Scale features, Variance Threshold)
- Train-Test Split (70/30 Stratified)
- Feature Selection (if applicable)
- Model Training(Decision Tree, Random Forest, Gradient Boosting, AdaBoost, SVM (RBF & Linear),Neural Net)
- Model Evaluation(Accuracy & AUC, ROC Curves,)

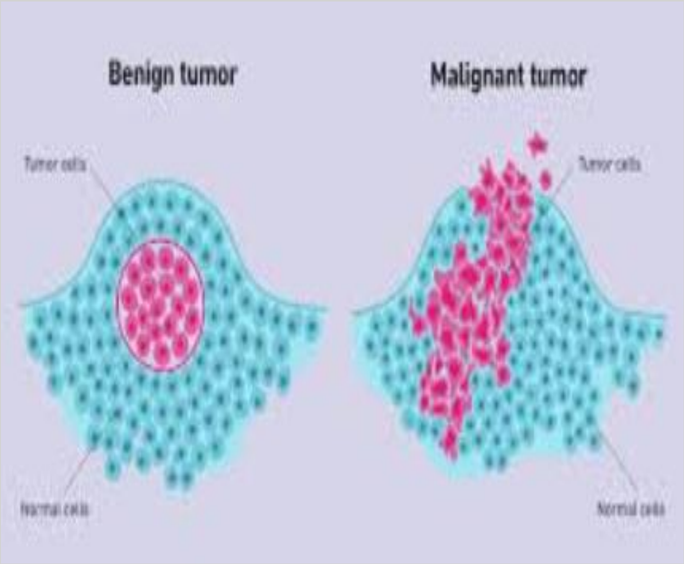
The Dataset

- 569 samples, 30 features, 2 classes (Benign, Malignant).
- Each tumor has 10 measured characteristics(radius, texture, perimeter, symmetry etc.)
- Each feature has 3 forms: *Mean*, *Standard Error*, and *Worst (in total 10*3 = 30 features)*
- overall behavior (mean), variability (SE), and extreme abnormality (worst)

Benign vs Malignant Tumors

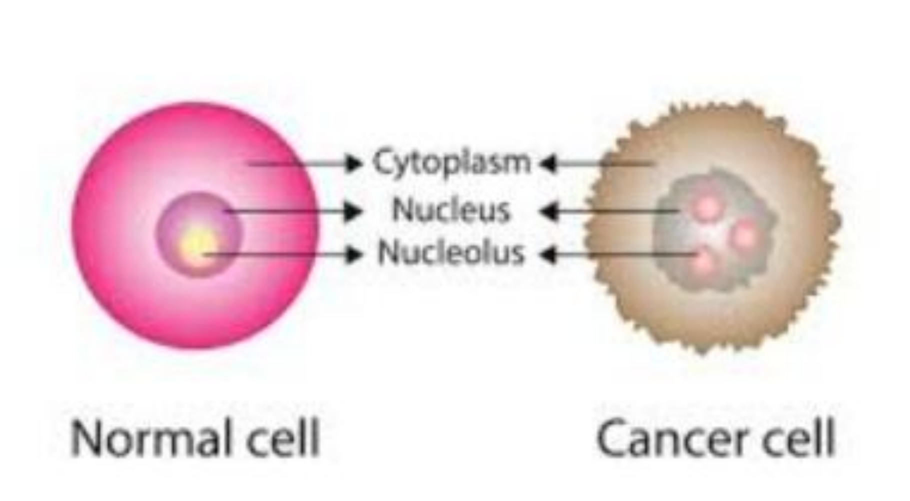
- Benign: Non-cancerous, does not spread, often removable.
- Malignant: Cancerous, invades tissues, can metasize through blood and lymph.





Biological Features

- There is an evident morphology change between normal and cancer cell (Nucleus shape, size, and texture indicate malignancy)
- Examples: Radius, Perimeter, Area, Smoothness, Concavity, Symmetry



Literature Insights

- Tree models like Random Forest, Gradient Boosting performed well, they have good explain ability too.
- SVM models also achieve high accuracy
- deep models(neural nets) need tuning & large data.

Model testing I (before FS)

- Decision tree : AUC ~0.91 - prone to overfitting
- Random forest : AUC ~0.98 - excellent performance, interpretable
- Gradient Boosting : AUC ~0.986 - strong performance, more complex

Model Testing II (before FS)

- **AdaBoost** : AUC ~0.988 - simple, very effective
- SVM RBF : AUC ~0.989 - excellent, but black-box
- **Neural Net** : AUC ~0.990 - highest AUC, black-box

Feature Selection

- SelectFromModel used; skipped for SVM RBF and Neural Net.
- concave points_mean, perimeter_mean, area_worst, radius_worst, concavity_mean.

