Machine Learning Analysis Report: Tumor Tissue Classification

This comprehensive report presents a machine learning analysis for tumor tissue classification. The study compares Logistic Regression, Random Forest, and XGBoost models using rigorous hyperparameter tuning, threshold optimization, feature importance analysis, and SHAP explainability. All models were evaluated for clinical applicability with emphasis on recall (sensitivity) for medical screening applications.

# Key Definitions and Mathematical Formulas

This section provides essential definitions and mathematical formulas for statistical concepts and techniques used throughout this analysis.

## Performance Metrics

Classification performance is evaluated using the following metrics, derived from the confusion matrix:

Confusion Matrix Components:  
• True Positives (TP): Correctly identified malignant cases  
• True Negatives (TN): Correctly identified benign cases  
• False Positives (FP): Incorrectly identified as malignant (Type I error)  
• False Negatives (FN): Incorrectly identified as benign (Type II error)

Accuracy: Overall correctness of the model  
Formula: Accuracy = (TP + TN) / (TP + TN + FP + FN)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Proportion of all cases correctly classified

Precision (Positive Predictive Value): Reliability of positive predictions  
Formula: Precision = TP / (TP + FP)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Of all cases predicted as malignant, what proportion actually are malignant?

Recall (Sensitivity): Ability to detect positive cases  
Formula: Recall = TP / (TP + FN)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Of all actual malignant cases, what proportion are correctly identified? Critical for medical screening to minimize missed diagnoses.

Specificity: Ability to correctly identify negative cases  
Formula: Specificity = TN / (TN + FP)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Of all actual benign cases, what proportion are correctly identified?

F1-Score: Harmonic mean of precision and recall  
Formula: F1 = 2 × (Precision × Recall) / (Precision + Recall)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Balanced measure when both precision and recall are equally important

F2-Score: Weighted harmonic mean favoring recall  
Formula: F2 = 5 × (Precision × Recall) / (4 × Precision + Recall)  
Range: 0 to 1 (higher is better)  
Clinical interpretation: Emphasizes recall over precision, preferred for medical screening where missing a positive case is more costly than a false alarm

Youden's J Index: Optimal balance between sensitivity and specificity  
Formula: J = Sensitivity + Specificity - 1  
Range: -1 to 1 (higher is better, 0 = random performance)  
Clinical interpretation: Maximizes the difference between true positive rate and false positive rate. Used to determine optimal classification thresholds.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve):  
Definition: Area under the curve plotting True Positive Rate vs False Positive Rate at various thresholds  
Range: 0 to 1 (0.5 = random, 1.0 = perfect)  
Clinical interpretation: Model's ability to distinguish between classes across all thresholds. AUC > 0.8 generally considered good, > 0.9 excellent for clinical applications.

## Dimensionality Reduction Techniques

Principal Component Analysis (PCA):  
Definition: Linear dimensionality reduction technique that projects data onto principal components that capture maximum variance.  
Mathematical basis: Eigenvalue decomposition of covariance matrix  
Purpose: Identify the most informative linear combinations of features  
Clinical interpretation: Reveals which feature combinations explain most variation in the data, helping identify key biomarker patterns.

t-Distributed Stochastic Neighbor Embedding (t-SNE):  
Definition: Non-linear dimensionality reduction technique that preserves local structure  
Purpose: Visualize high-dimensional data in 2D/3D while maintaining similarity relationships  
Mathematical basis: Minimizes divergence between probability distributions in high and low dimensions  
Clinical interpretation: Reveals natural clustering patterns and separability between classes that may not be apparent in linear projections.

## Machine Learning Algorithms

Logistic Regression:  
Definition: Linear classifier using logistic function to model probability of binary outcomes  
Formula: P(y=1|x) = 1 / (1 + e^(-(β₀ + β₁x₁ + β₂x₂ + ... + βₙxₙ)))  
Advantages: Interpretable coefficients, fast training, probabilistic outputs  
Clinical application: Provides interpretable feature weights indicating biomarker importance

Random Forest:  
Definition: Ensemble of decision trees using bootstrap aggregating (bagging)  
Method: Trains multiple decision trees on random subsets of data and features  
Prediction: Averages predictions across all trees (regression) or majority vote (classification)  
Advantages: Handles non-linear relationships, feature interactions, robust to overfitting  
Clinical application: Captures complex biomarker interactions without assuming linear relationships

XGBoost (Extreme Gradient Boosting):  
Definition: Gradient boosting framework that sequentially builds weak learners  
Method: Each new tree corrects errors made by previous trees  
Optimization: Minimizes regularized objective function using gradient descent  
Advantages: High performance, handles missing values, built-in regularization  
Clinical application: Often achieves highest accuracy by learning complex patterns through iterative error correction

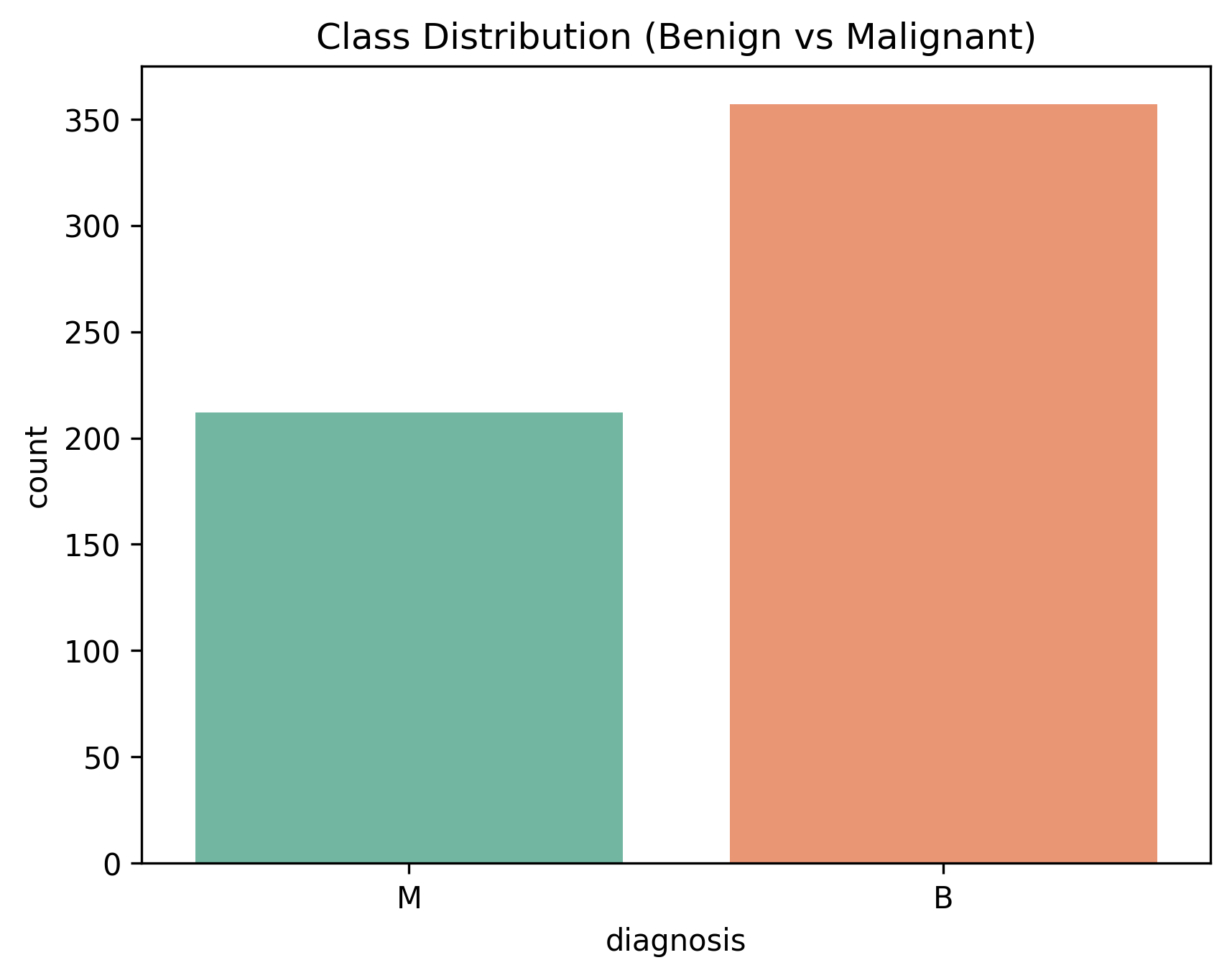
## Explainability Methods

SHAP (SHapley Additive exPlanations):  
Definition: Game-theoretic approach to explain individual predictions  
Mathematical basis: Shapley values from cooperative game theory  
Formula: φᵢ = Σ |S|!(|N|-|S|-1)! / |N|! × [f(S ∪ {i}) - f(S)]  
where φᵢ is feature i's SHAP value, S is subset of features, N is all features  
Properties: Additivity, symmetry, dummy feature, efficiency  
Clinical interpretation: Shows exactly how each biomarker contributes to individual patient predictions, enabling transparent and trustworthy AI-assisted diagnosis.

# 1. Data Overview and Exploratory Analysis

## 1.1 Class Distribution

Distribution of target classes in the dataset:



Sample data preview (first 5 rows):

Sample data preview (showing first 5 rows and first 6 columns out of 33 total columns):

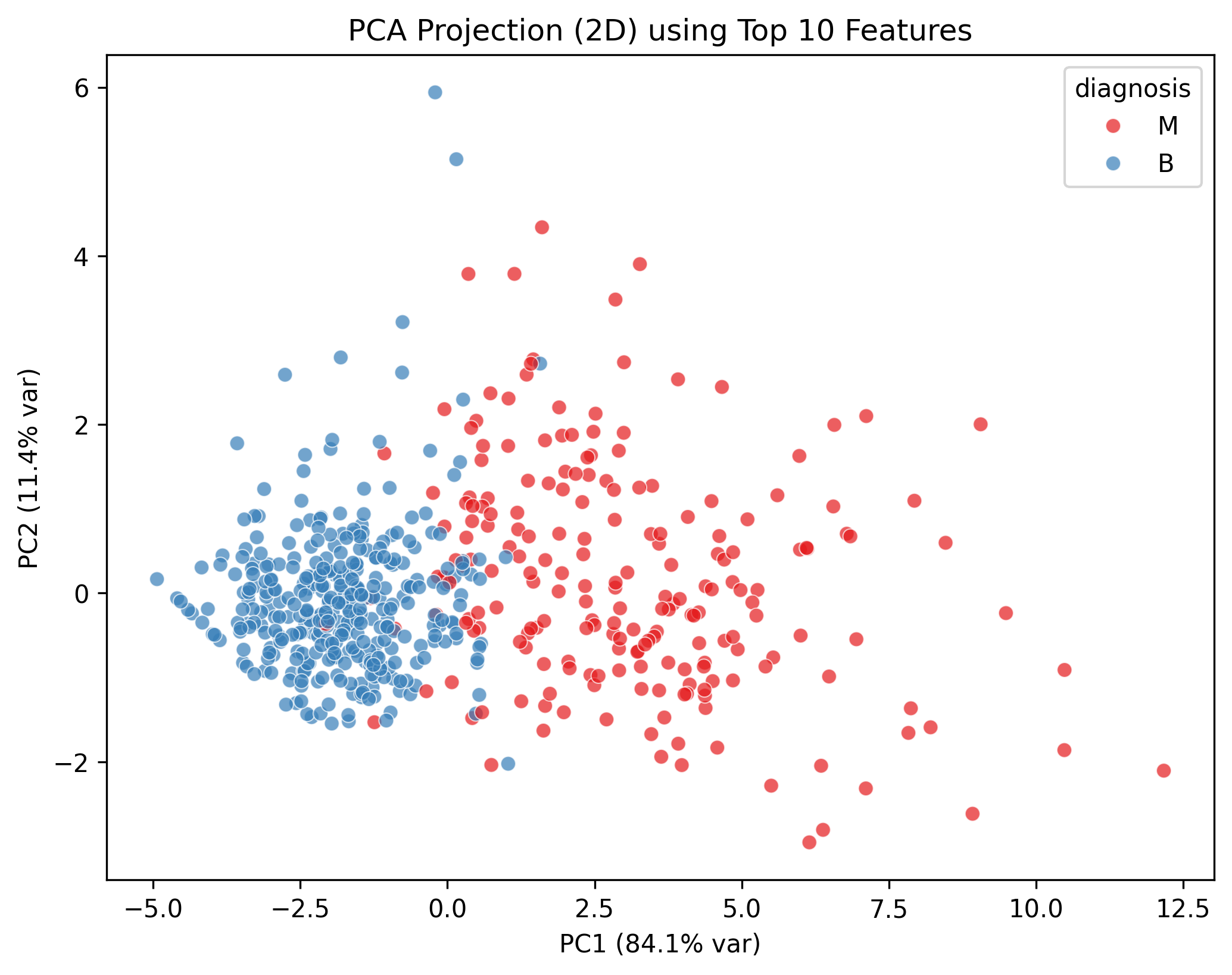
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **id** | **diagnosis** | **radius\_mean** | **texture\_mean** | **perimeter\_mean** | **area\_mean** |
| 842302 | M | 17.99 | 10.38 | 122.8 | 1001.0 |
| 842517 | M | 20.57 | 17.77 | 132.9 | 1326.0 |
| 84300903 | M | 19.69 | 21.25 | 130.0 | 1203.0 |
| 84348301 | M | 11.42 | 20.38 | 77.58 | 386.1 |
| 84358402 | M | 20.29 | 14.34 | 135.1 | 1297.0 |

Class distribution statistics:

|  |  |
| --- | --- |
| **diagnosis** | **count** |
| B | 357 |
| M | 212 |

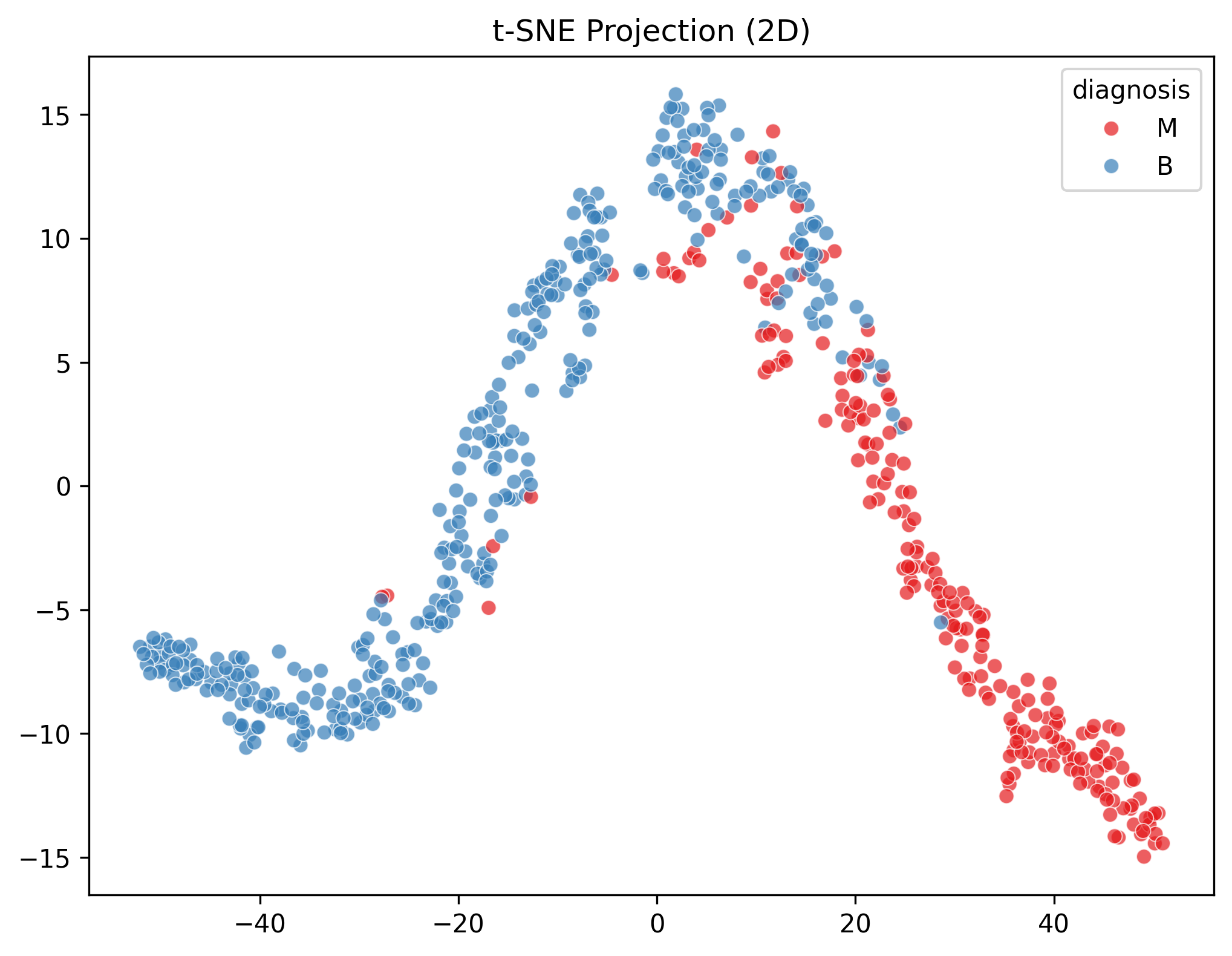
## 1.2 Principal Component Analysis

PCA visualization of the top 10 predictive features:



