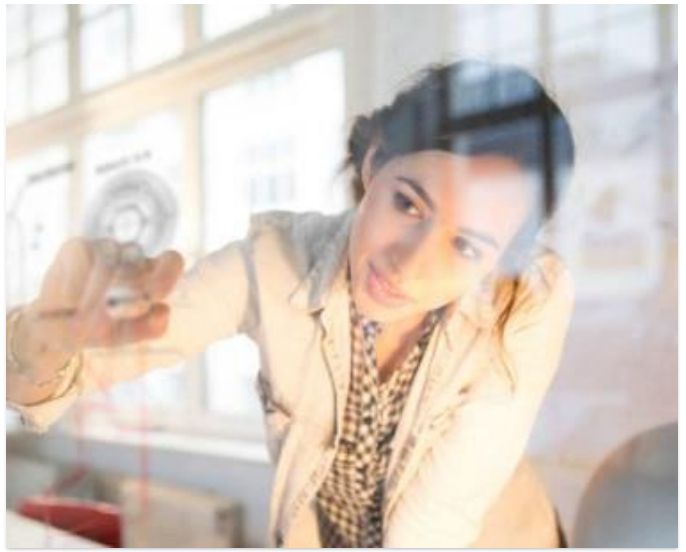
Using Composite Features Engineering & Ensemble Models

ML CLASSIFICATION: TUMOR DIAGNOSIS

OBJECTIVE

- Diagnose breast-tumor: ‘B’, ‘M’
- Minimize False Negatives
- Maintain High Interpretability



DATA OVERVIEW

- Wisconsin University Cancer Diagnosis Data Set
- 569 Records
- 30 Numeric Features
- No missing values or duplicates

TEST TRAIN SPLIT

Done before EDA

Copy used for EDA

Same split preserved for pipeline and then training

STATISTICAL PROFILE

Variation in range:

- fractal_dimension_mean: ~0.05–0.1
- area_worst: ~185–3432
- 3 sub-categories:
  - _mean
  - _se
  - _worst
- 10 features per sub-category

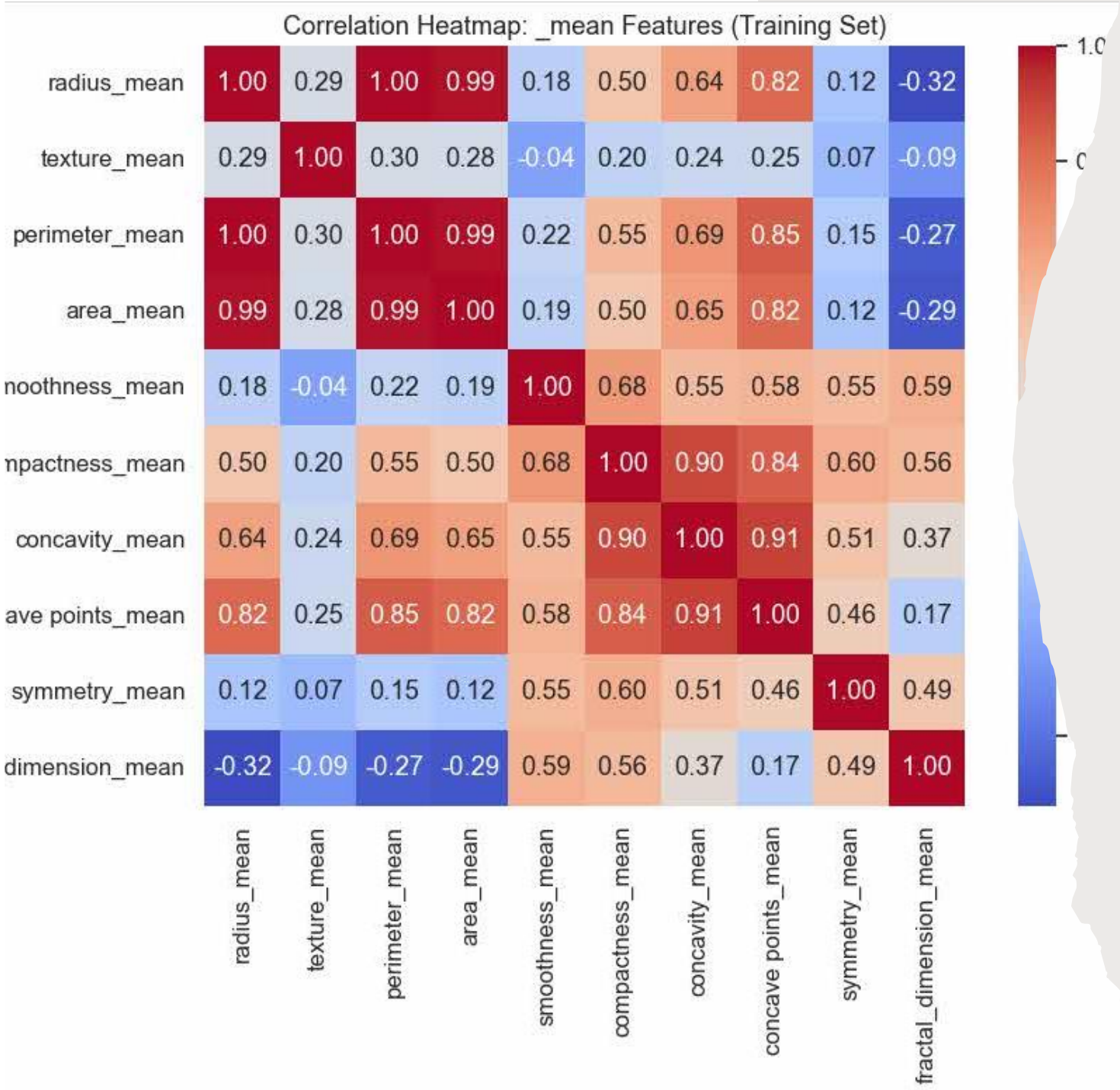
CORRELATION

Two feature sets highly correlated:

radius_, perimeter_, area_: 0.94 to 1.00

And

compactness_, concavity_, concave points_: 0.77 to 0.91



HANDLING

Group 1: keep area_ discard others

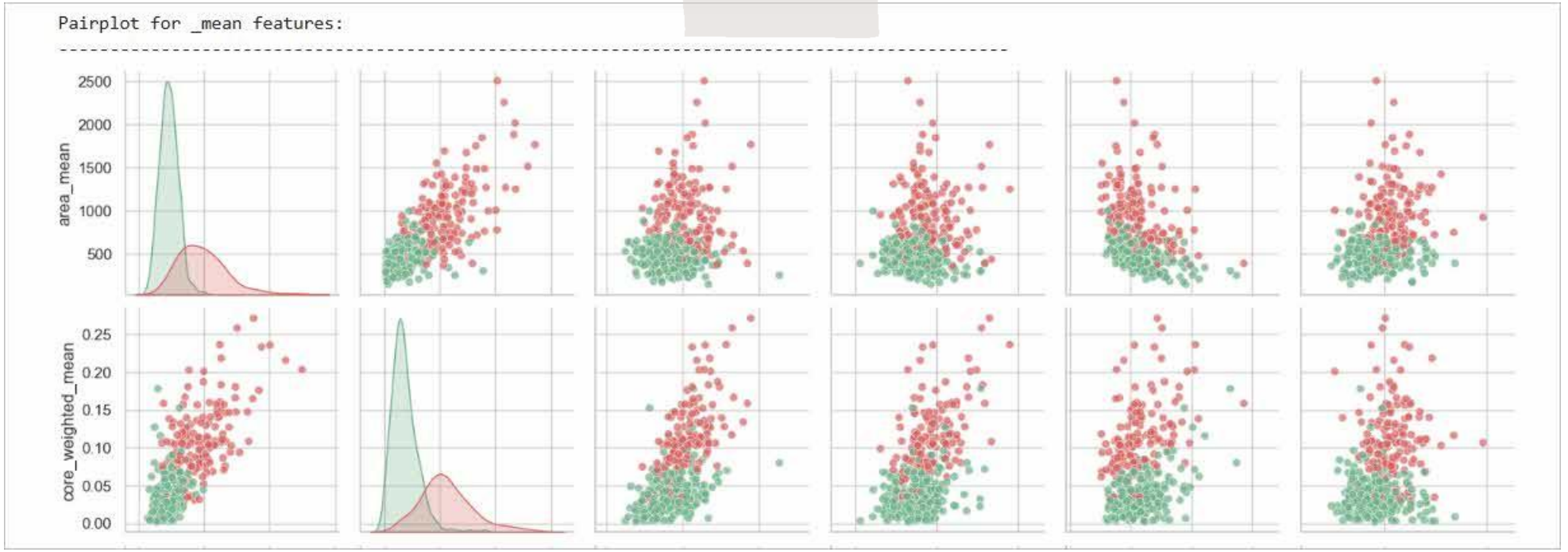
Group 2: add a composite feature core_weighted_* using

+ One composite of unrelated features: mass_nature_ = area & texture & smoothness