Model	Selected Features
Decision Tree	concave points_mean, area_worst, concave points_worst, perimeter_mean, radius_worst
Random Forest	perimeter_mean, area_mean, concavity_mean, concave points_mean, radius_worst
Gradient Boosting	concave points_mean, radius_worst, area_worst, perimeter_worst, area_mean
AdaBoost	texture_mean, smoothness_mean, compactness_mean, symmetry_mean, concave points_mean

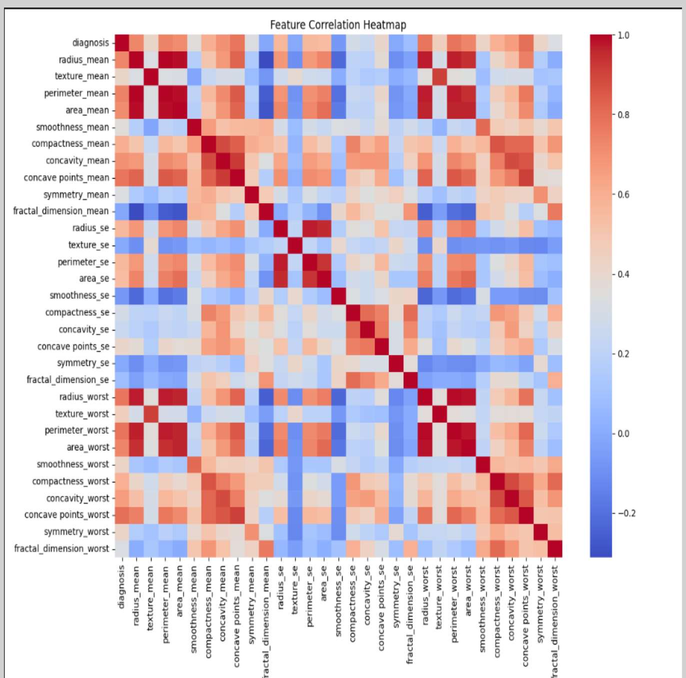
Impact of Feature Selection on Model Performance

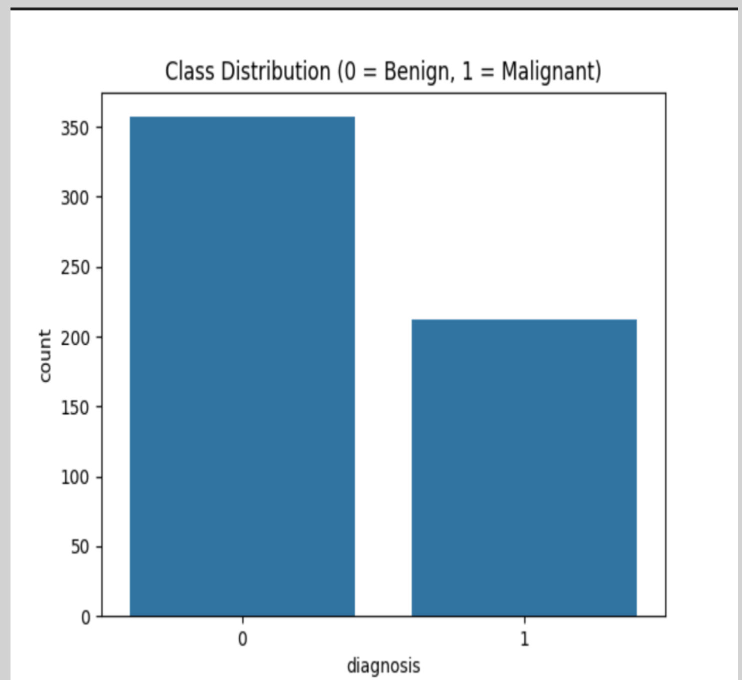
- **Performance remained stable** for Random Forest and AdaBoost
- **Slight AUC drop** in Decision Tree and Gradient Boosting, expected due to their reliance on full feature sets.

Model	Before FS Test AUC	After FS Test AUC	Change
Decision Tree	~0.917	~0.900	Slightly decreased
Random Forest	~0.981	~0.980	Essentially same
Gradient Boosting	~0.986	~0.971	Slightly decreased
AdaBoost	~0.988	~0.987	Essentially same

EDA Insights

- **Correlation:** Features like *radius*, *area*, and *perimeter* are highly correlated.
- **Class Distribution:** Slight imbalance between benign and malignant samples





Summary Statistics

- Features vary widely, For example, *radius_mean* and *area_mean* had large ranges, while *smoothness_mean* and *concave points_mean* had very small values.
- We used **StandardScaler** to standardize features before model training.