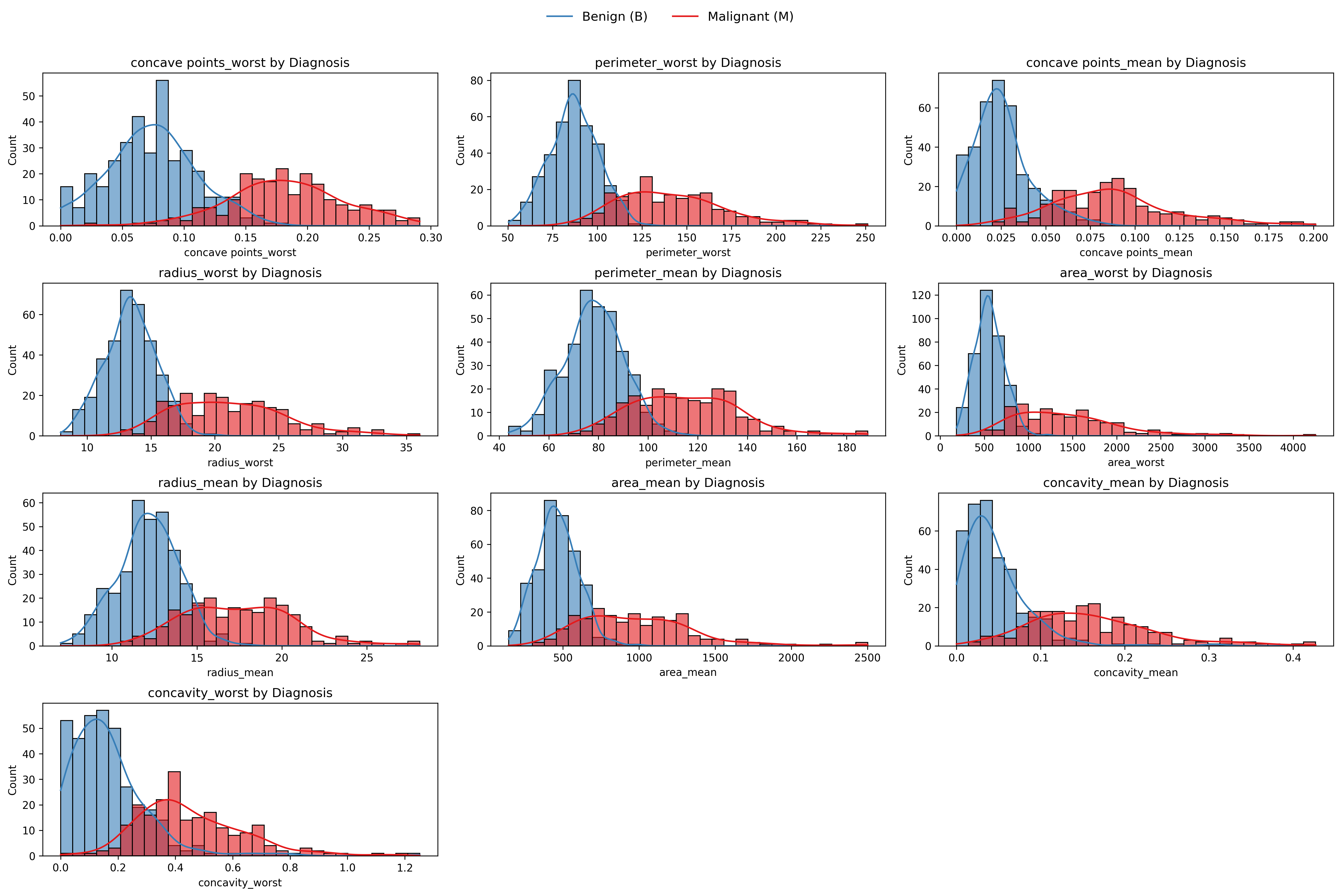
## 1.3 t-SNE Projection

t-SNE dimensional reduction visualization:



## 1.4 Feature Distributions

Distribution of the top 10 most predictive features:



# 2. Model Development and Hyperparameter Tuning

Three machine learning algorithms were evaluated: Logistic Regression, Random Forest, and XGBoost. Each model underwent extensive hyperparameter tuning using 5-fold cross-validation with ROC-AUC as the optimization metric.

Hyperparameter tuning results:

|  |  |  |
| --- | --- | --- |
| **Model** | **Best Params** | **Best ROC-AUC** |
| Logistic Regression | {'clf\_\_C': 10, 'clf\_\_penalty':... | 0.9861 |
| Random Forest | {'clf\_\_max\_depth': None, 'clf\_... | 0.9847 |
| XGBoost | {'clf\_\_colsample\_bytree': 0.8,... | 0.9831 |

Detailed hyperparameter tuning metrics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **CV\_ROC\_AUC** | **Best\_Params** | **Hyperparameter\_Search\_Candidates** | **Mean\_Fit\_Time** | **Mean\_Score\_Time** |
| Logistic Regression | 0.9861 | {'clf\_\_C': 10, 'clf\_\_penalty':... | 8 | 0.0385 | 0.0062 |
| Random Forest | 0.9847 | {'clf\_\_max\_depth': None, 'clf\_... | 48 | 0.5760 | 0.0297 |
| XGBoost | 0.9831 | {'clf\_\_colsample\_bytree': 0.8,... | 72 | 0.1559 | 0.0096 |

# 3. Model Performance Evaluation

## 3.1 Overall Performance Metrics

Key performance metrics for all models:

(Showing first 6 columns of 7 total)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1\_Score** |
| Logistic Regression | 0.9649 | 0.9524 | 0.9524 | 0.9722 | 0.9524 |
| Random Forest | 0.9737 | 1.00 | 0.9286 | 1.00 | 0.9630 |
| XGBoost | 0.9912 | 1.00 | 0.9762 | 1.00 | 0.9880 |

Comprehensive model results including threshold analysis:

Table summary (original size: 3 rows × 18 columns):

###### Row 1:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | Logistic Regression |
| Accuracy | 0.9649 |
| Precision | 0.9524 |
| Recall | 0.9524 |
| Specificity | 0.9722 |
| F1\_Score | 0.9524 |
| ROC\_AUC | 0.9983 |
| CV\_ROC\_AUC | 0.9861 |
| Best\_Youden\_Threshold | 0.2560 |
| Best\_F1\_Threshold | 0.5570 |
| Best\_F2\_Threshold | 0.2560 |
| Youden\_Score | 0.9583 |
| Max\_F1\_Score | 0.9756 |
| Max\_F2\_Score | 0.9859 |
| Hyperparameter\_Search\_Candidates | 8 |
| Best\_Params | {'clf\_\_C': 10, 'clf\_\_penalty': 'l2', 'clf\_\_solver': 'saga'} |
| Mean\_Fit\_Time | 0.0385 |
| Mean\_Score\_Time | 0.0062 |

###### Row 2:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | Random Forest |
| Accuracy | 0.9737 |
| Precision | 1.00 |
| Recall | 0.9286 |
| Specificity | 1.00 |
| F1\_Score | 0.9630 |
| ROC\_AUC | 0.9947 |
| CV\_ROC\_AUC | 0.9847 |
| Best\_Youden\_Threshold | 0.4720 |
| Best\_F1\_Threshold | 0.4720 |
| Best\_F2\_Threshold | 0.1330 |
| Youden\_Score | 0.9286 |
| Max\_F1\_Score | 0.9630 |
| Max\_F2\_Score | 0.9677 |
| Hyperparameter\_Search\_Candidates | 48 |
| Best\_Params | {'clf\_\_max\_depth': None, 'clf\_\_max\_features': 'sqrt', 'clf\_\_min\_samples\_leaf': 2, 'clf\_\_min\_samples\_split': 5, 'clf\_\_n\_estimators': 200} |
| Mean\_Fit\_Time | 0.5760 |
| Mean\_Score\_Time | 0.0297 |

###### Row 3:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | XGBoost |
| Accuracy | 0.9912 |
| Precision | 1.00 |
| Recall | 0.9762 |
| Specificity | 1.00 |
| F1\_Score | 0.9880 |
| ROC\_AUC | 0.9977 |
| CV\_ROC\_AUC | 0.9831 |
| Best\_Youden\_Threshold | 0.3600 |
| Best\_F1\_Threshold | 0.3600 |
| Best\_F2\_Threshold | 0.3600 |
| Youden\_Score | 0.9762 |
| Max\_F1\_Score | 0.9880 |
| Max\_F2\_Score | 0.9809 |
| Hyperparameter\_Search\_Candidates | 72 |
| Best\_Params | {'clf\_\_colsample\_bytree': 0.8, 'clf\_\_learning\_rate': 0.05, 'clf\_\_max\_depth': 3, 'clf\_\_n\_estimators': 100, 'clf\_\_subsample': 1.0} |
| Mean\_Fit\_Time | 0.1559 |
| Mean\_Score\_Time | 0.0096 |

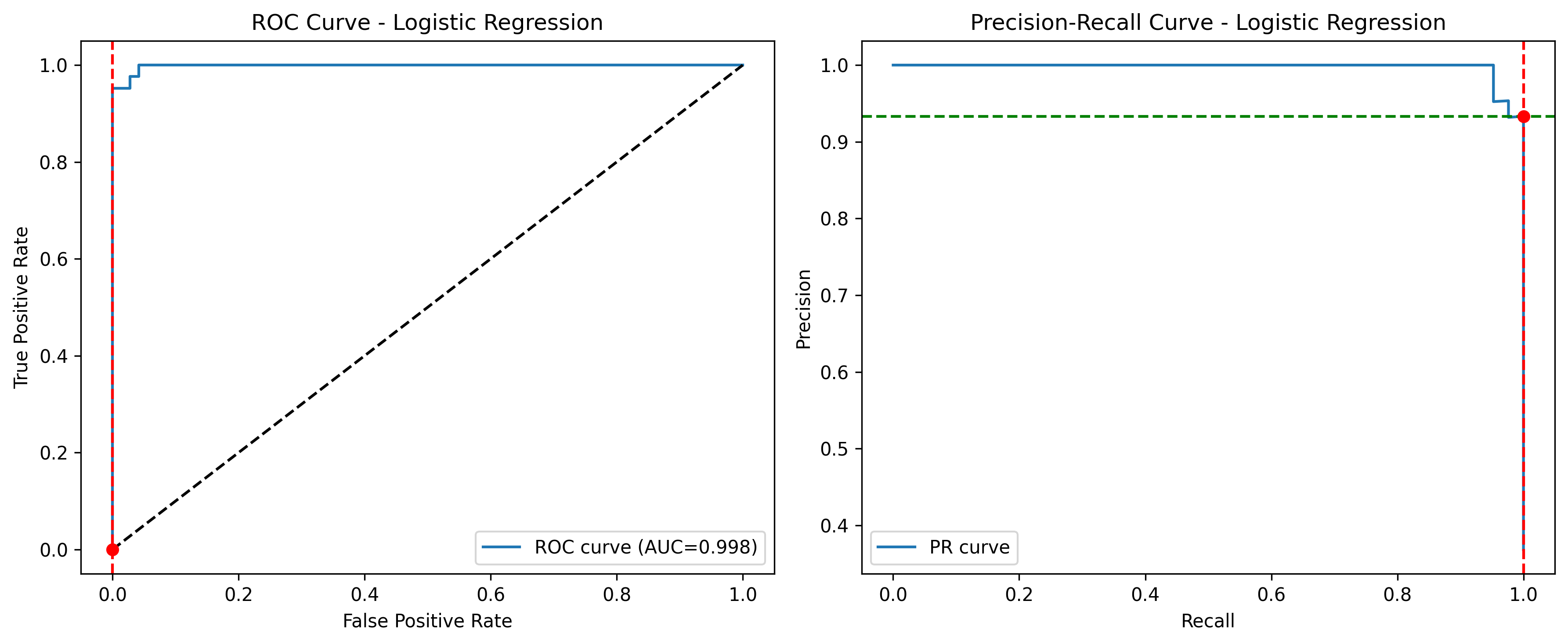
Model rankings by different performance metrics:

(Showing first 6 columns of 7 total)

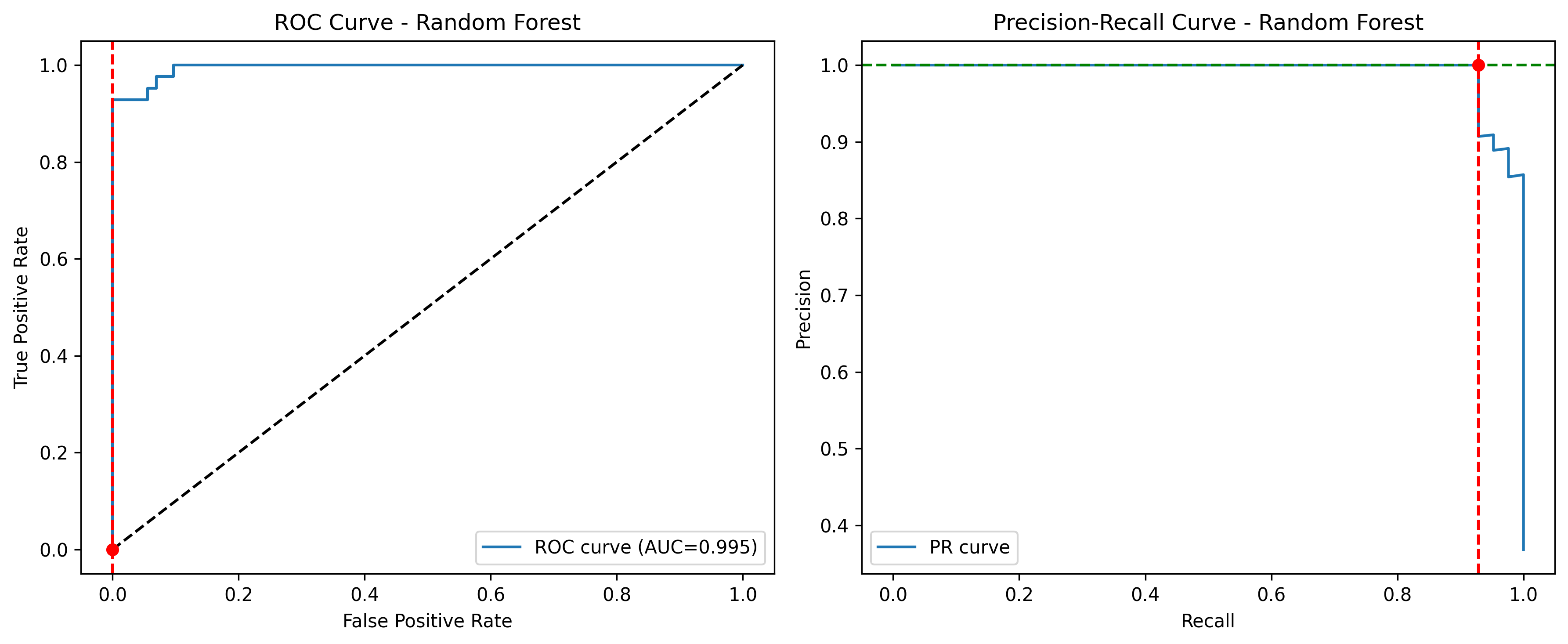
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Unnamed: 0** | **Accuracy\_Rank** | **Precision\_Rank** | **Recall\_Rank** | **Specificity\_Rank** | **F1\_Score\_Rank** |
| 1st Place | XGBoost | XGBoost | XGBoost | XGBoost | XGBoost |
| 2nd Place | Random Forest | Random Forest | Logistic Regression | Random Forest | Random Forest |
| 3rd Place | Logistic Regression | Logistic Regression | Random Forest | Logistic Regression | Logistic Regression |