Weights derived from feature importance

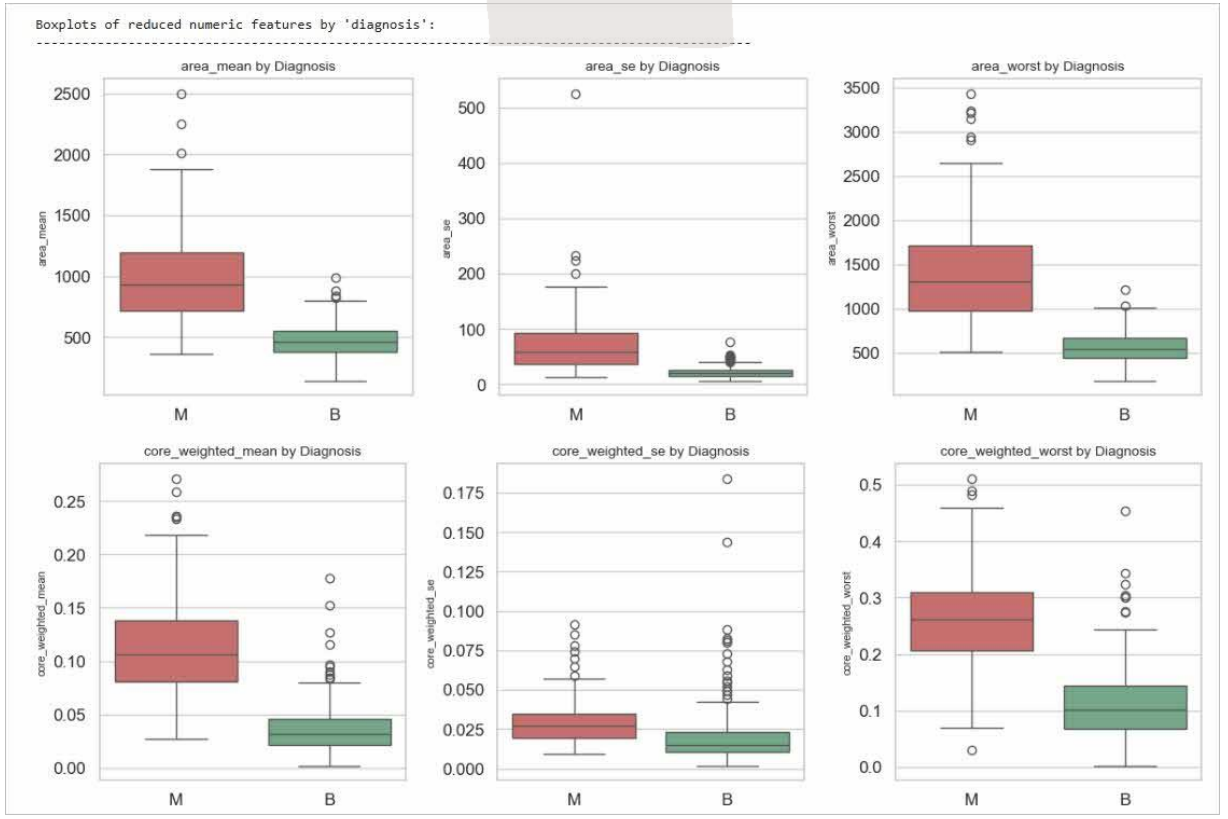
VISUAL ANALYSIS

Example 1: density distribution – especially in the composite feature



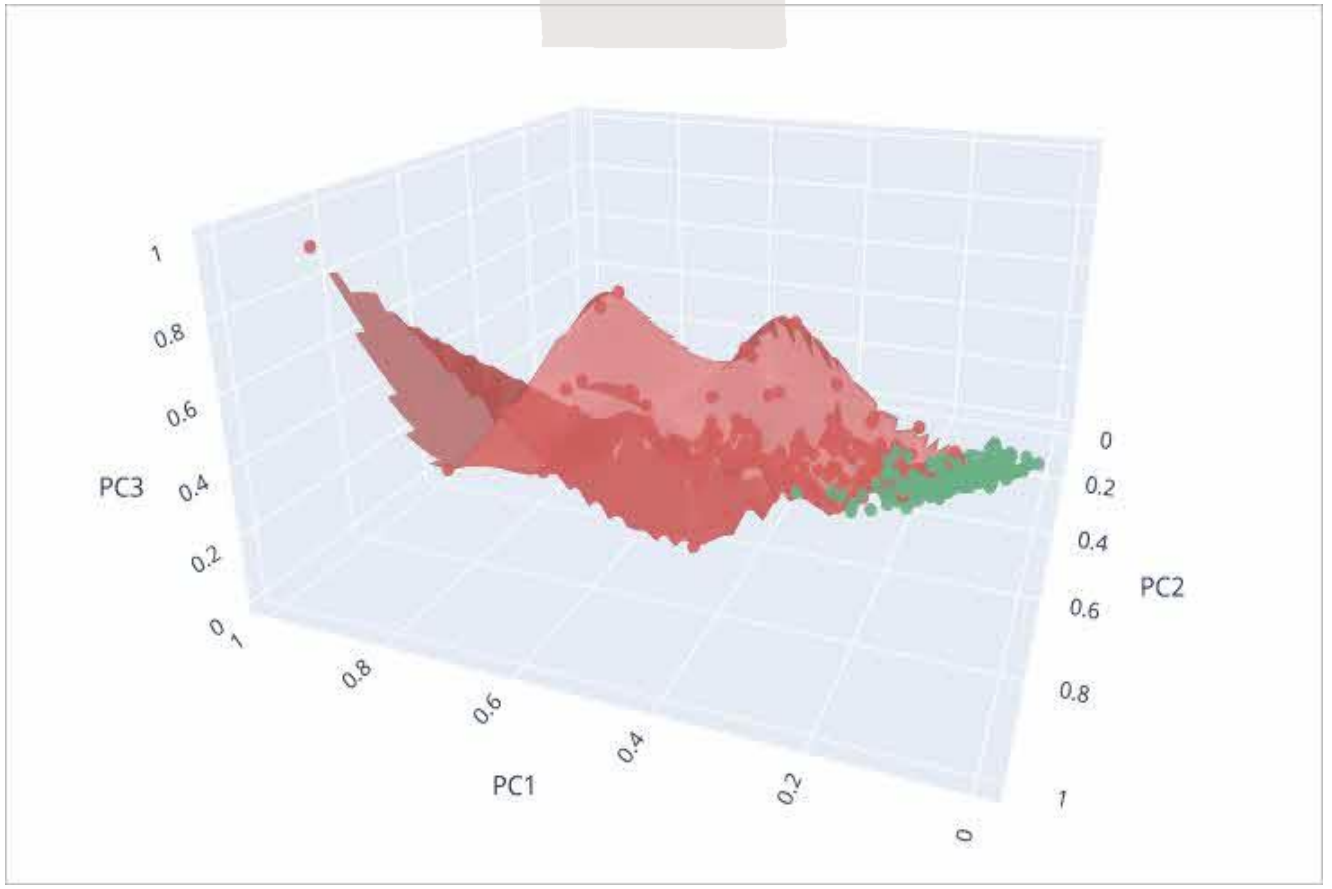
VISUAL ANALYSIS

Example 2: class separation



VISUAL ANALYSIS

3D PCA: for visualization only



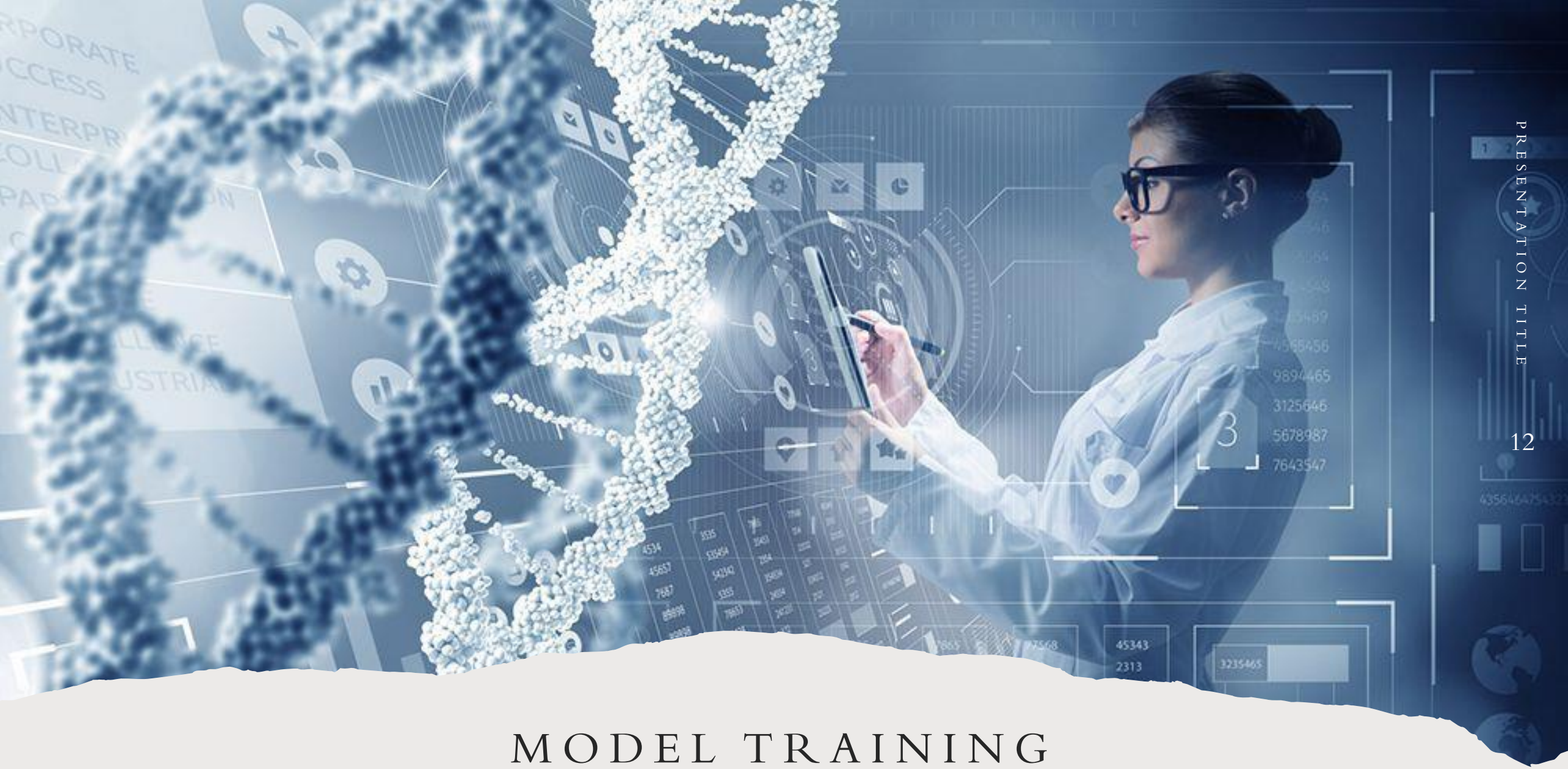
Not for modelling because can't be interpreted.

However, shows a promise that the classes might be well-separable.

PROCESSING

Fresh Copy of X_train, y_train

- Target encoding
- Feature engineering
- Outlier handing
- Robust scaling
- Etc.



MODEL TRAINING

CHOICE OF MODELS

Overview

7 models selected.

Mix of ensemble models, linear, non-linear and neural networks.

All models trained using 5-fold stratified cross-validation.

