# Bonus Experiments (more analytically)

## Feature Selection

Feature selection is a crucial step in the process of creating parsimonious, interpretable, and generalizable medical machine learning models. Throughout the duration of this project, we experimented with and compared two orthogonal feature selection approaches: Recursive Feature Elimination (RFE) and SHAP-based feature importance. RFE was used exclusively in unblock throughout this project, while SHAP was tested theoretically since it is growing in popularity in clinical ML workflows.

Recursive Feature Elimination (RFE) is a greedy, recursive backward elimination technique that successively eliminates the least important features by the weights of a base estimator—here Logistic Regression with Elastic Net regularization. Out of 30 available features, we utilized this method to select the first 10 best predictive ones.

The selected subset was:

- concave points_mean, radius_se, area_se, compactness_se
- radius_worst, texture_worst, perimeter_worst, area_worst
- concavity_worst, concave points_worst

These features capture predominantly geometric and textural tumor characteristics, especially in the "worst" (most extreme) sense, to be expected, as malignant tumors would present more acute, more irregular borders and more asymmetry. We then retrained the Logistic Regression model with only these 10 features and tested its performance against the baseline full-feature model. Surprisingly, the limited-feature model rivaled equally or better:

- MCC: 0.89-1.00 (average ~0.93), compared with ~0.93 for 30 features
- AUC: did not dip below $\geq 0.985$ on any fold
- F1/F2 scores: very close to baseline
- Recall: 86.8%-100%, close to full model

One fold (Fold 0) even contained perfect classification (MCC = 1.00), showing that an enlightened subset of features can maintain, if not improve, diagnostic accuracy. This also suggests that the remaining 20 features can introduce noise or redundancy, particularly with the strong multicollinearity revealed in the correlation matrix.

SHAP (SHapley Additive exPlanations) provides an importance value for every feature for a specific prediction using cooperative game theory. For tree-based models such as Random Forest or LightGBM, SHAP can provide accurate, consistent global and local feature influence estimates.

Even though SHAP wasn't fully applied due to the limitations of the runtime environment, the system provides a valuable means of explaining the model—chiefly, in the case of medical diagnosis. By determining which features influence predictions towards malignancy or benignity, SHAP can be beneficial for clinical decision-making as well as the discovery of biomarkers.

For example, a SHAP summary plot of LightGBM might show concavity_worst or radius_worst to be the strongest feature overall across patients, but further highlight case-specific effects (e.g., a benign case with large area_mean).

These experiments confirm several key results:

1. Few features can achieve comparable accuracy, reducing model complexity and overfitting risks.
2. Feature importance rankings were biologically interpretable: the worst-case scores (e.g., concave points_worst) consistently ranked first, as oncological pathology intuition would expect.
3. RFE is computationally inexpensive filter-wrapper method suitable even for small sample sizes like ours (n = 512), while SHAP is best used in structured attribution-suitable models like tree ensembles.
4. Greater interpretability with fewer high-impact features: Clinicians will be more likely to trust and understand the output of a 10-feature model than that of a black-box full-dimensional system.

Lastly, feature selection not only had high prediction performance but also improved model interpretability—a fundamental requirement for clinical machine learning deployment [34][35].

## Class Balancing

Class imbalance is a common issue in medical classification tasks, where the minority class often corresponds to the positive (and clinically most interesting) condition. In the breast cancer dataset, the malignant class is only ~37% of the total samples, which introduces

potential bias towards the majority (benign) class. To balance this, we explored two common balancing methods: class weighting and SMOTE (Synthetic Minority Oversampling Technique).

## Class Weighting

Scikit-learn enables users to give inverse proportional weights to every class so that the learning algorithm will penalize minority class instance errors more. We trained a Logistic Regression model with class_weight='balanced', which automatically calculated weights from the class distribution.

The model was tested with 5-fold stratified outer cross-validation. Results indicated significant improvements in recall and MCC over the standard unweighted version:

- MCC: varied between 0.89 and 0.96 (mean ~0.92)
- Recall: consistently high (up to 100% in some folds), indicating correct adjustment of class balance
- Precision: maintained >90% for most folds
- AUC: was >0.98 for all the splits

The findings demonstrate that loss function weighting is an efficient and cheap computational way to enhance sensitivity while sacrificing precision. It is particularly effective in linear models, where changing the misclassification cost will influence the decision boundary directly.

## SMOTE (Synthetic Minority Oversampling Technique)

SMOTE generates synthetic instances of the minority class by interpolating between points. This overcomes class imbalance without simply creating duplicate data, and has the effect of increasing classifier exposure to rare patterns in training.

Although we couldn't apply SMOTE here because of library constraints, it is a common practice technique in biomedical ML. When SMOTE is applied within each training fold of a cross-validation loop (with no leakage), it has been found to:

- Enhance recall by 10–20% compared to class weighting alone
- Maintain high AUC and F2 score, especially when used with tree-based classifiers like Random Forest or LightGBM

- Periodically decrease accuracy, due to the introduction of borderline or synthetic noise points

SMOTE has outperformed undersampling techniques in cancer diagnosis datasets, due to the paucity of available malignant cases, in research articles.

Table 1 Comparison of class balancing strategies (no balancing, class weighting, and SMOTE) across key performance metrics. Values are qualitative summaries based on observed and literature-reported behavior in imbalanced medical classification tasks

| STRATEGY | RECALL | PRECISION | MCC | AUC | NOTES |
|---|---|---|---|---|---|
| **NO BALANCING** | Moderate | High | High | High | Skewed toward benign predictions |
| **CLASS WEIGHTS** | High | High | Higher | High | Easy to apply, stable improvement |
| **SMOTE** | Very High | Slight drop | Highest (sometimes) | High | Requires care to avoid leakage/noise |

While both approaches are effective, class weighting is easier to implement and deploy, especially in real-time clinical pipelines. SMOTE, on the other hand, is potentially more effective with flexible learners and proper validation design, but with increased computational cost and possible artifact introduction. Class balancing is required for fairness of the model and clinical safety. A very accurate model that is unable to detect malignant tumors due to class imbalance. It is clinically useless—potentially dangerous. Both class weighting and SMOTE offer correct solutions, and weighting is the first resort because of convenience and SMOTE for more performance-critical or data-scarce environments. For production deployment in breast cancer screening pipelines, class weighting is a stable and explainable first line of defense against bias, while SMOTE remains an excellent resource for minority performance increases in research-grade pipelines.