## 3.2 ROC and Precision-Recall Curves

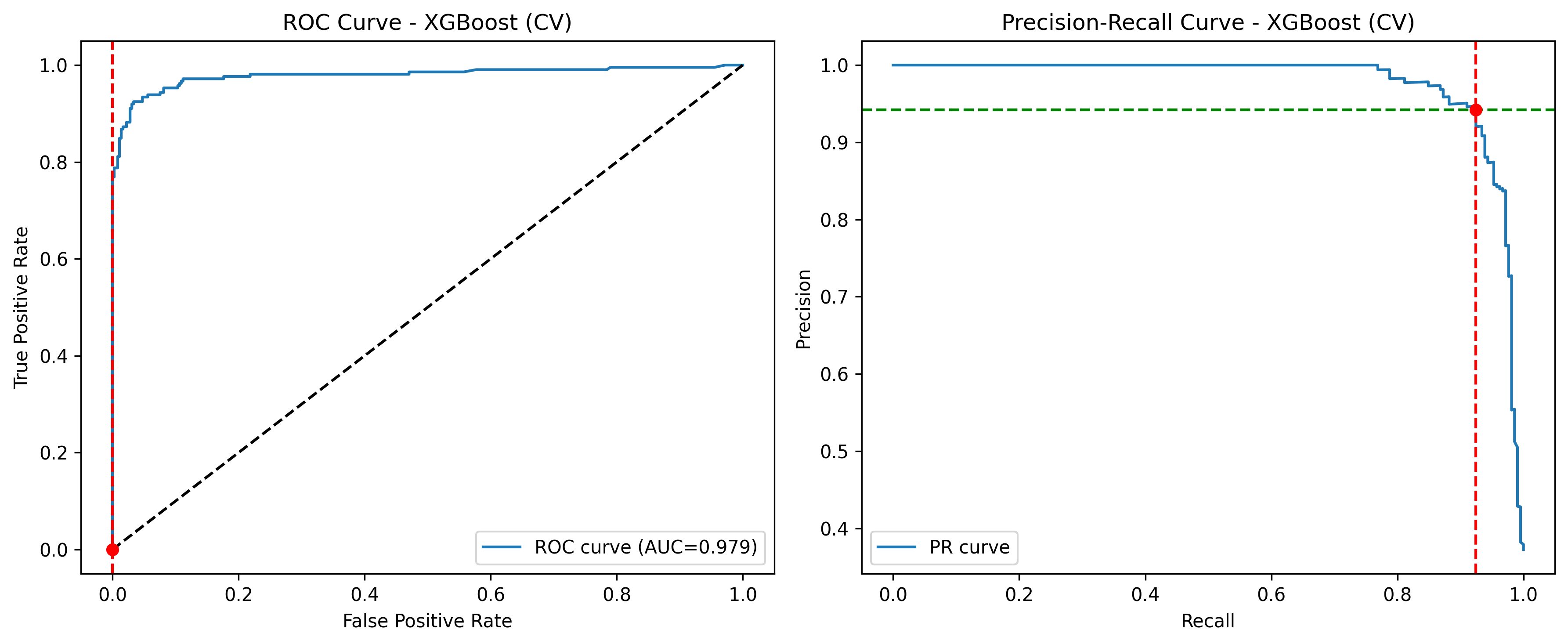
Logistic Regression ROC and Precision-Recall curves:



Random Forest ROC and Precision-Recall curves:



XGBoost ROC and Precision-Recall curves:



# 4. Threshold Optimization Analysis

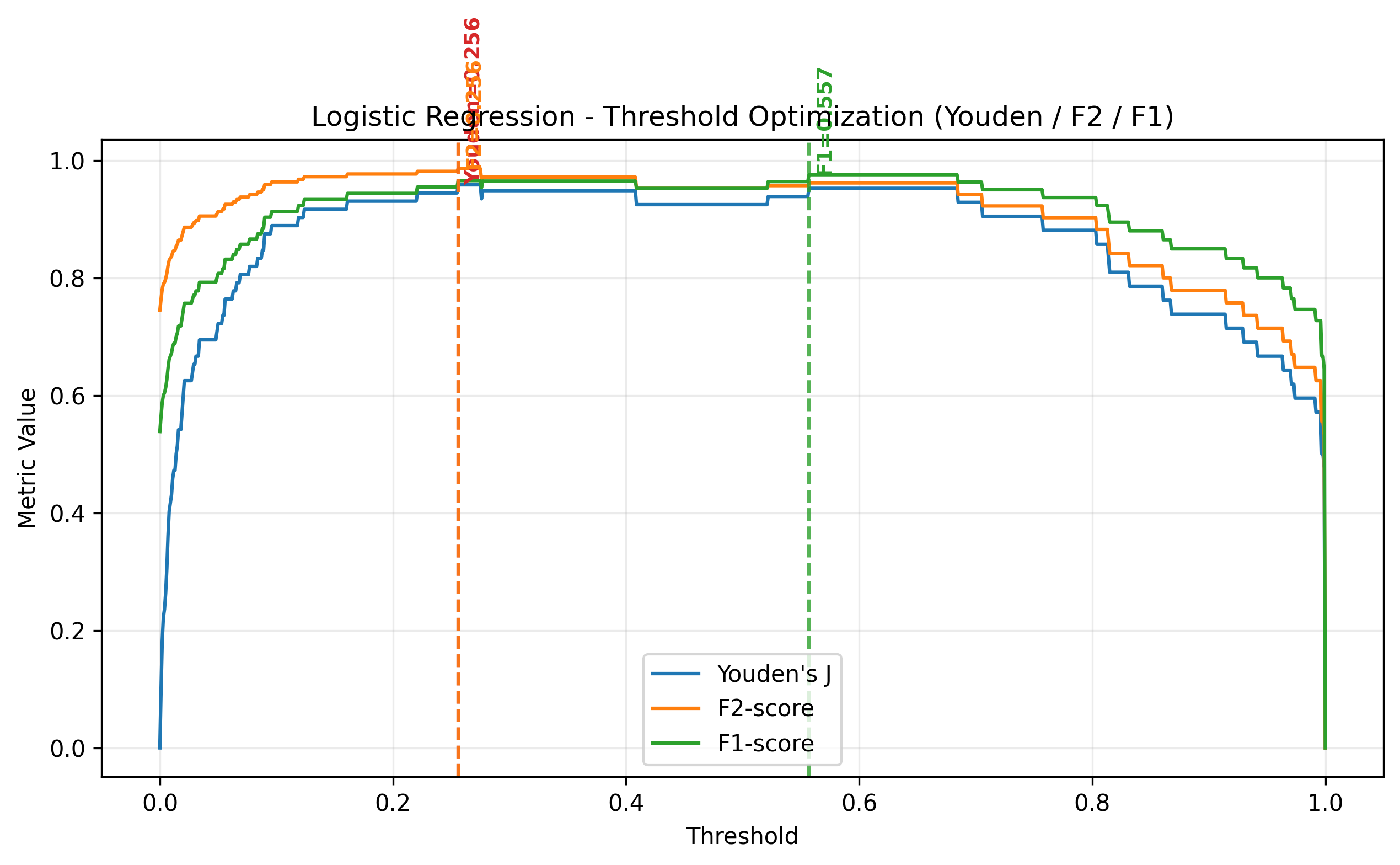
Optimal classification thresholds were determined using three metrics:  
  
• F1-score: Harmonic mean of precision and recall  
• F2-score: Weighted harmonic mean favoring recall  
• Youden's J: Sensitivity + Specificity - 1 (optimal balance)

Optimal thresholds for each model:

(Showing first 6 columns of 7 total)

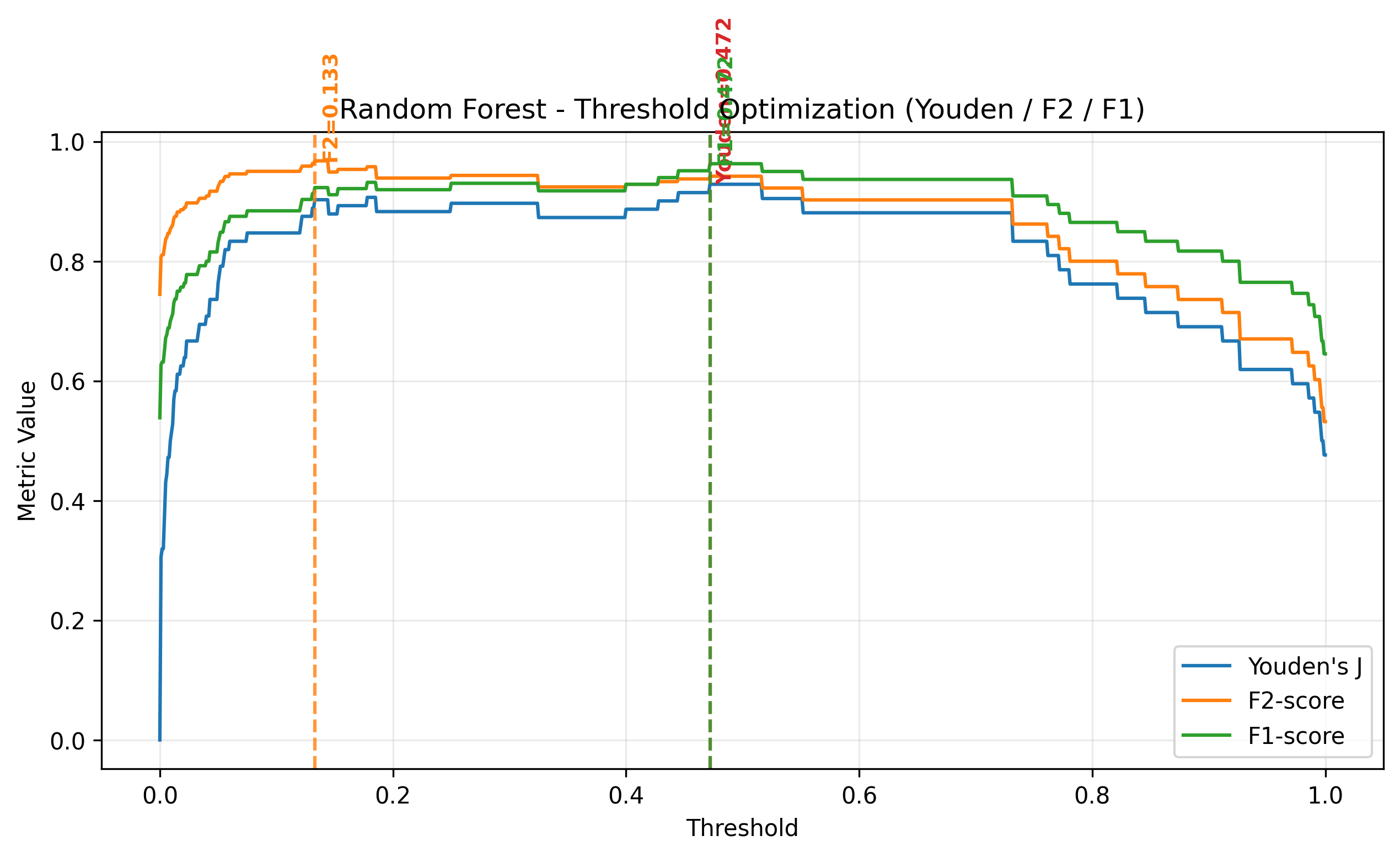
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Best\_Youden\_Threshold** | **Best\_F1\_Threshold** | **Best\_F2\_Threshold** | **Youden\_Score** | **Max\_F1\_Score** |
| Logistic Regression | 0.2560 | 0.5570 | 0.2560 | 0.9583 | 0.9756 |
| Random Forest | 0.4720 | 0.4720 | 0.1330 | 0.9286 | 0.9630 |
| XGBoost | 0.3600 | 0.3600 | 0.3600 | 0.9762 | 0.9880 |

Logistic Regression threshold optimization:



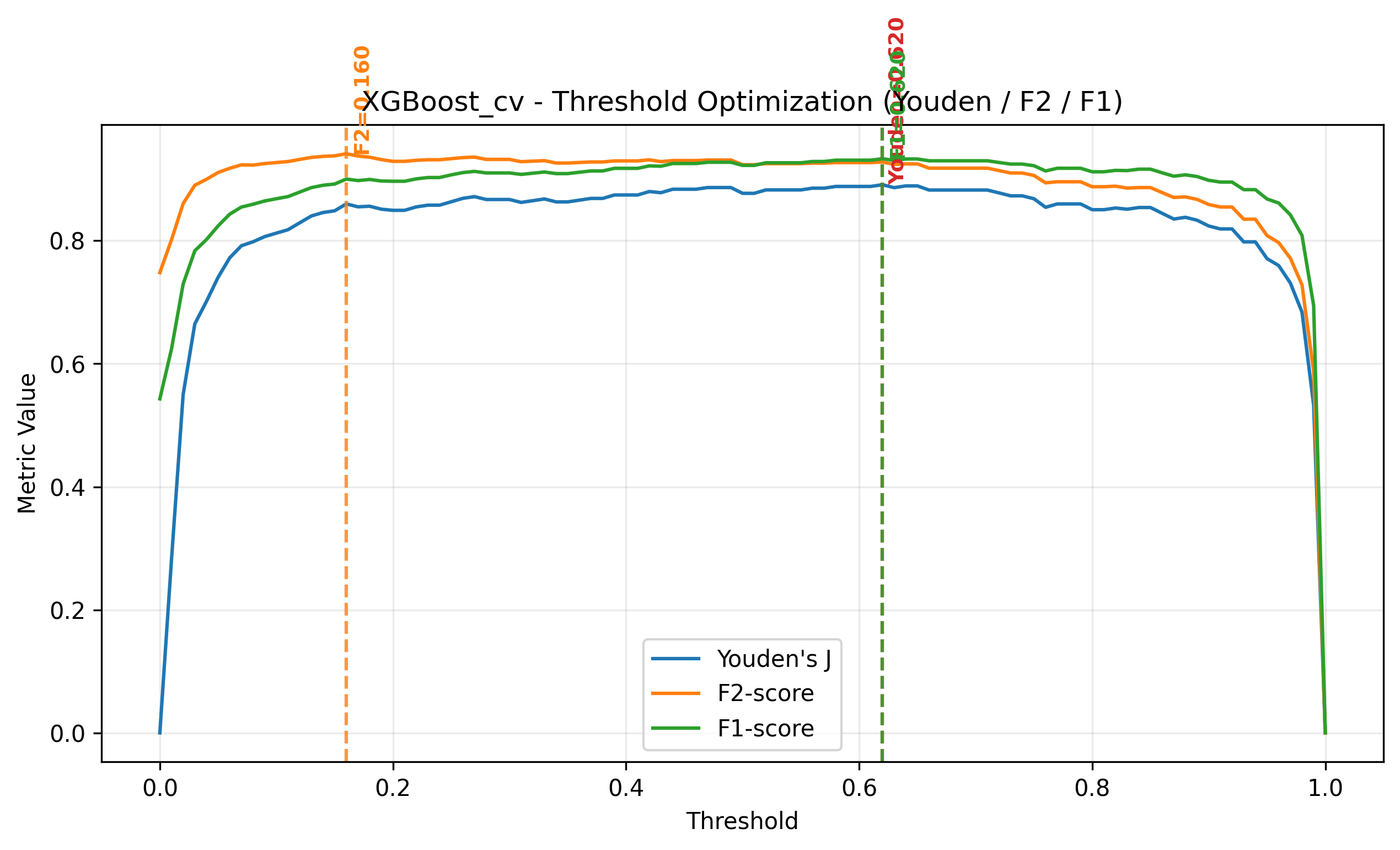
This plot shows how different classification thresholds affect Logistic Regression performance metrics. The optimal thresholds are determined by maximizing F1-score (balanced precision-recall), F2-score (recall-focused), and Youden's J index (sensitivity-specificity balance). For medical screening, F2-score and recall optimization are typically preferred to minimize missed diagnoses.

Random Forest threshold optimization:



This plot shows how different classification thresholds affect Random Forest performance metrics. The optimal thresholds are determined by maximizing F1-score (balanced precision-recall), F2-score (recall-focused), and Youden's J index (sensitivity-specificity balance). For medical screening, F2-score and recall optimization are typically preferred to minimize missed diagnoses.