The Models

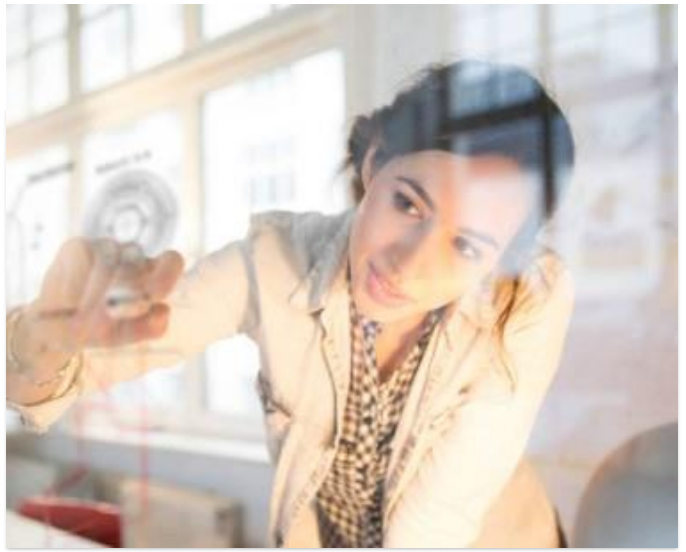
- Baseline: Logistic Regression
- Tree Based Ensembles: Random Forest, XGBoost, LightGBM
- Kernel Method: SVM (RBF)
- Neural Network: MLPClassifier
- Interpretable: Decision Tree

PERFORMANCE METRICS

Metrics Used:

- Accuracy
- Precision
- Recall
- FI
- ROC-AUC

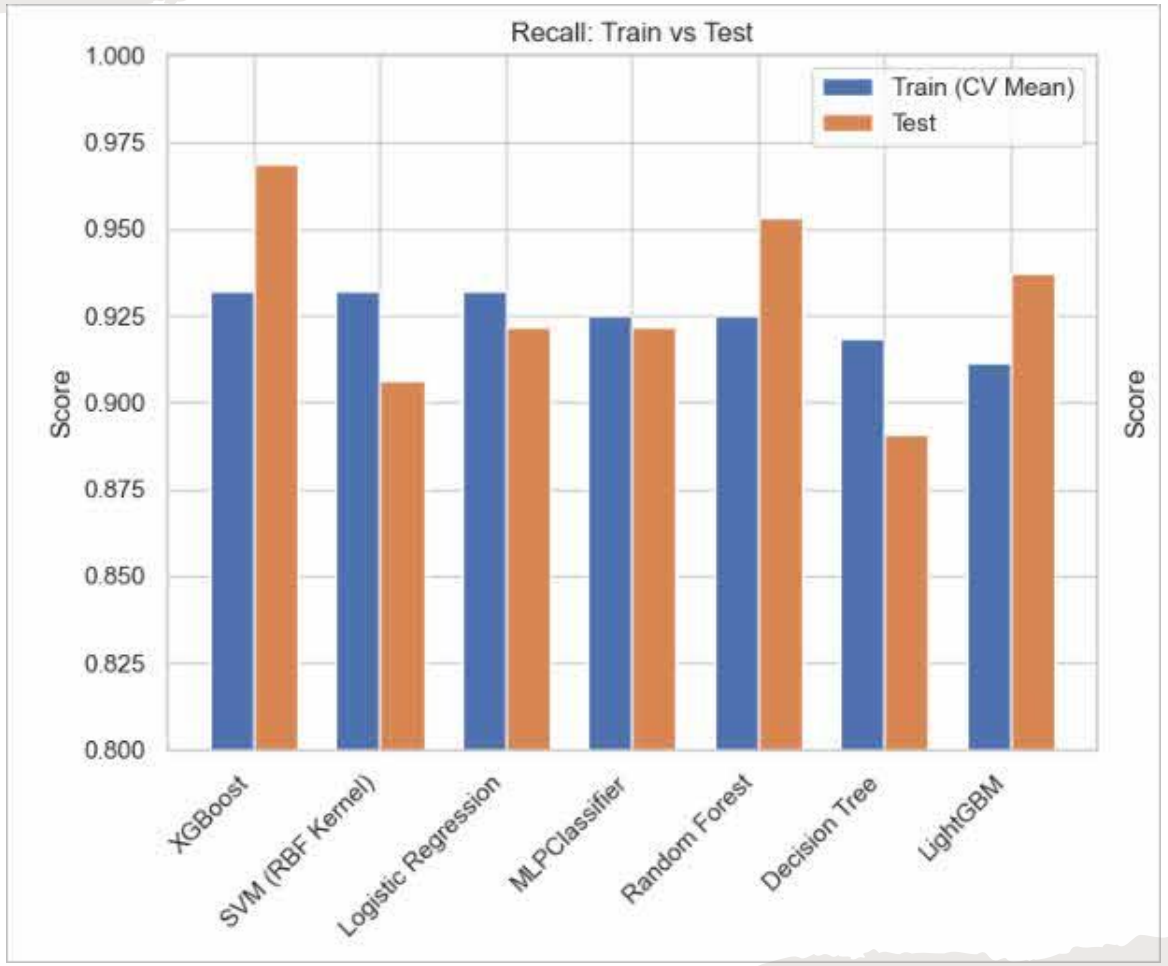
Recall on class 'M' is the most important metric: the goal is not to misclassify malignant



TOP 3 MODELS

Chosen based on recall on CV Mean:

- XGBoost
- SVM
- Logistics Regression

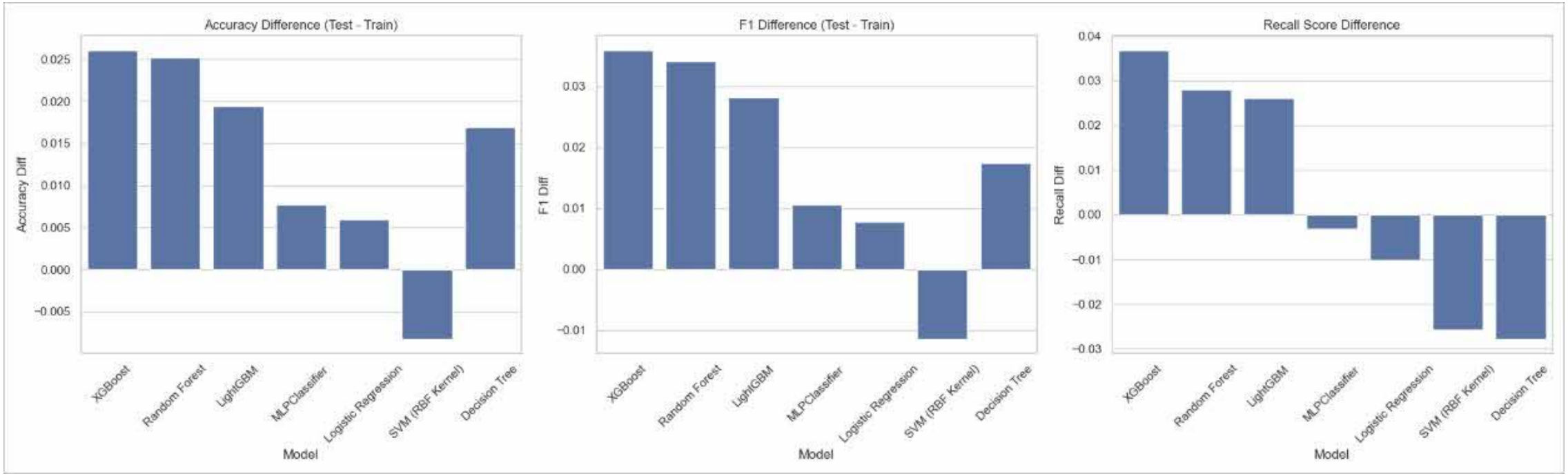


OVERFITTING? UNDERFITTING?