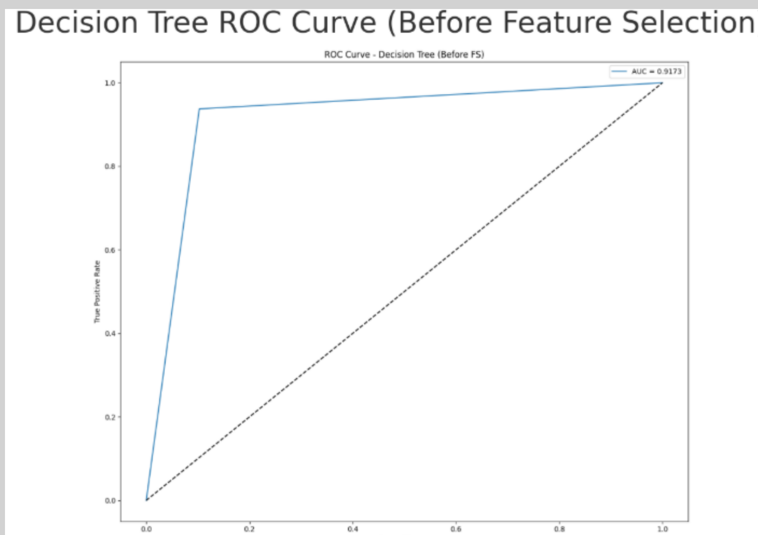
Feature	Mean	Std Dev	Min	Max
radius_mean	14.13	3.52	6.98	28.11
perimeter_mean	91.97	24.30	43.79	188.50
area_mean	654.89	351.91	143.50	2501.00
smoothness_mean	0.096	0.014	0.0526	0.1634
concave points_mean	0.0489	0.0388	0.0000	0.2012
symmetry_mean	0.1812	0.0274	0.1060	0.3040

Model Performance

- All models performed well.
- **Random Forest** and **AdaBoost** selected: excellent AUC, interpretable, clinically practical.
- Neural Network achieved highest AUC (~0.99) but was not selected due to lack of interpretability → black-box model.

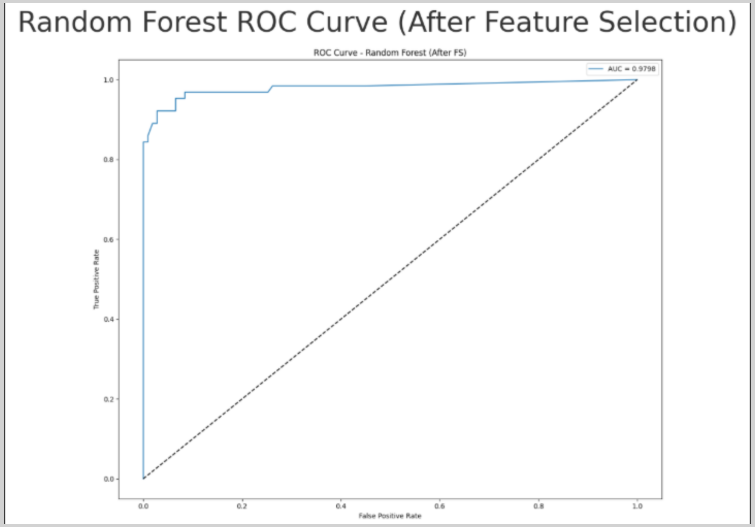
Decision Tree ROC

- Baseline model, decent AUC - 0.91, but prone to overfitting.
- Curve rises relatively smoothly but is **not very sharp** toward the top-left corner → meaning the model is not perfectly discriminating between classes.



Random Forest ROC

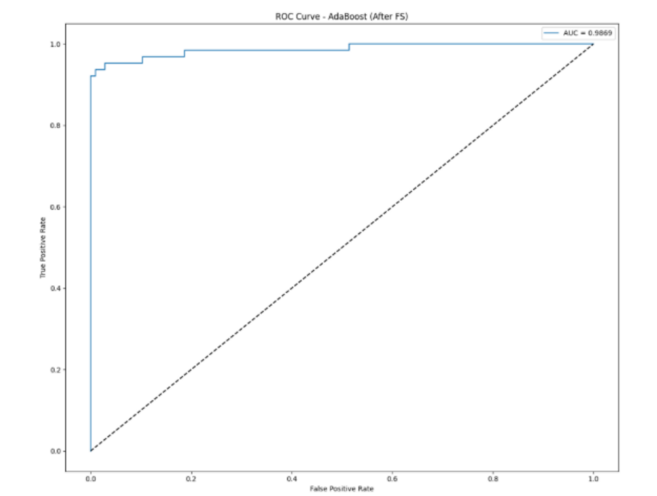
- Excellent AUC – 0.98 & interpretability — suitable for clinical use.
- Sharp rise towards top left, this means the model is achieving a **high true positive rate.**
- not overfitting, AUC remained similar to before feature selection



AdaBoost ROC

- Strong AUC – 0.987, simple & interpretable model.
- curve stays well above the diagonal i.e. making good predictions.
- performed well despite being simpler than random forest
- performance maintained even after feature selection

AdaBoost ROC Curve (After Feature Selection)



Why These 3 ROC Curves?

- Decision tree = baseline performance
- RF and Adaboost = strong + interpretable
- Didn't show neural and SVM = black-box models + less interpretability.

Interpretability is important!

Future Work & Conclusion

- Larger datasets, multi-modal data
- Calibration & XAI
- Clinician collaboration

RF & AdaBoost recommended!

less features

good performance

good interpretability