XGBoost threshold optimization:

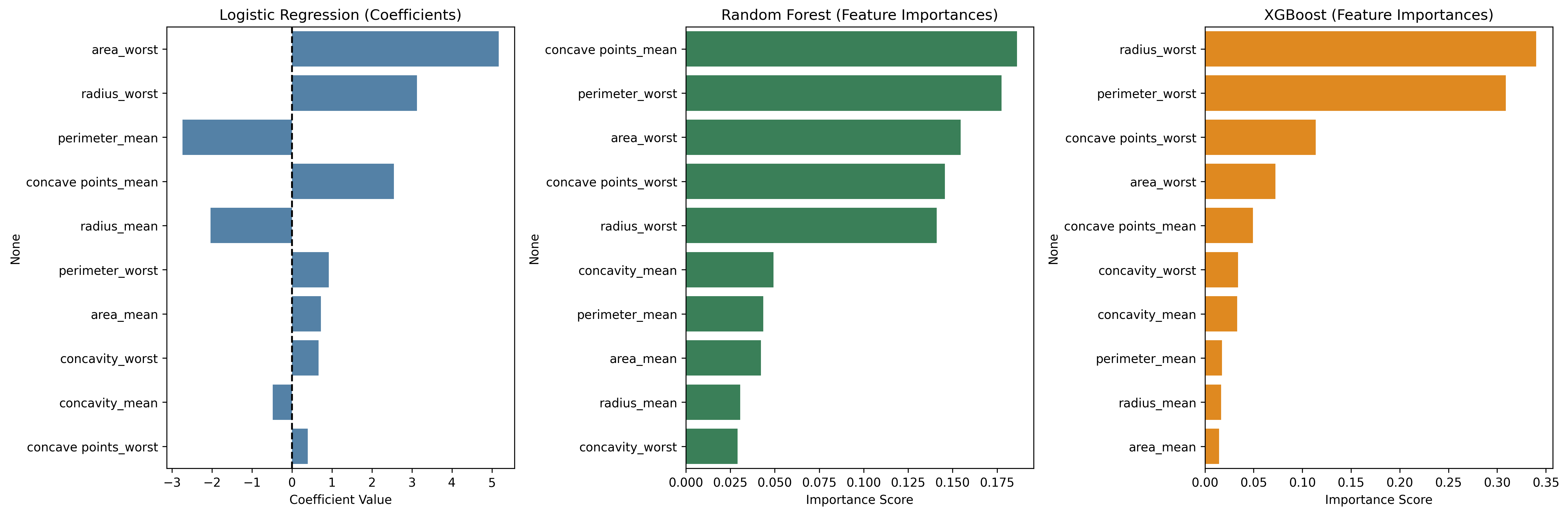


This plot shows how different classification thresholds affect XGBoost performance metrics. The optimal thresholds are determined by maximizing F1-score (balanced precision-recall), F2-score (recall-focused), and Youden's J index (sensitivity-specificity balance). For medical screening, F2-score and recall optimization are typically preferred to minimize missed diagnoses.

# 5. Traditional Feature Importance Analysis

## 5.1 Feature Importance Comparison

Comparison of traditional feature importances across all three models:

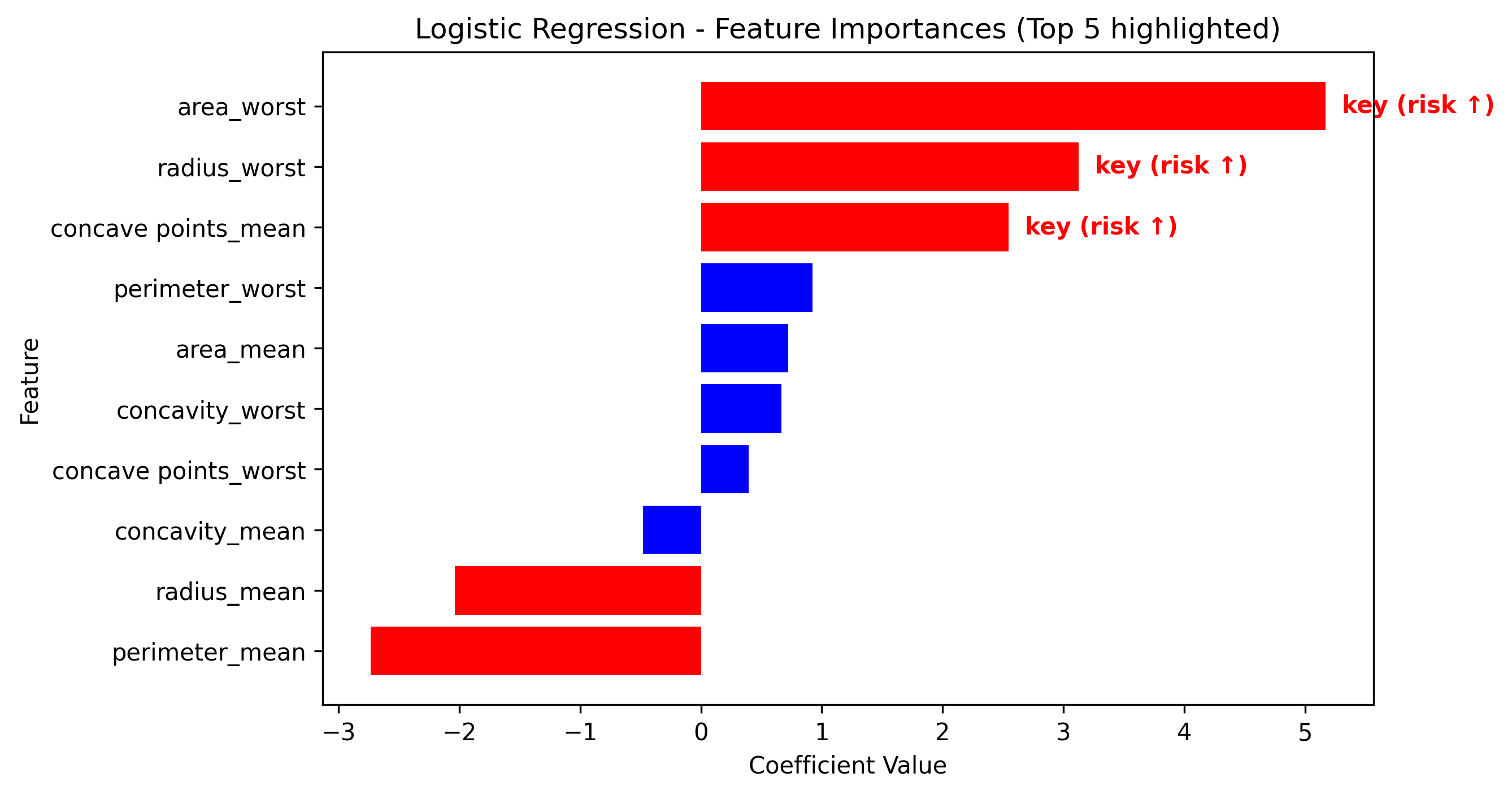


Top 5 most important features for each model (traditional methods):

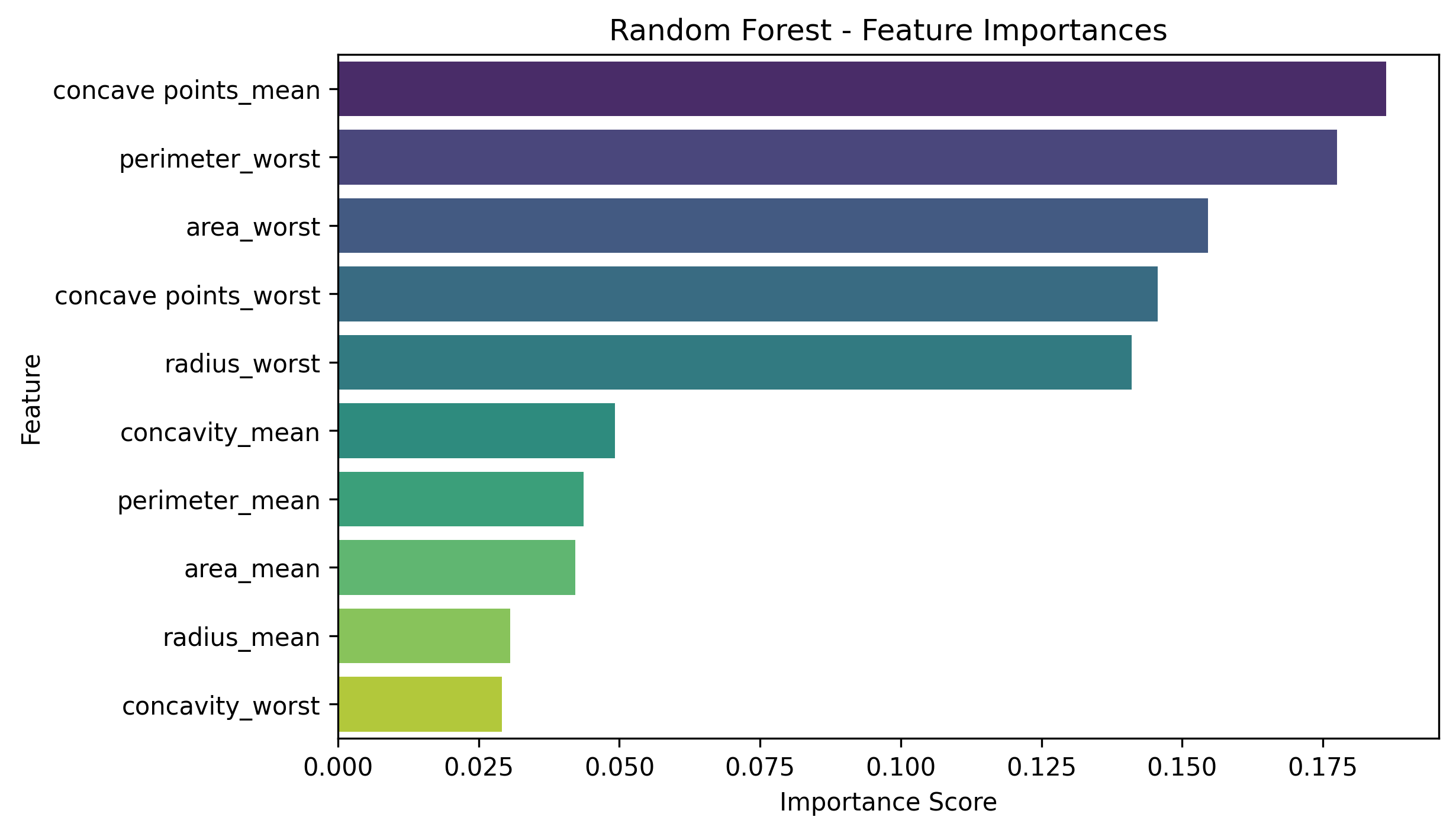
|  |  |  |
| --- | --- | --- |
| **Logistic Regression** | **Random Forest** | **XGBoost** |
| area\_worst | concave points\_mean | radius\_worst |
| radius\_worst | perimeter\_worst | perimeter\_worst |
| perimeter\_mean | area\_worst | concave points\_worst |
| concave points\_mean | concave points\_worst | area\_worst |
| radius\_mean | radius\_worst | concave points\_mean |

## 5.2 Individual Model Feature Importance

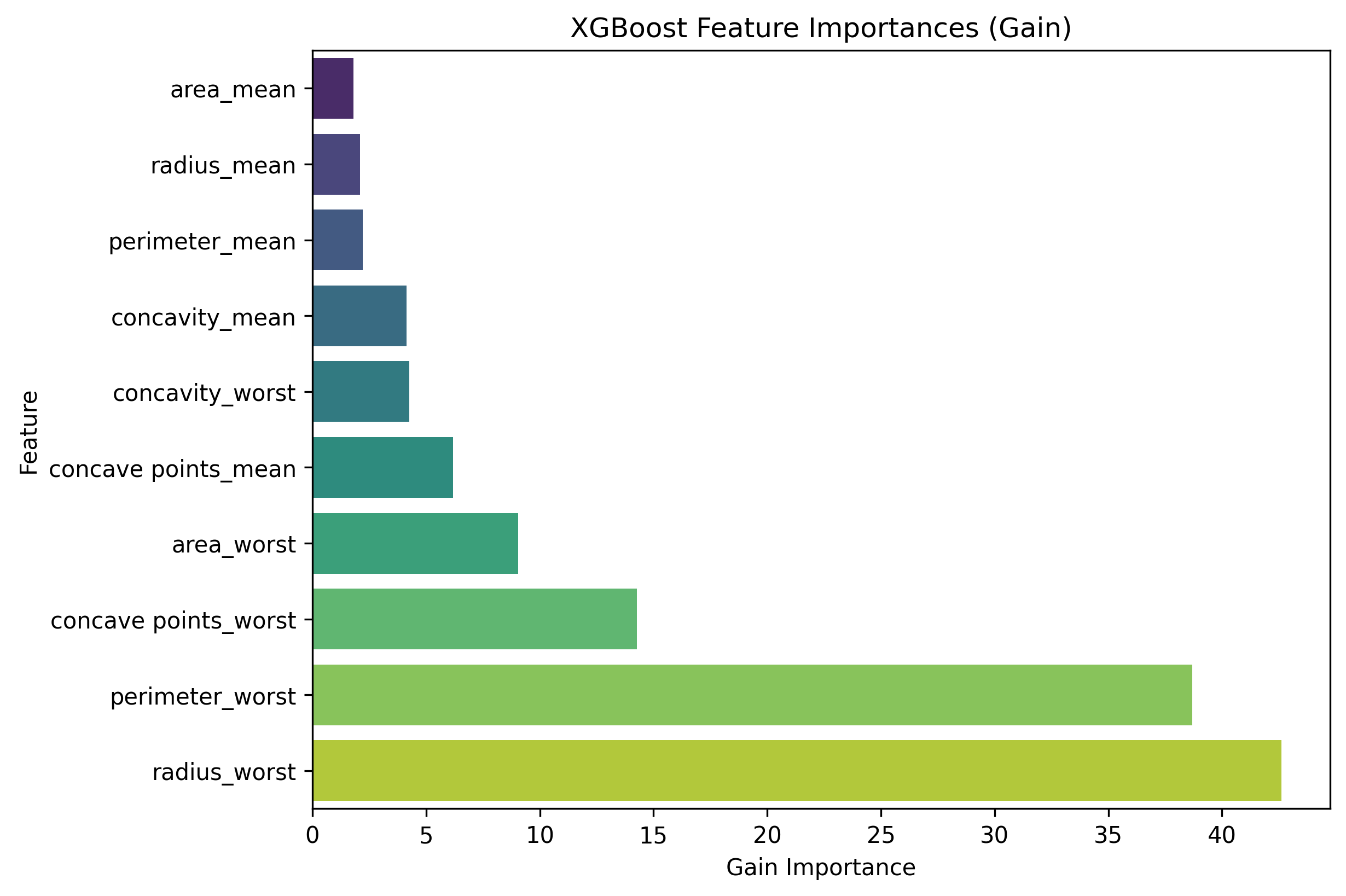
Logistic Regression traditional feature importance analysis:



Random Forest traditional feature importance analysis:



XGBoost traditional feature importance analysis:



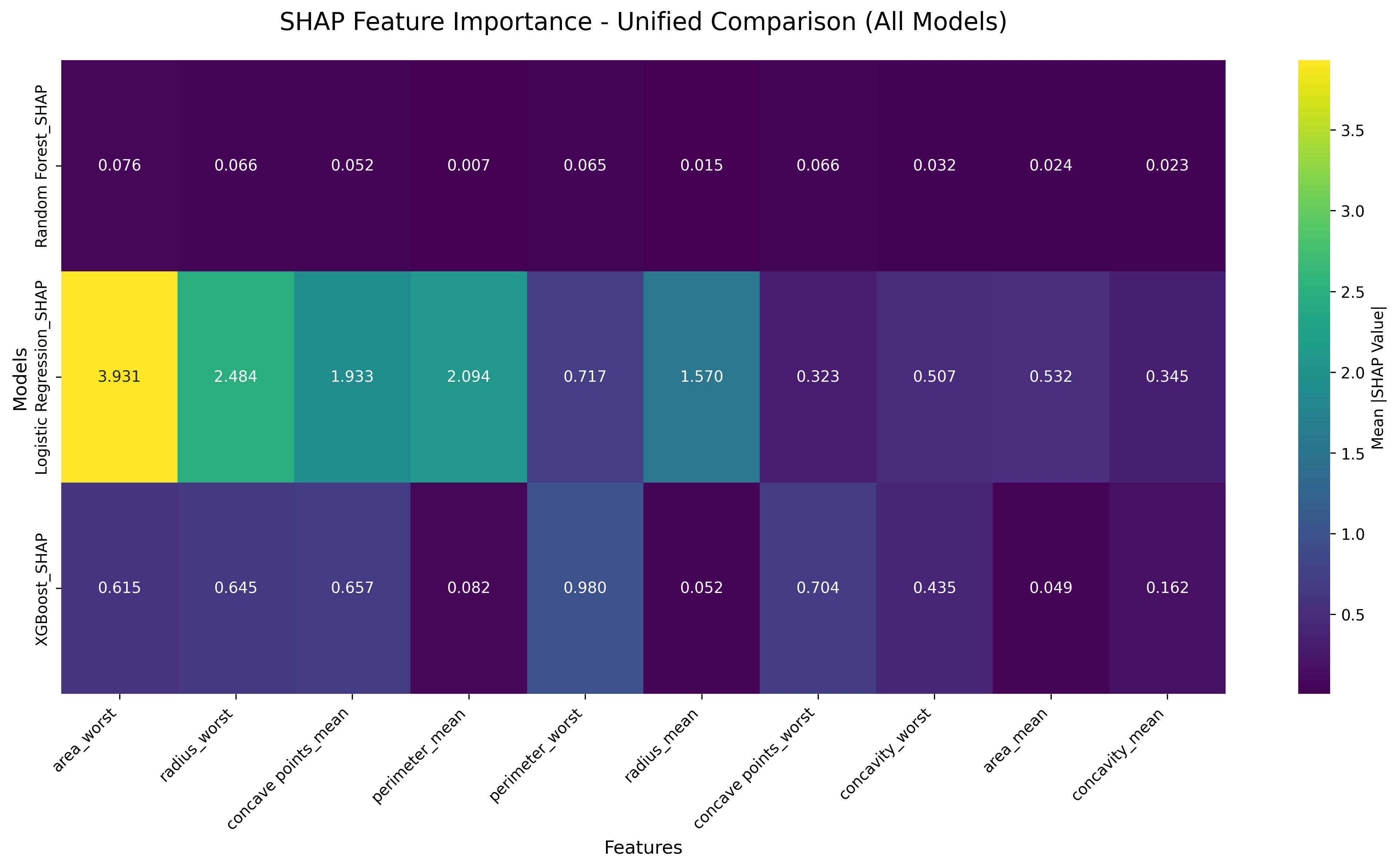
# 6. SHAP Explainability Analysis

SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain individual predictions by computing the contribution of each feature. Unlike traditional feature importance methods that provide global rankings, SHAP values explain how each feature contributes to moving the prediction away from the expected (baseline) value for each individual case. This is particularly valuable in clinical applications where understanding why a model made a specific prediction for a particular patient is crucial for trust and decision-making.