	Model	Recall Diff	F1 Diff	ROC-AUC Diff	Precision Diff	Accuracy Diff
3	XGBoost	0.036796	0.035888	0.001832	0.033565	0.026057
2	Random Forest	0.028068	0.034127	0.007104	0.037104	0.025209
4	LightGBM	0.026006	0.028134	0.007474	0.029333	0.019425
6	MLPClassifier	-0.003182	0.010638	0.002101	0.024286	0.007697
0	Logistic Regression	-0.010079	0.007834	0.004804	0.026667	0.005950
5	SVM (RBF Kernel)	-0.025704	-0.011360	0.004411	0.004306	-0.008214
1	Decision Tree	-0.027766	0.017362	0.008098	0.057371	0.016932

GENERALIZATION VISUALIZED

(Test – Train) difference range: -0.03 to 0.04 = good generalization!



MODEL TUNING

XGBOOST TUNING

Best Parameters

`n_estimators`: number of trees → more boosting rounds improve detection

`learning_rate`: 0.07 → makes model sensitive to subtle tumor patterns

`max_depth`: 3 → prevents overfitting noisy biopsy features

`subsample & colsample_bytree`: control robustness

Helps catch complex malignant patterns while keeping stability

How it Helps

- Boosting learns difficult malignant patterns iteratively
- Depth control avoids overfitting small benign variations
- Higher recall after tuning: 0.93 → 1.00
- Great for medical datasets with non-linear boundaries

SVM (RBF KERNEL) TUNING

Best Parameters

C controls margin softness → higher C reduces FN

gamma controls curve of boundary → captures subtle malignant signatures

Best params: C=5, gamma='scale'

Helps create flexible boundaries around malignant clusters

How it Helps

- RBF kernel discovers non-linear tumor separation
- High recall after tuning: 0.93 → 0.97
- Balances sensitivity and generalization

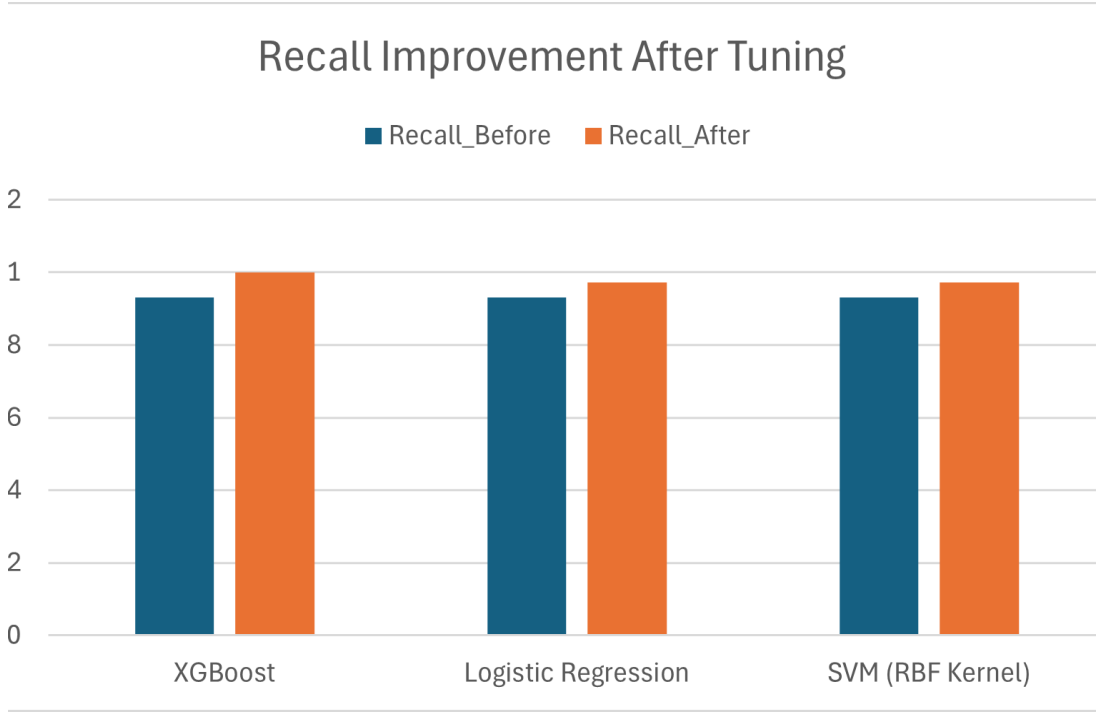
LOGISTICS REGRESSION TUNING

Best Parameters

- `C` controls regularization → larger C reduces underfitting
- `penalty='l2'` → stable boundary
- `solver='lbfgs'` → optimized for small/medium datasets
- Good for clinical interpretability

RECALL IMPROVEMENT

- Adjusts model behavior to reduce FN
- Focus on malignant class (pos_label=1)
- All 3 models improved recall significantly
- XGBoost emerged as final best model
- These tuned models feed into an Ensemble for higher robustness



ENSEMBLE

ENSEMBLE

- Soft-voting Ensemble
- VotingClassifier
- Got very close to XGBoost, but came No. 2

- Estimators
  - XGBoost
  - Logistics Regression
  - SVM (RBF Kernel)