## 6.1 SHAP Global Feature Importance

Global SHAP importance summarizes feature contributions across all predictions by taking the mean absolute SHAP value for each feature. This provides a model-agnostic way to rank feature importance that accounts for both positive and negative contributions.

SHAP feature importance comparison across all models:



How to interpret this heatmap:  
• Darker colors indicate higher SHAP importance values  
• Each row represents a different model's SHAP analysis  
• Each column represents a feature from the dataset  
• Values show the mean absolute SHAP value across all test predictions  
• Features are ordered by overall importance across models  
• This allows comparison of which features each model considers most important

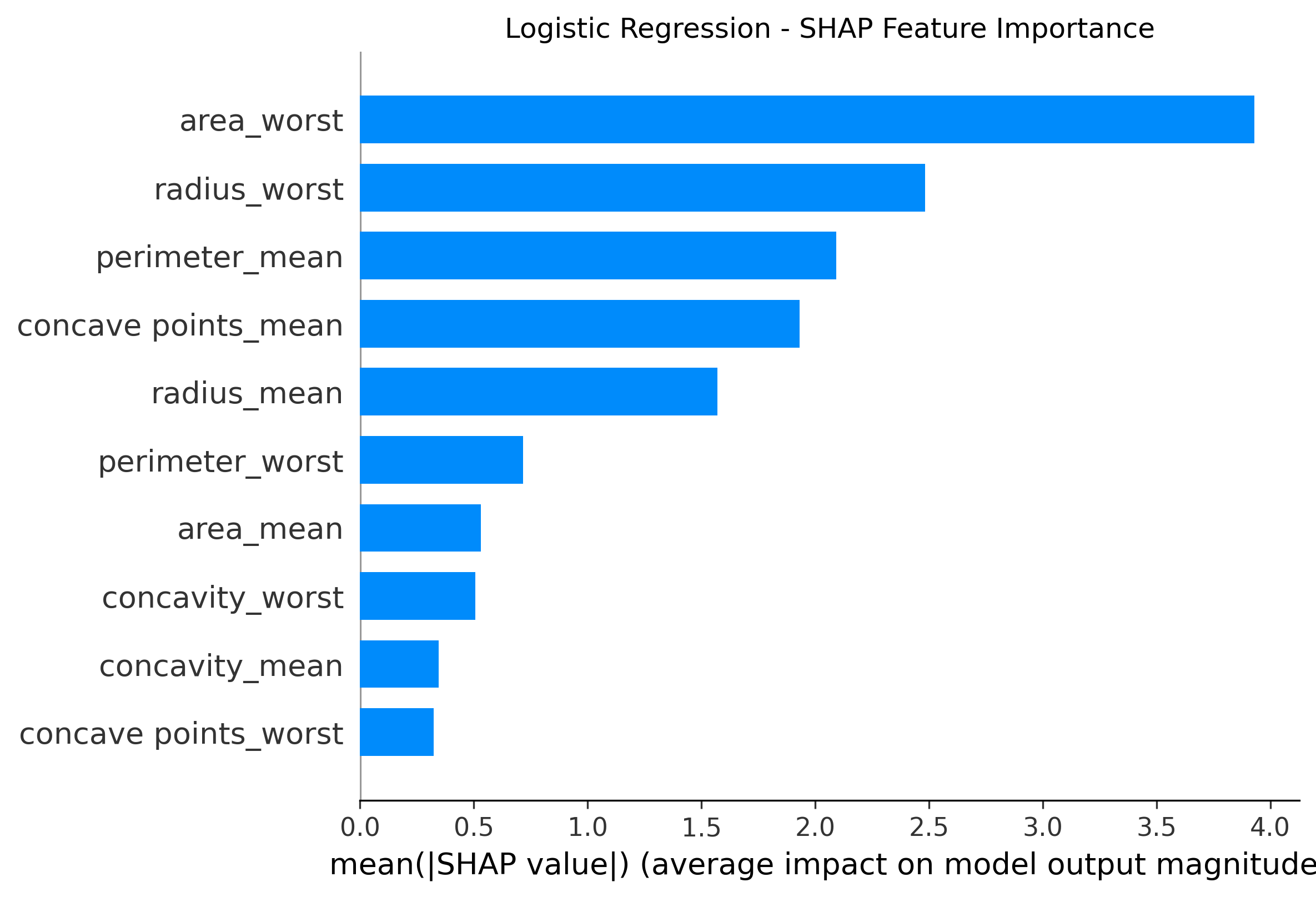
SHAP importance values for top features across models:

|  |  |  |
| --- | --- | --- |
| **feature** | **Logistic Regression\_SHAP** | **XGBoost\_SHAP** |
| area\_worst | 3.93 | 0.6151 |
| radius\_worst | 2.48 | 0.6451 |
| perimeter\_mean | 2.09 | 0.0820 |
| concave points\_mean | 1.93 | 0.6566 |
| radius\_mean | 1.57 | 0.0515 |
| perimeter\_worst | 0.7173 | 0.9803 |
| area\_mean | 0.5320 | 0.0491 |
| concavity\_worst | 0.5066 | 0.4349 |
| concavity\_mean | 0.3451 | 0.1615 |
| concave points\_worst | 0.3233 | 0.7040 |

## 6.2 Individual Model SHAP Analysis

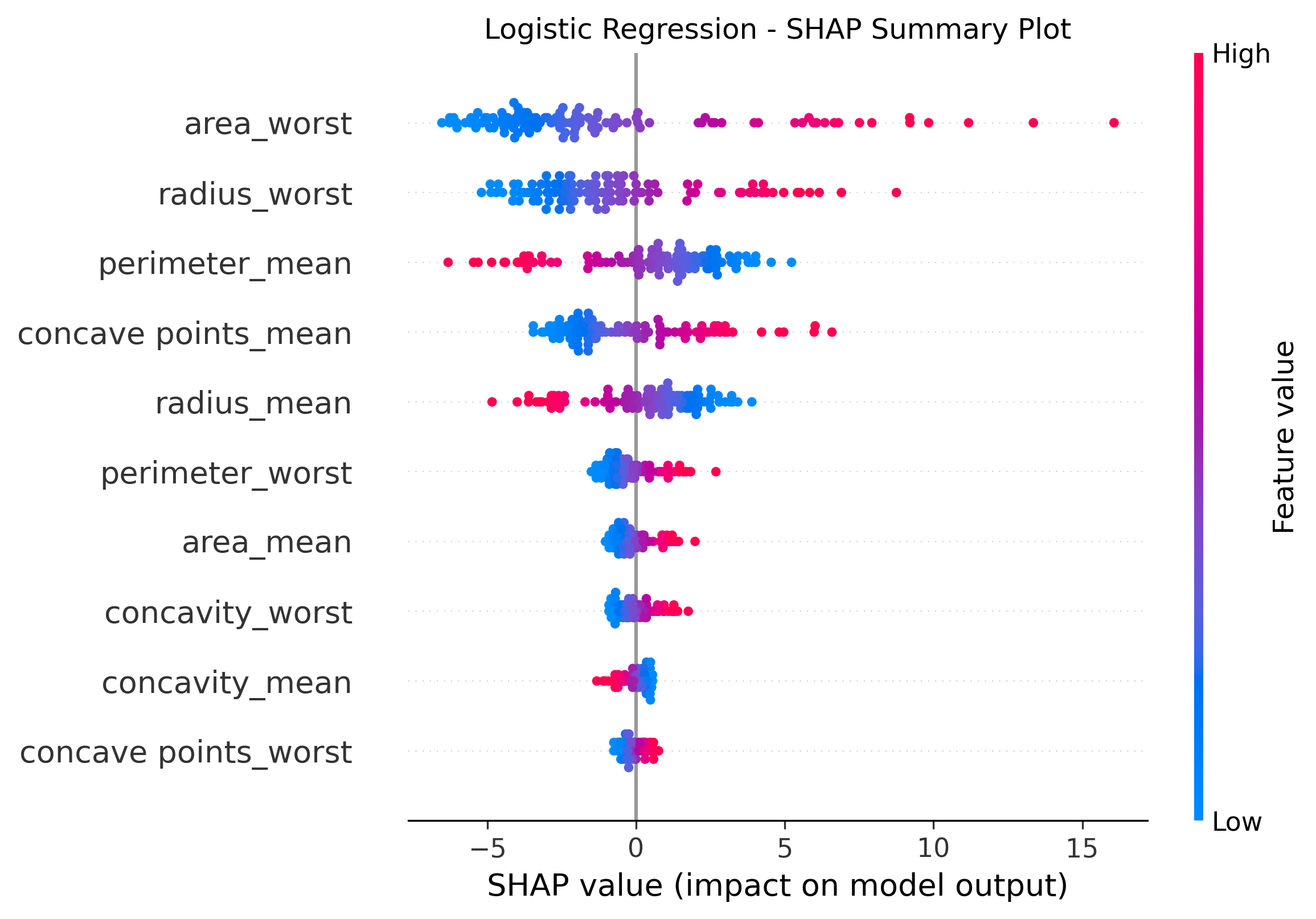
### 6.2.1 Logistic Regression SHAP Analysis

Logistic Regression SHAP feature importance (global ranking):



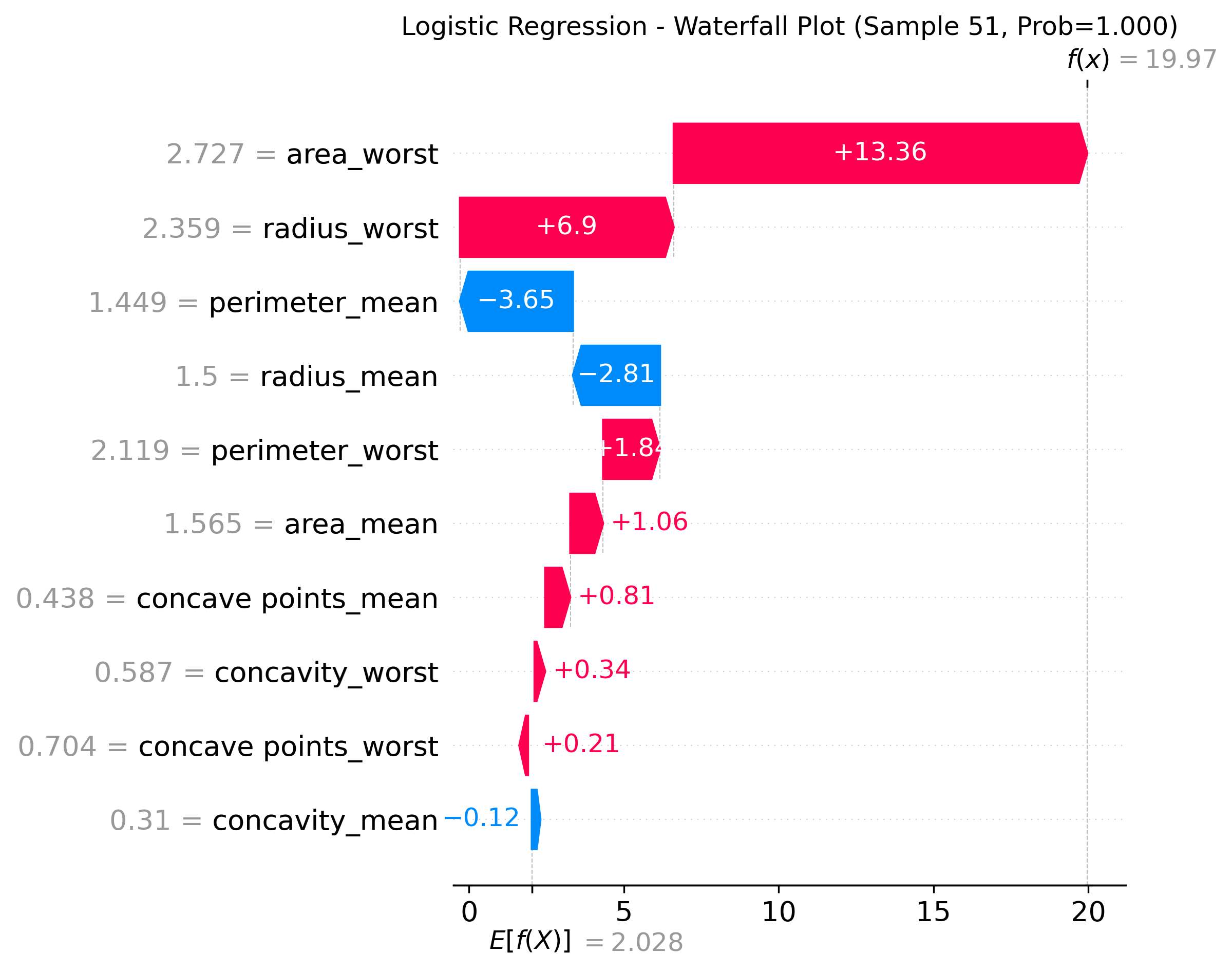
Logistic Regression - SHAP Feature Importance Bar Plot Interpretation:  
• Each bar represents the mean absolute SHAP value for that feature  
• Longer bars indicate features that have larger average impact on predictions  
• This shows which features the Logistic Regression model relies on most heavily  
• Unlike traditional feature importance, SHAP values are additive and sum to the difference between prediction and baseline  
• Features are ranked from most to least impactful on model decisions

Logistic Regression SHAP summary plot showing feature impact distribution:



Logistic Regression - SHAP Summary Plot (Beeswarm) Interpretation:  
• Each dot represents one patient's SHAP value for that feature  
• X-axis shows SHAP value (positive = increases malignancy probability, negative = decreases)  
• Y-axis lists features ordered by importance (top = most important)  
• Color indicates feature value (red = high feature value, blue = low feature value)  
• Width of dot cloud shows how much that feature's impact varies across patients  
• This reveals not just importance, but HOW each feature influences predictions