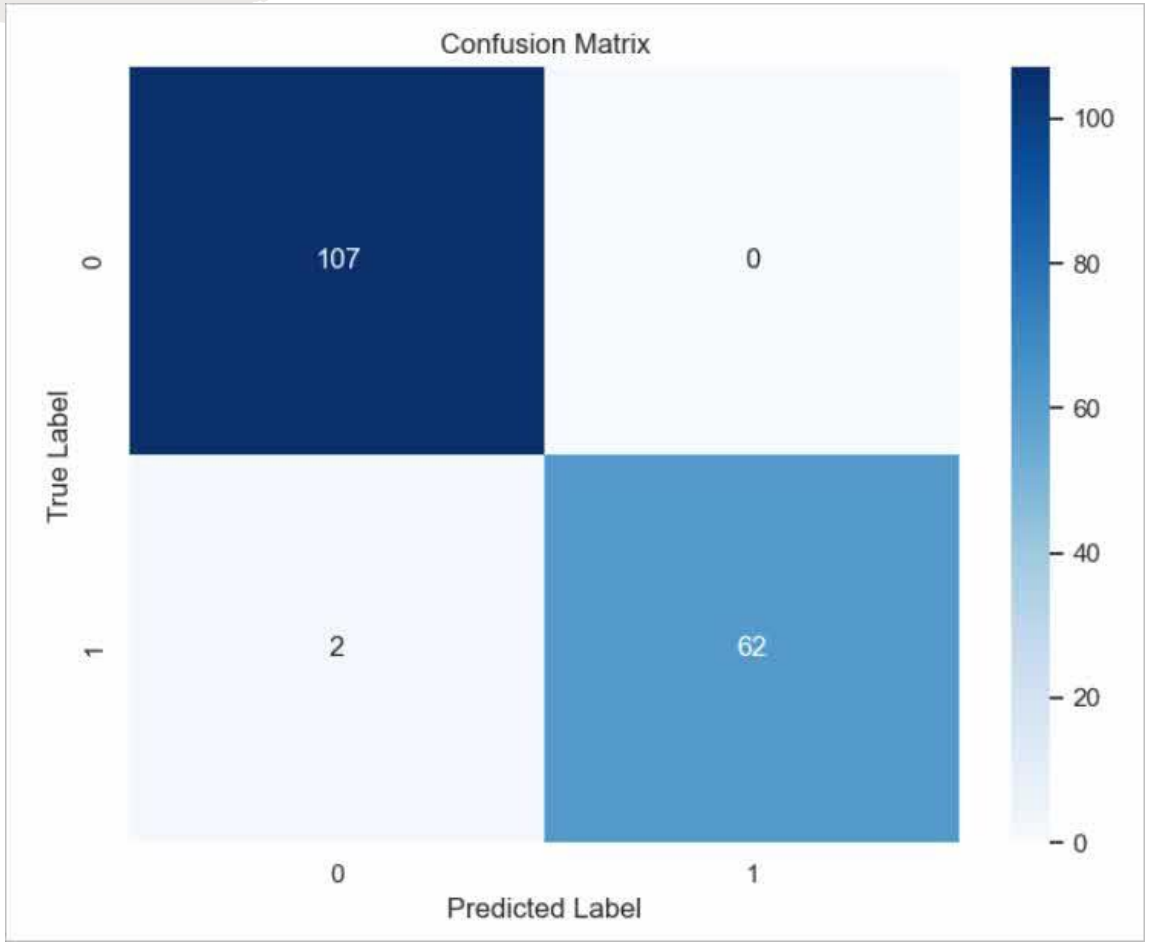
	Model	Accuracy	Precision	Recall	F1	AUC
0	XGBoost (Tuned)	1.000000	1.000000	1.000000	1.000000	1.000000
3	Voting Ensemble (Top-3 Tuned)	0.992462	1.000000	0.979730	0.989761	0.999973
1	SVM (RBF Kernel) (Tuned)	0.989950	1.000000	0.972973	0.986301	0.999676
2	Logistic Regression (Tuned)	0.989950	1.000000	0.972973	0.986301	0.996514

THE WINNER: XGBOOST

WINNING MODEL ON TEST DATA

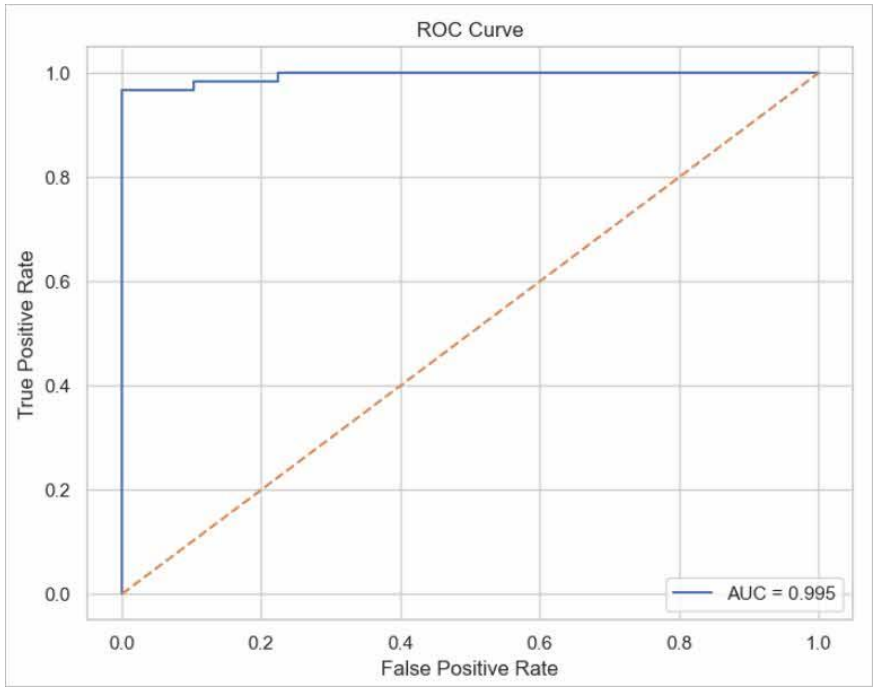
CONFUSION MATRIX

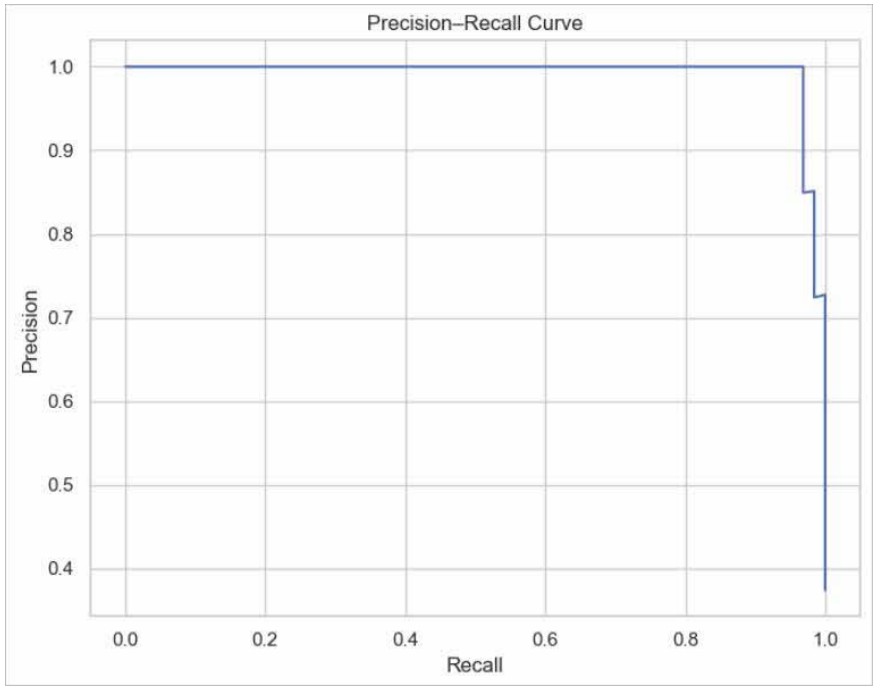
- The model never falsely classifies a benign case as malignant
- Two malignant cases misclassified out of 64 , low but clinically important



ROC-AUC & PRECISION-RECALL

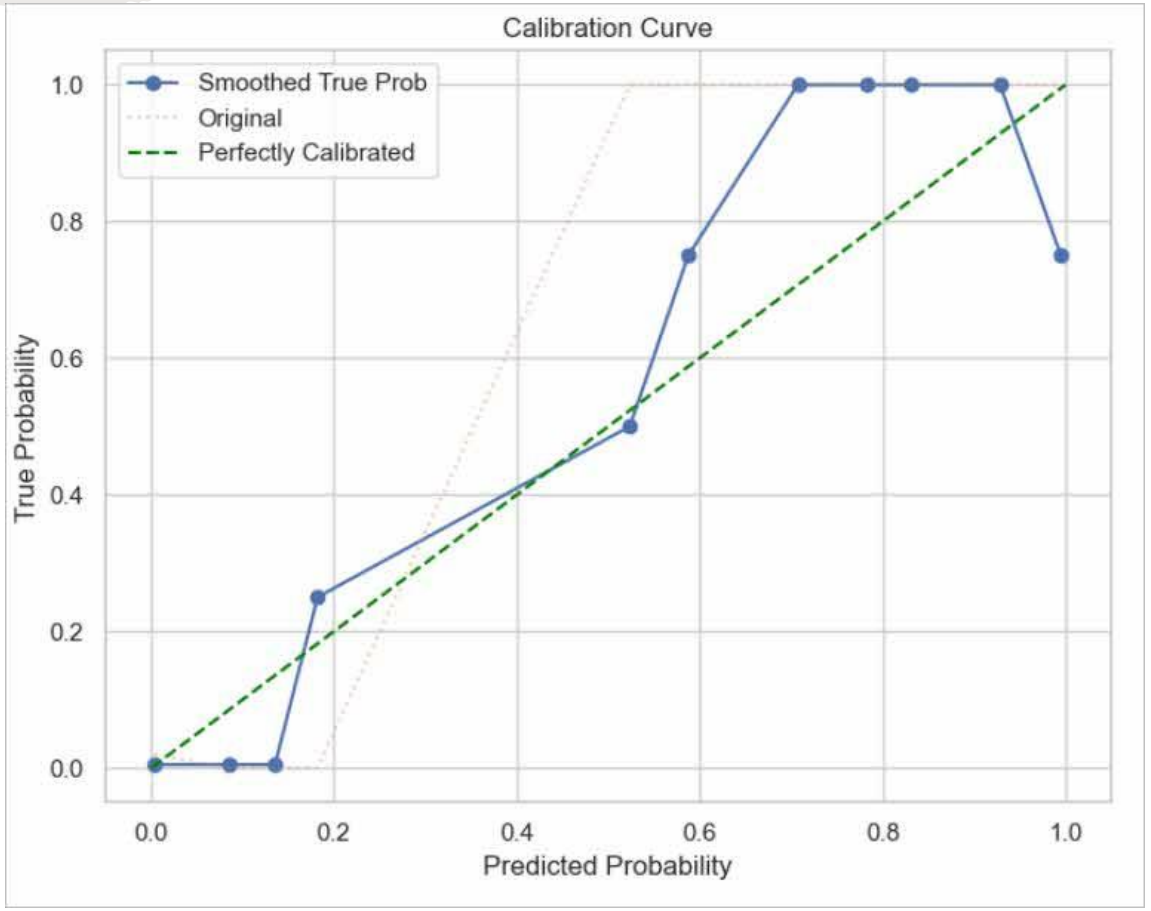
- ROC-AUC = 0.995
- Almost perfect ranking
- Excellent for threshold tuning
- Precision is near 1.0 across most recall levels





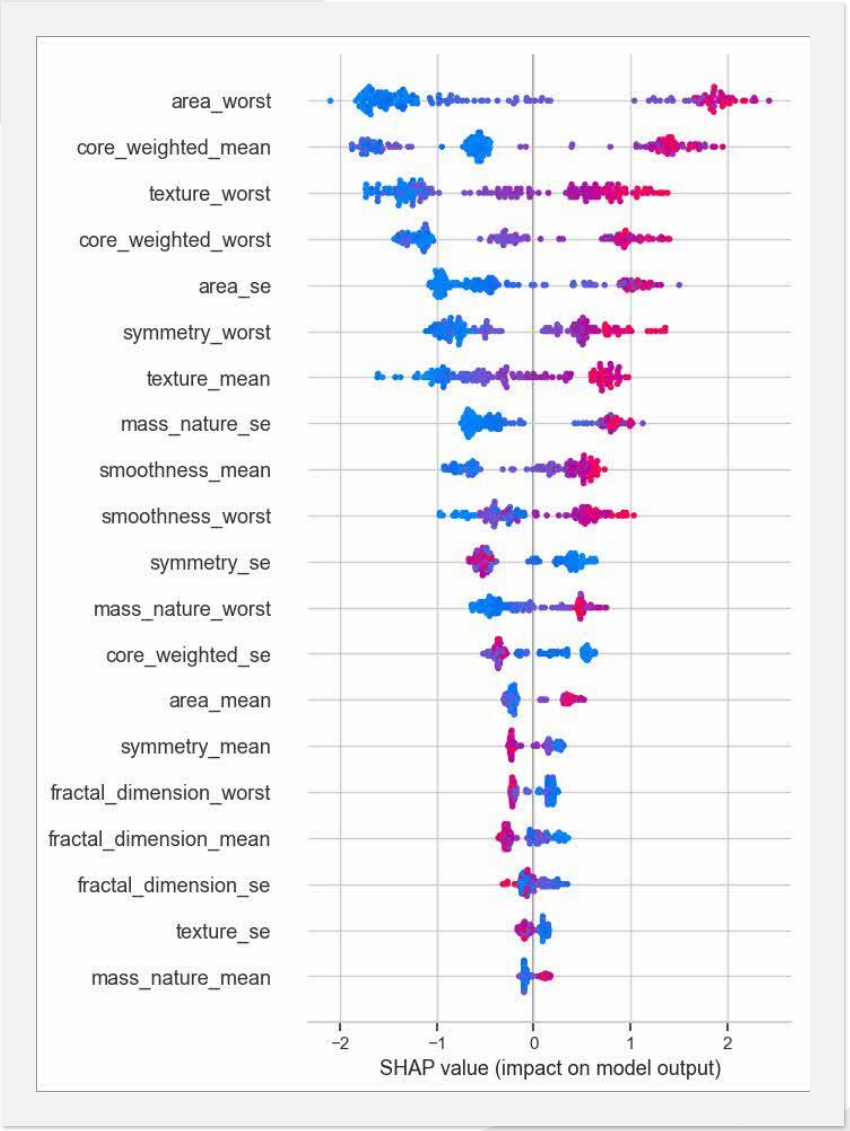
CALIBRATION CURVE

- Small number if instances available in each probability bin
- Good practical calibration
- Sample size limitations