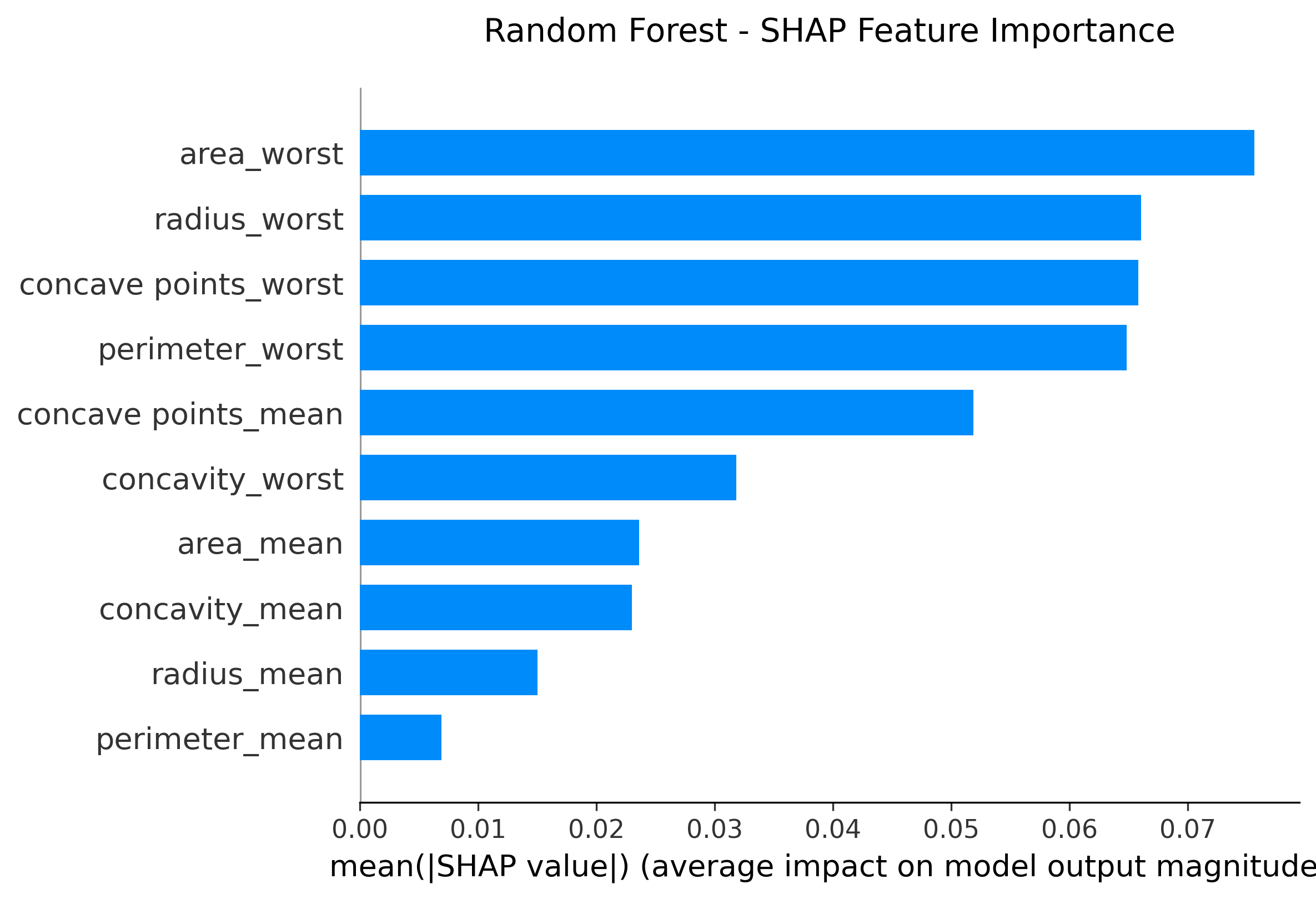
Logistic Regression SHAP waterfall plots for individual predictions:



Logistic Regression - SHAP Waterfall Plot Interpretation:  
• Shows exactly how the model arrived at this specific patient's prediction  
• Starts from the baseline (expected value across all patients)  
• Each bar shows how much each feature pushes the prediction up (red) or down (blue)  
• Features are ordered by magnitude of contribution to this specific prediction  
• Final prediction is the sum of baseline + all feature contributions  
• This provides complete transparency for individual clinical decisions

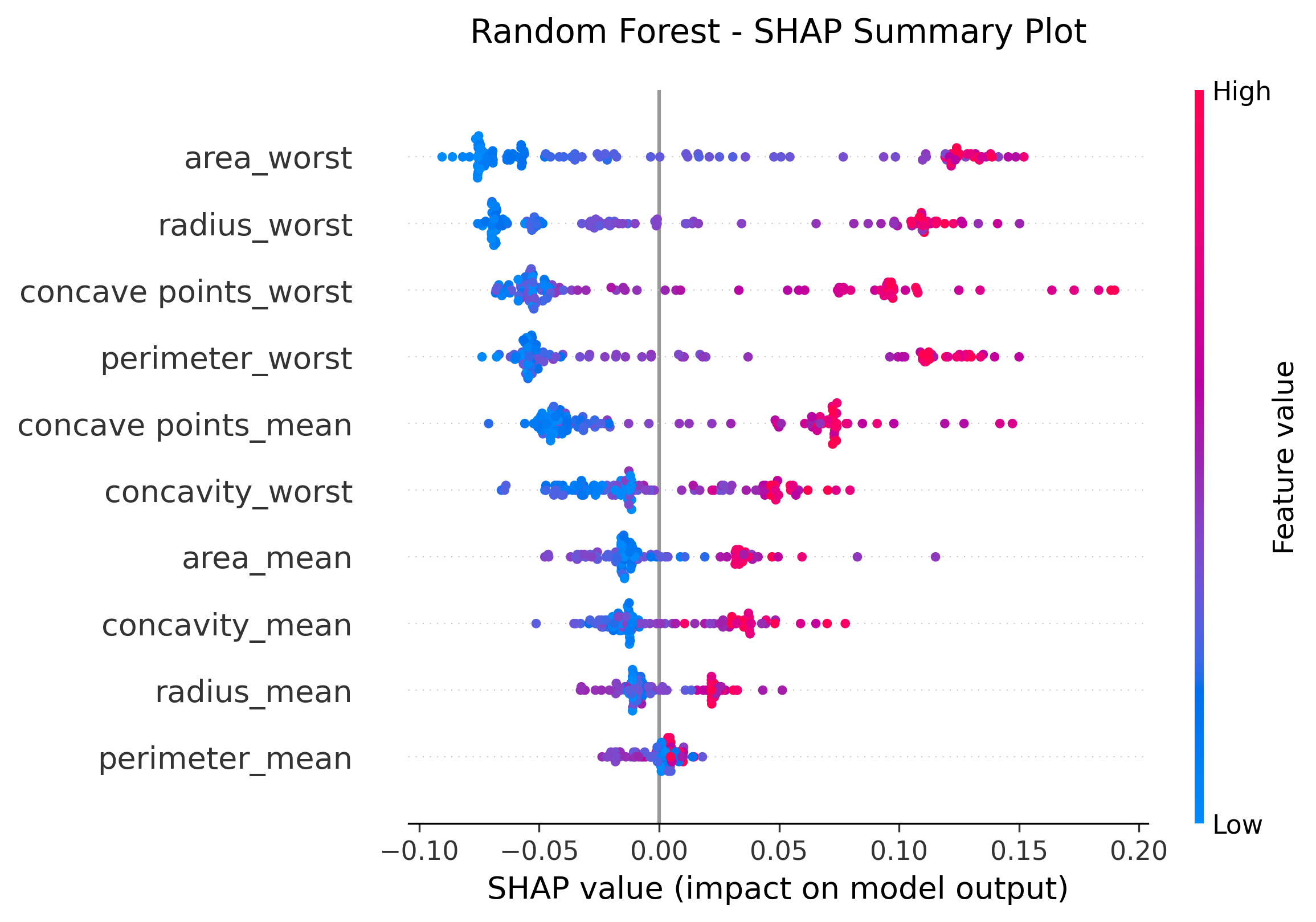
### 6.2.2 Random Forest SHAP Analysis

Random Forest SHAP feature importance (global ranking):



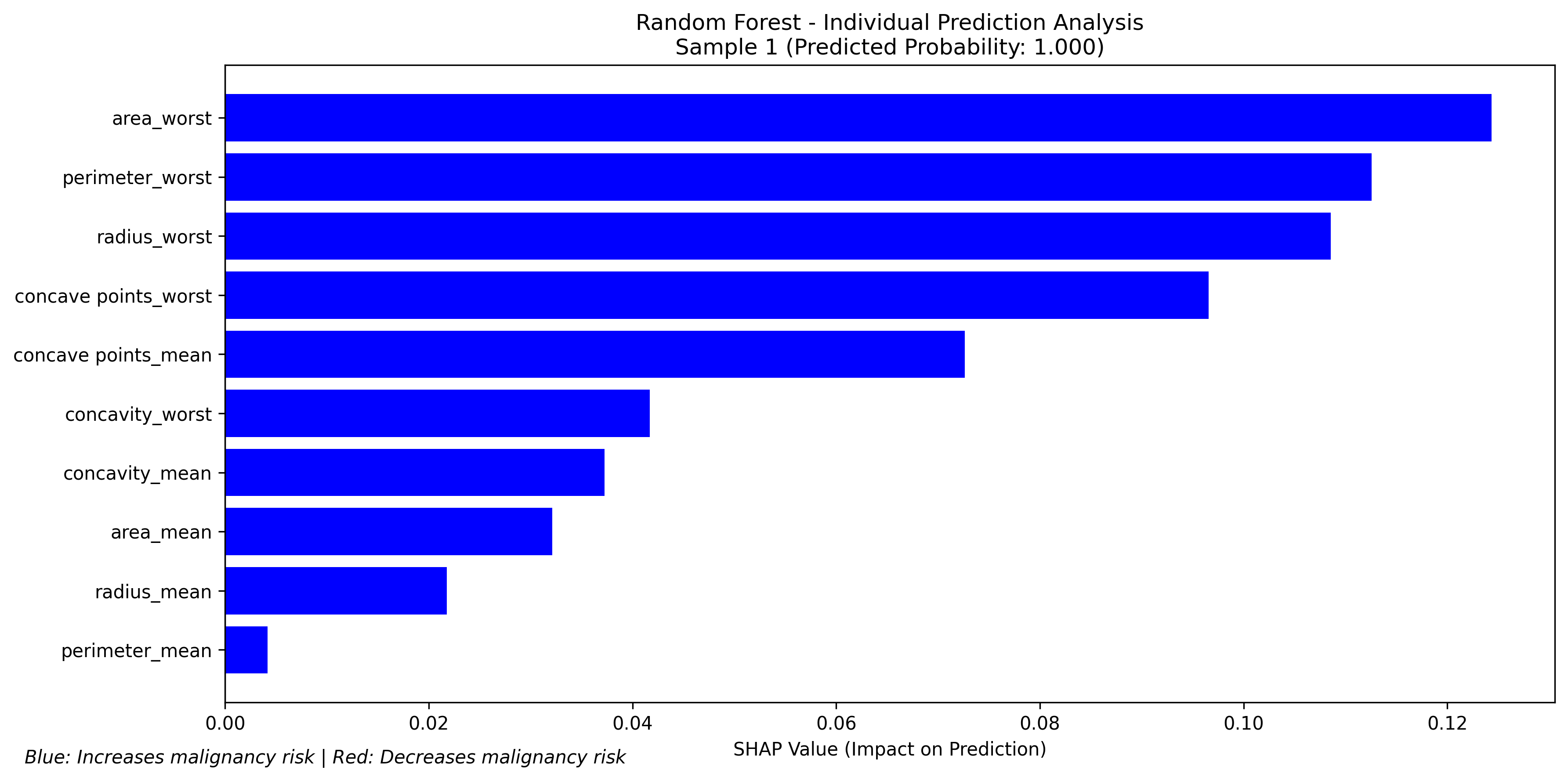
Random Forest - SHAP Feature Importance Bar Plot Interpretation:  
• Each bar represents the mean absolute SHAP value for that feature  
• Tree-based models like Random Forest can capture complex non-linear relationships  
• SHAP values account for feature interactions automatically captured by the trees  
• Features are ranked by their average contribution magnitude across all predictions  
• Unlike Gini importance, SHAP values are consistent and theoretically grounded

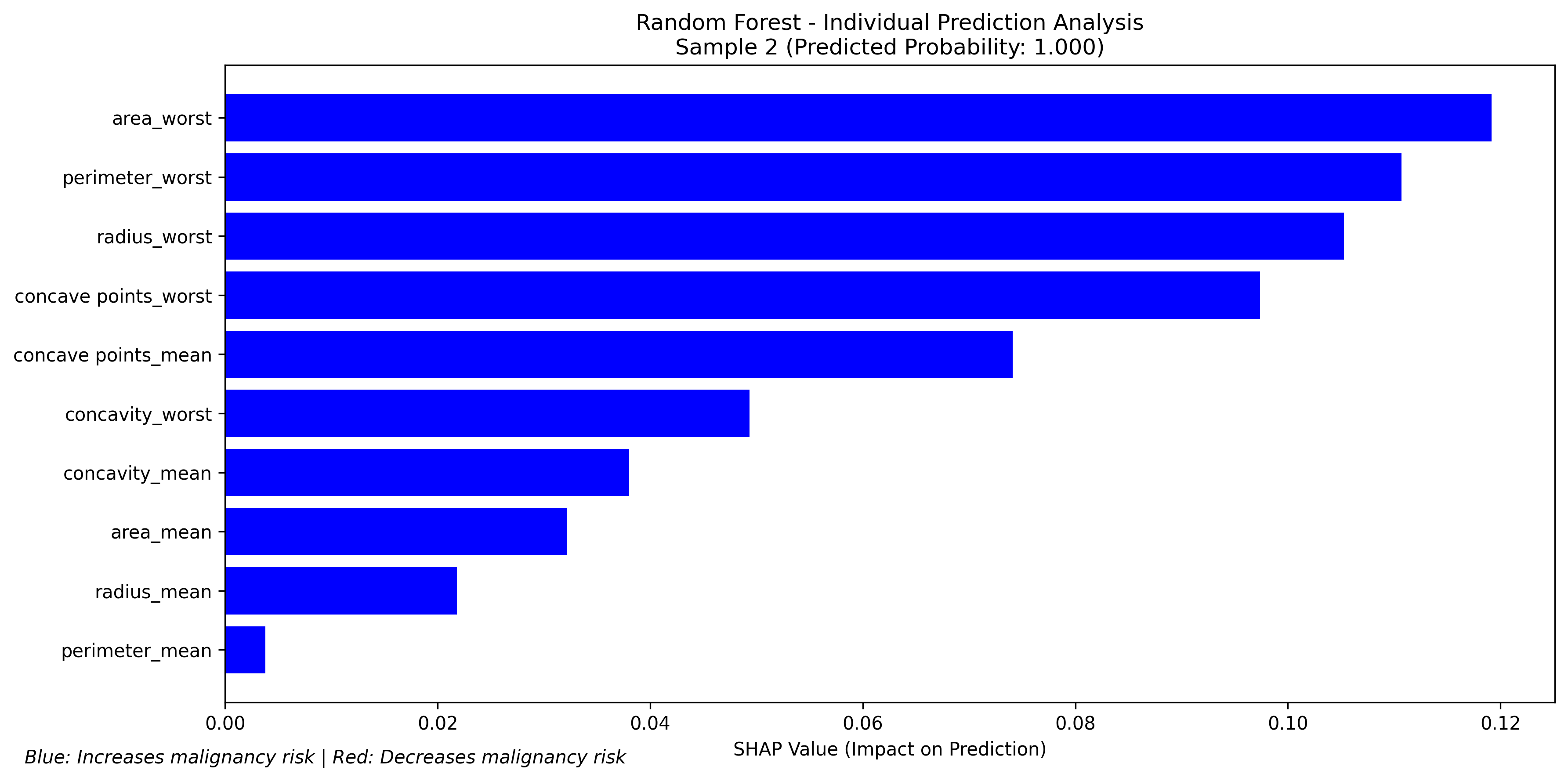
Random Forest SHAP summary plot showing feature impact distribution:



Random Forest - SHAP Summary Plot Interpretation:  
• Shows the relationship between feature values and their impact on predictions  
• For Random Forest, can reveal non-linear relationships and thresholds  
• Look for vertical spreads indicating feature interactions  
• Color patterns show if high/low feature values consistently increase or decrease risk