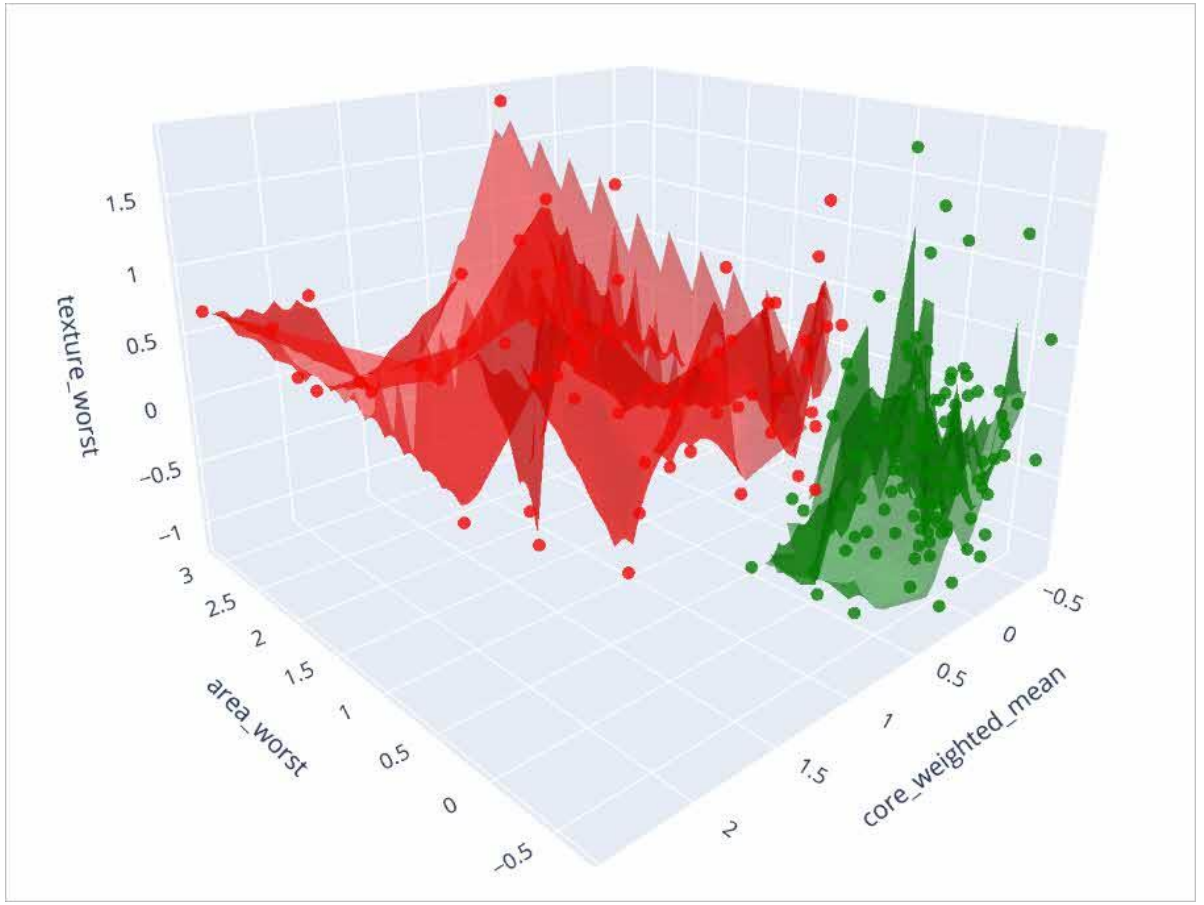
SHAP EXPLAINABILITY

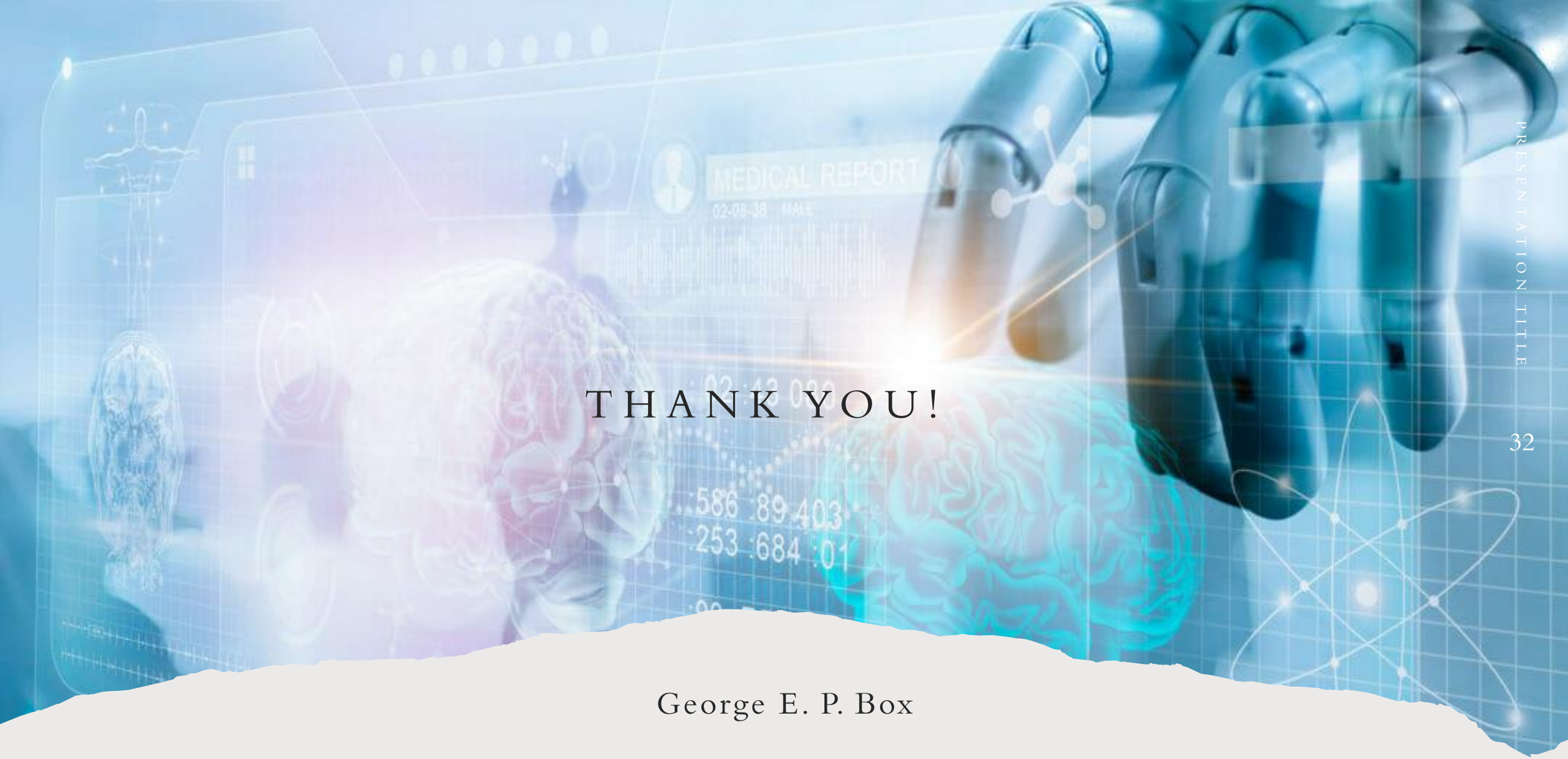
- Applied to the tuned XGBoost model.
- Larger, more irregular, and more texturally complex tumours are more likely to be malignant.
- This is clinically intuitive and increases trust in the model.



3D SCATTER MESH WITH TOP 3 FEATURES

- Benign and malignant points forming distinct, non-overlapping manifolds.
- The classification problem is geometrically well-posed.
- The selection features capture the essential decision boundary.
- Engineered feature made it to top 3





THANK YOU!

George E. P. Box

ALL MODELS ARE WRONG, SOME ARE USEFUL.