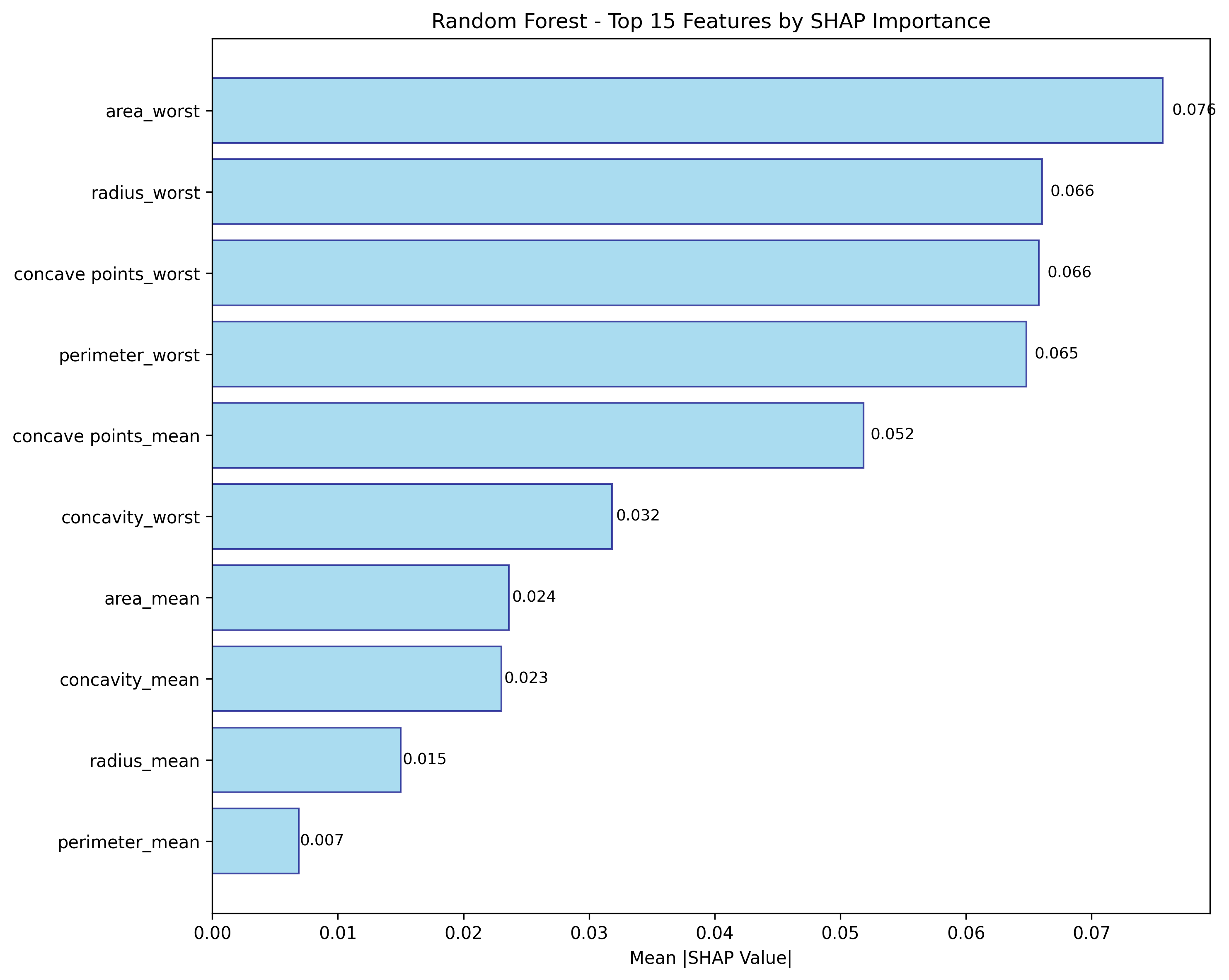
Random Forest SHAP individual prediction explanations:





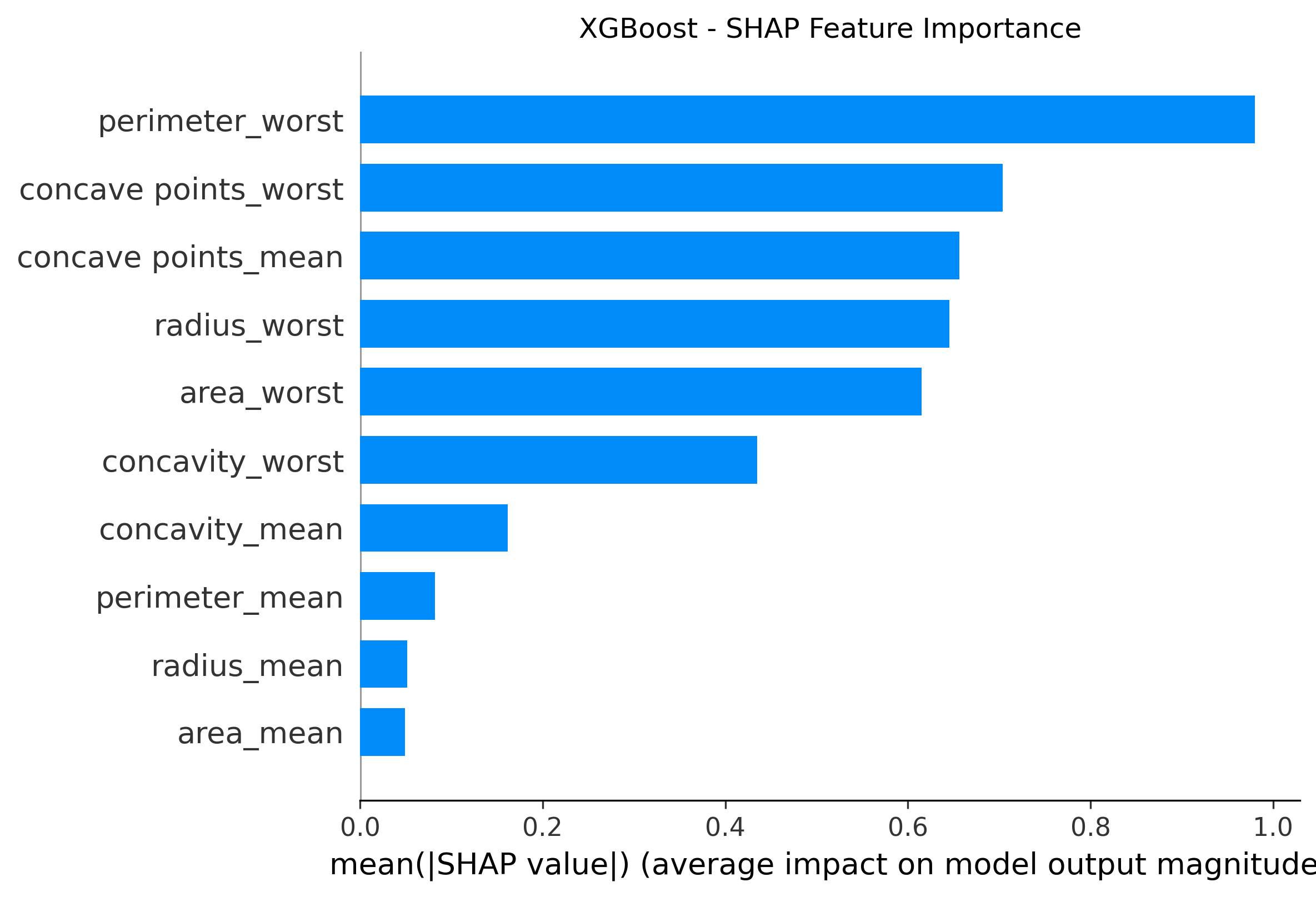
Random Forest - Individual Prediction Analysis:  
• Bar charts show top 10 features by magnitude for specific patients  
• Blue bars increase malignancy risk, red bars decrease risk  
• Random Forest can identify complex feature combinations  
• Each patient's explanation may highlight different feature combinations

Random Forest SHAP importance (alternative visualization):

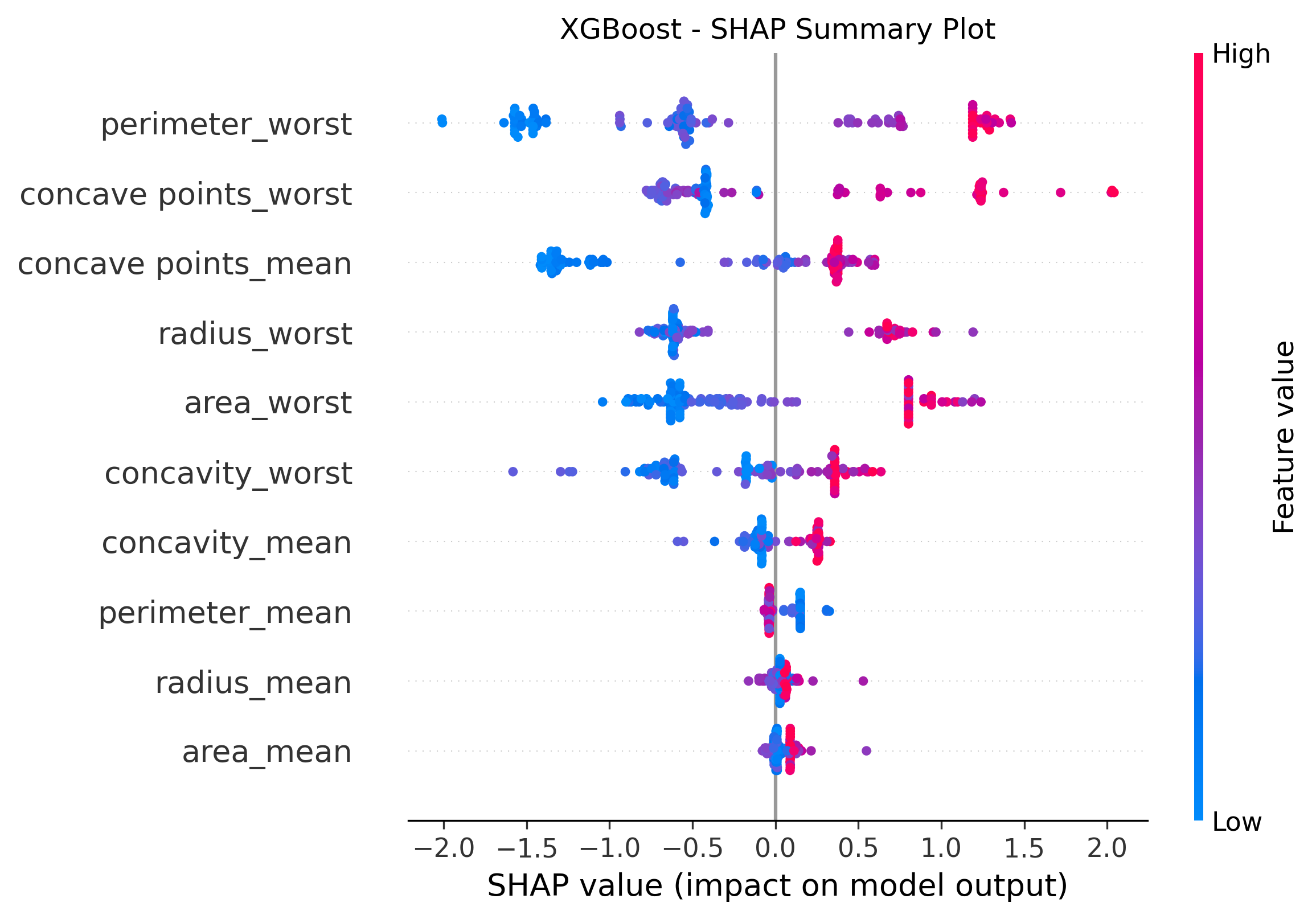


### 6.2.3 XGBoost SHAP Analysis

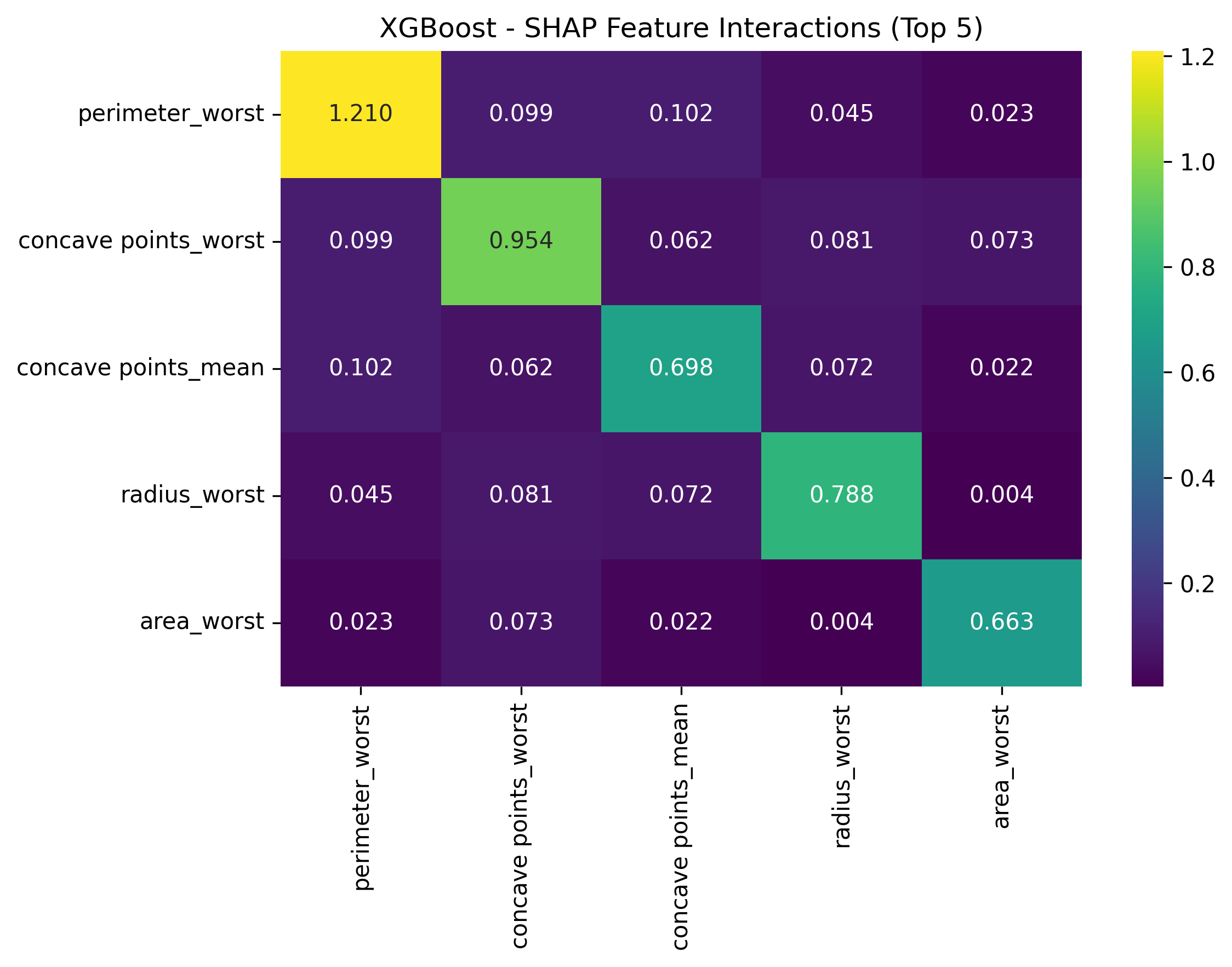
XGBoost SHAP feature importance (global ranking):



XGBoost - SHAP Feature Importance Interpretation:  
• XGBoost often identifies different important features than other models  
• Gradient boosting can detect subtle patterns missed by other algorithms  
• SHAP values account for the sequential nature of boosted trees  
• Feature importance reflects both individual effects and interaction contributions



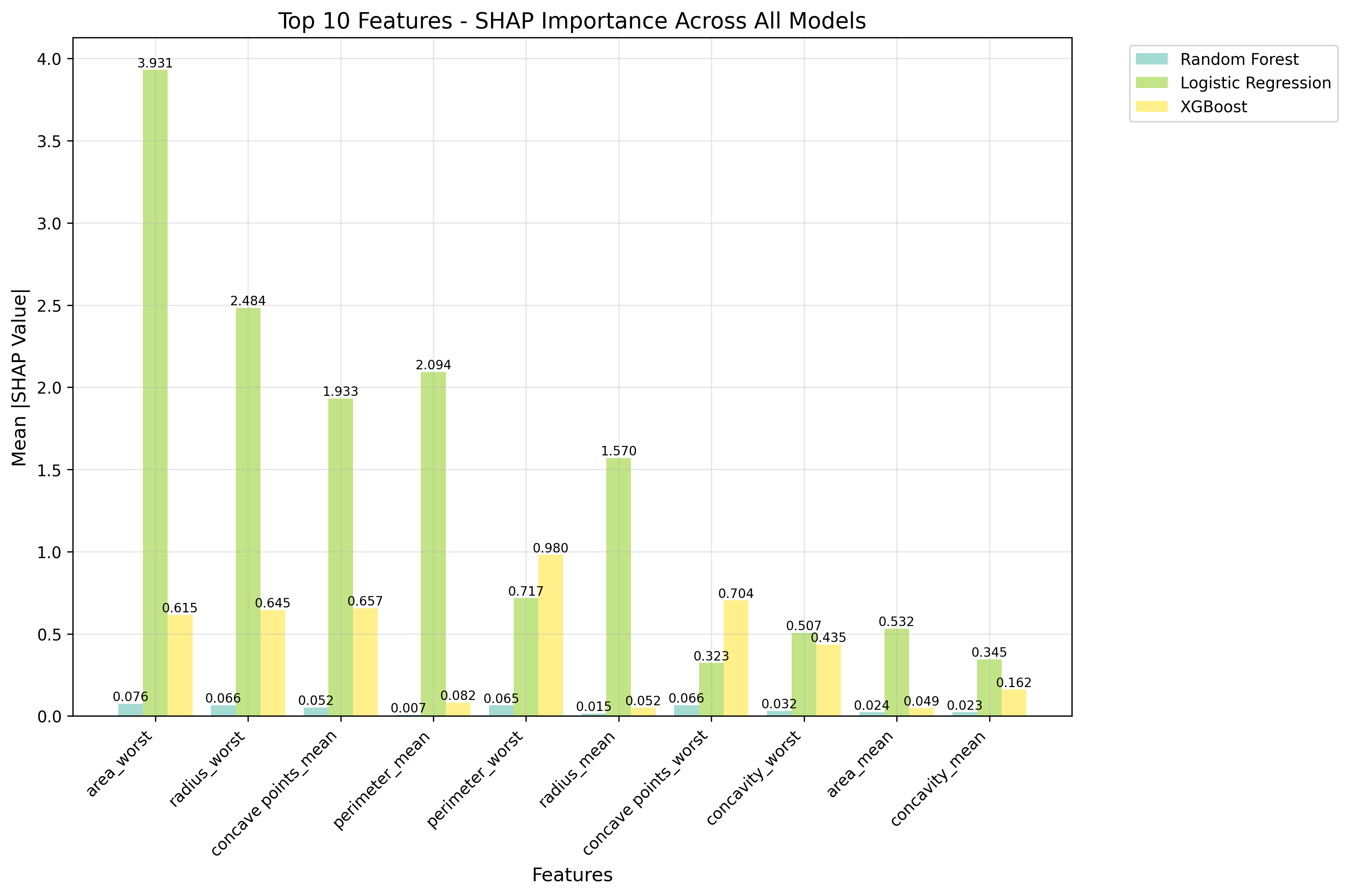
XGBoost SHAP feature interactions heatmap:



XGBoost - SHAP Feature Interactions Interpretation:  
• Shows how pairs of features interact to influence predictions  
• Off-diagonal cells show interaction strength between feature pairs  
• Darker colors indicate stronger interactions  
• High interaction values suggest combined effects different from individual effects  
• Identifies biomarker combinations that together provide unique diagnostic information

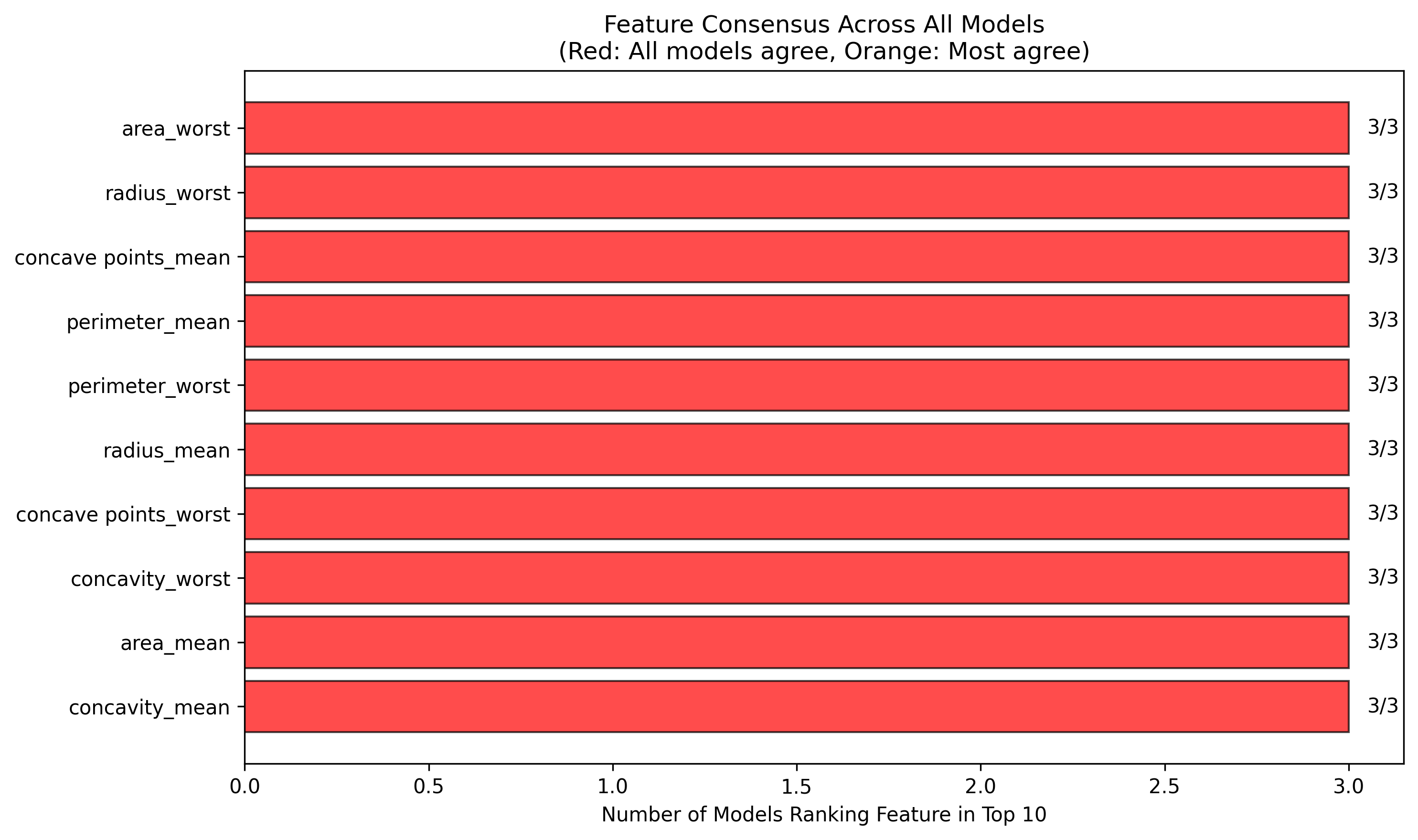
## 6.3 Unified SHAP Analysis and Clinical Insights

Top 10 features comparison across all models using SHAP values:



Top 10 Features Comparison Interpretation:  
• Each group of bars represents one feature  
• Different colored bars show different models' SHAP importance for that feature  
• Height of bars indicates mean absolute SHAP value  
• Features with consistent agreement across models are likely most reliable for clinical use  
• Large differences between models may indicate feature instability or model-specific biases

Feature consensus analysis across all models:



Feature Consensus Chart Interpretation:  
• Shows how many models rank each feature in their top 10 most important  
• Red bars: Features ALL models consider highly important (strongest clinical evidence)  
• Orange bars: Features MOST models consider important (good clinical candidates)  
• Light blue bars: Features only some models consider important (require careful evaluation)  
• This analysis helps prioritize biomarkers for clinical implementation

## 6.4 Clinical Implementation Recommendations

Based on the comprehensive SHAP analysis, the following clinical implementation strategy is recommended:  
  
1. Prioritize biomarkers with high consensus scores across all models for clinical deployment  
  
2. Use SHAP waterfall plots to explain individual patient predictions to clinical staff, enhancing trust and interpretability  
  
3. Monitor for cases where SHAP explanations conflict with clinical intuition, which may indicate model limitations or novel biomarker relationships  
  
4. Consider feature interactions when interpreting complex cases, particularly for tree-based models  
  
5. Regularly validate that SHAP explanations remain clinically meaningful as models are updated or retrained  
  
6. Focus on features that show consistent directional effects across models for reliable clinical guidance

# 7. Feature Selection and Consistency

Features consistently selected by all three models (threshold > 0.05 importance):

|  |
| --- |
| **Common Features** |
| area\_worst |
| perimeter\_worst |
| concave points\_worst |
| radius\_worst |

Logistic Regression selected features:

|  |
| --- |
| **Logistic Regression Selected** |
| concavity\_mean |
| radius\_worst |
| concavity\_worst |
| area\_mean |
| area\_worst |
| perimeter\_mean |
| radius\_mean |
| perimeter\_worst |
| concave points\_mean |
| concave points\_worst |

Random Forest selected features:

|  |
| --- |
| **Random Forest Selected** |
| radius\_worst |
| area\_worst |
| perimeter\_worst |
| concave points\_mean |
| concave points\_worst |

XGBoost selected features:

|  |
| --- |
| **XGBoost Selected** |
| area\_worst |
| radius\_worst |
| perimeter\_worst |
| concave points\_worst |

# 8. Clinical Decision Support

For medical screening applications, recall (sensitivity) is typically prioritized to minimize false negatives. The SHAP analysis provides additional clinical insights by explaining individual predictions, which enhances trust and interpretability in clinical settings. The following analysis provides clinical recommendations:

Clinical decision support analysis:

(Showing first 6 columns of 10 total)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Recall** | **Precision** | **Specificity** | **F1\_Score** | **Clinical\_Best\_Recall** |
| Logistic Regression | 0.9524 | 0.9524 | 0.9722 | 0.9524 | nan |
| Random Forest | 0.9286 | 1.00 | 1.00 | 0.9630 | nan |
| XGBoost | 0.9762 | 1.00 | 1.00 | 0.9880 | ✅ Best Sensitivity |

Model results with clinical preferences highlighted:

Table summary (original size: 3 rows × 23 columns):

###### Row 1:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | Logistic Regression |
| Accuracy | 0.9649 |
| Precision | 0.9524 |
| Recall | 0.9524 |
| Specificity | 0.9722 |
| F1\_Score | 0.9524 |
| ROC\_AUC | 0.9983 |
| CV\_ROC\_AUC | 0.9861 |
| Best\_Youden\_Threshold | 0.2560 |
| Best\_F1\_Threshold | 0.5570 |
| Best\_F2\_Threshold | 0.2560 |
| Youden\_Score | 0.9583 |
| Max\_F1\_Score | 0.9756 |
| Max\_F2\_Score | 0.9859 |
| Hyperparameter\_Search\_Candidates | 8 |
| Best\_Params | {'clf\_\_C': 10, 'clf\_\_penalty': 'l2', 'clf\_\_solver': 'saga'} |
| Mean\_Fit\_Time | 0.0385 |
| Mean\_Score\_Time | 0.0062 |
| Clinical\_Best\_Recall | nan |
| Clinical\_Best\_Precision | nan |
| Clinical\_Best\_ROC\_AUC | ✅ Best ROC-AUC |
| Clinical\_Best\_F1 | nan |
| Clinical\_Recommendation | nan |

###### Row 2:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | Random Forest |
| Accuracy | 0.9737 |
| Precision | 1.00 |
| Recall | 0.9286 |
| Specificity | 1.00 |
| F1\_Score | 0.9630 |
| ROC\_AUC | 0.9947 |
| CV\_ROC\_AUC | 0.9847 |
| Best\_Youden\_Threshold | 0.4720 |
| Best\_F1\_Threshold | 0.4720 |
| Best\_F2\_Threshold | 0.1330 |
| Youden\_Score | 0.9286 |
| Max\_F1\_Score | 0.9630 |
| Max\_F2\_Score | 0.9677 |
| Hyperparameter\_Search\_Candidates | 48 |
| Best\_Params | {'clf\_\_max\_depth': None, 'clf\_\_max\_features': 'sqrt', 'clf\_\_min\_samples\_leaf': 2, 'clf\_\_min\_samples\_split': 5, 'clf\_\_n\_estimators': 200} |
| Mean\_Fit\_Time | 0.5760 |
| Mean\_Score\_Time | 0.0297 |
| Clinical\_Best\_Recall | nan |
| Clinical\_Best\_Precision | ✅ Best Precision |
| Clinical\_Best\_ROC\_AUC | nan |
| Clinical\_Best\_F1 | nan |
| Clinical\_Recommendation | nan |

###### Row 3:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Model | XGBoost |
| Accuracy | 0.9912 |
| Precision | 1.00 |
| Recall | 0.9762 |
| Specificity | 1.00 |
| F1\_Score | 0.9880 |
| ROC\_AUC | 0.9977 |
| CV\_ROC\_AUC | 0.9831 |
| Best\_Youden\_Threshold | 0.3600 |
| Best\_F1\_Threshold | 0.3600 |
| Best\_F2\_Threshold | 0.3600 |
| Youden\_Score | 0.9762 |
| Max\_F1\_Score | 0.9880 |
| Max\_F2\_Score | 0.9809 |
| Hyperparameter\_Search\_Candidates | 72 |
| Best\_Params | {'clf\_\_colsample\_bytree': 0.8, 'clf\_\_learning\_rate': 0.05, 'clf\_\_max\_depth': 3, 'clf\_\_n\_estimators': 100, 'clf\_\_subsample': 1.0} |
| Mean\_Fit\_Time | 0.1559 |
| Mean\_Score\_Time | 0.0096 |
| Clinical\_Best\_Recall | ✅ Best Sensitivity |
| Clinical\_Best\_Precision | nan |
| Clinical\_Best\_ROC\_AUC | nan |
| Clinical\_Best\_F1 | ✅ Best F1-Score |
| Clinical\_Recommendation | 🏥 RECOMMENDED (Best Recall) |

# 9. Detailed Model Results

## 9.1 Logistic Regression Detailed Results

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| **Unnamed: 0** | **Pred\_0** | **Pred\_1** |
| Actual\_0 | 69 | 3 |
| Actual\_1 | 0 | 42 |

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unnamed: 0** | **precision** | **recall** | **f1-score** | **support** |
| 0 | 1.00 | 0.9583 | 0.9787 | 72.00 |
| 1 | 0.9333 | 1.00 | 0.9655 | 42.00 |
| accuracy | 0.9737 | 0.9737 | 0.9737 | 0.9737 |
| macro avg | 0.9667 | 0.9792 | 0.9721 | 114.00 |
| weighted avg | 0.9754 | 0.9737 | 0.9739 | 114.00 |

Best Thresholds Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **metric** | **threshold** | **precision** | **recall** | **score** |
| YOUDEN | 0.2560 | 0.9333 | 1.00 | 0.9583 |
| F1 | 0.5570 | 1.00 | 0.9524 | 0.9756 |
| F2 | 0.2560 | 0.9333 | 1.00 | 0.9859 |

## 9.2 Random Forest Detailed Results

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| **Unnamed: 0** | **Pred\_0** | **Pred\_1** |
| Actual\_0 | 72 | 0 |
| Actual\_1 | 3 | 39 |

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unnamed: 0** | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.9600 | 1.00 | 0.9796 | 72.00 |
| 1 | 1.00 | 0.9286 | 0.9630 | 42.00 |
| accuracy | 0.9737 | 0.9737 | 0.9737 | 0.9737 |
| macro avg | 0.9800 | 0.9643 | 0.9713 | 114.00 |
| weighted avg | 0.9747 | 0.9737 | 0.9735 | 114.00 |

Best Thresholds Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **metric** | **threshold** | **precision** | **recall** | **score** |
| YOUDEN | 0.4720 | 1.00 | 0.9286 | 0.9286 |
| F1 | 0.4720 | 1.00 | 0.9286 | 0.9630 |
| F2 | 0.1330 | 0.8571 | 1.00 | 0.9677 |

## 9.3 XGBoost Detailed Results

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| **Unnamed: 0** | **Pred\_0** | **Pred\_1** |
| Actual\_0 | 345 | 12 |
| Actual\_1 | 16 | 196 |

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unnamed: 0** | **precision** | **recall** | **f1-score** | **support** |
| 0 | 0.9557 | 0.9664 | 0.9610 | 357.00 |
| 1 | 0.9423 | 0.9245 | 0.9333 | 212.00 |
| accuracy | 0.9508 | 0.9508 | 0.9508 | 0.9508 |
| macro avg | 0.9490 | 0.9455 | 0.9472 | 569.00 |
| weighted avg | 0.9507 | 0.9508 | 0.9507 | 569.00 |

Best Thresholds Analysis:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **metric** | **threshold** | **precision** | **recall** | **score** |
| YOUDEN | 0.6200 | 0.9434 | 0.9246 | 0.8910 |
| F1 | 0.6200 | 0.9434 | 0.9246 | 0.9332 |
| F2 | 0.1600 | 0.8408 | 0.9717 | 0.9414 |

# 10. Conclusions and Recommendations

This comprehensive analysis evaluated three machine learning models for tumor tissue classification using both traditional and advanced explainability methods. Key findings include:  
  
• All models achieved good performance with proper hyperparameter tuning  
• Threshold optimization revealed trade-offs between precision and recall  
• Traditional feature importance identified key predictive biomarkers  
• SHAP analysis provided deeper insights into individual predictions and feature interactions  
• Clinical recommendations prioritize models with highest recall for screening applications  
• SHAP explainability enhances clinical trust and interpretability  
  
The combination of rigorous model evaluation and advanced explainability provides a robust foundation for clinical decision-making with appropriate model selection based on specific use case requirements. The SHAP analysis particularly enhances the clinical applicability by providing interpretable explanations for individual patient predictions.