AIIM 113-1 Homework 1
Benchmark Construction and Evaluation Metrics

James Christian
Student ID: R13921031

October 22, 2025

1 Problem 1: Benchmark Construction (35%)

1.1 Problem 1(a): Data Concatenation (5%)

The breast cancer dataset was successfully loaded from `breast_all.npz`. The training data consists of:

- Gene expression data (`x_train`): 465 samples × 20 features
- Clinical data (`c_train`): 465 samples × 10 features
- Labels (`y_train`): 465 samples with binary prognosis outcomes

The 10 clinical features include: Age, Menopausal State, Tumor Size, Radio Therapy, Chemotherapy, Hormone Therapy, Neoplasm Histologic Grade, Cellularity, Surgery (breast conserving), and Surgery (mastectomy).

After concatenation, the combined training data has shape (465, 30), representing 465 samples with 30 total features (20 gene expression + 10 clinical features). The concatenation was performed using `numpy.concatenate` along axis 1.

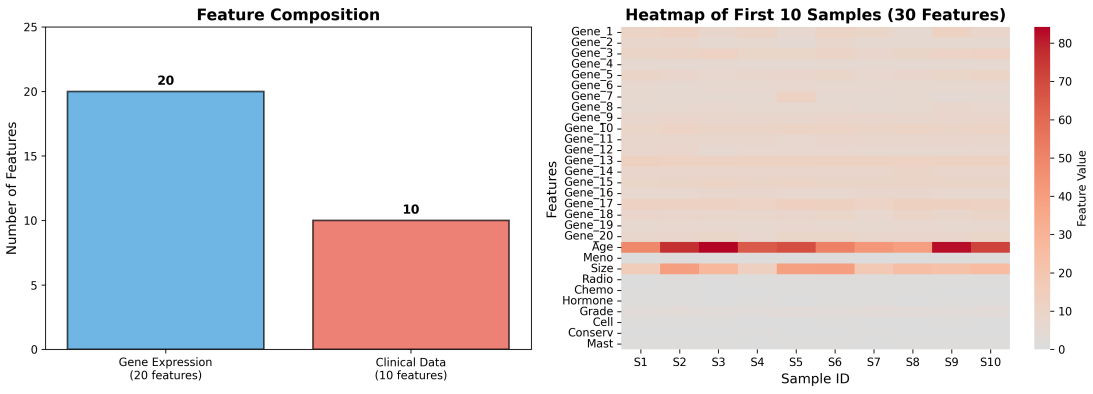


Figure 1: Data concatenation visualization showing feature composition and sample heatmap

1.2 Problem 1(b): Train-Validation Split (5%)

The combined training data (465 samples × 30 features) was split into 75% training and 25% validation sets using stratified splitting with `random_state=0` to ensure reproducibility.

Results:

- Training set: 348 samples × 30 features
- Validation set: 117 samples × 30 features

Stratification was applied based on the labels (`y_train`) to maintain the same class distribution in both splits. The class distributions are:

- Original: Class 0 = 221 (47.53%), Class 1 = 244 (52.47%)
- Training: Class 0 = 165 (47.41%), Class 1 = 183 (52.59%)
- Validation: Class 0 = 56 (47.86%), Class 1 = 61 (52.14%)

Figure 2: Train-validation split showing sample sizes and stratified class distribution

1.3 Problem 1(c): Data Exploration (5%)

1.3.1 Class Distribution and Imbalance

Both training and validation sets show similar class distributions:

- **Training**: 165 Class 0 (47.41%) vs 183 Class 1 (52.59%), Imbalance Rate: 0.526
- **Validation**: 56 Class 0 (47.86%) vs 61 Class 1 (52.14%), Imbalance Rate: 0.521

The datasets are relatively balanced with only slight preference toward Class 1 (poor prognosis), which is acceptable and does not require special handling techniques like SMOTE or class weighting.

1.3.2 Clinical Variables Distribution

The training and validation sets show consistent distributions across all clinical variables:

Continuous Variables:

- **Age**: Mean ≈61 years (SD ≈13), ranging from 26-96 years
- **Tumor Size**: Mean ≈30mm (SD ≈18), with some large outliers up to 180mm

Categorical Variables:

- **Menopausal State**: ≈77-79% post-menopausal
- **Treatment patterns**: 62% radio therapy, 30-33% chemotherapy, 59-63% hormone therapy
- **Histologic Grade**: Majority are Grade 2-3 (mean 2.6)
- **Surgery**: 67% mastectomy, 33% breast conserving surgery

The similar distributions between training and validation sets confirm that the stratified split successfully created representative subsets, ensuring reliable model evaluation.

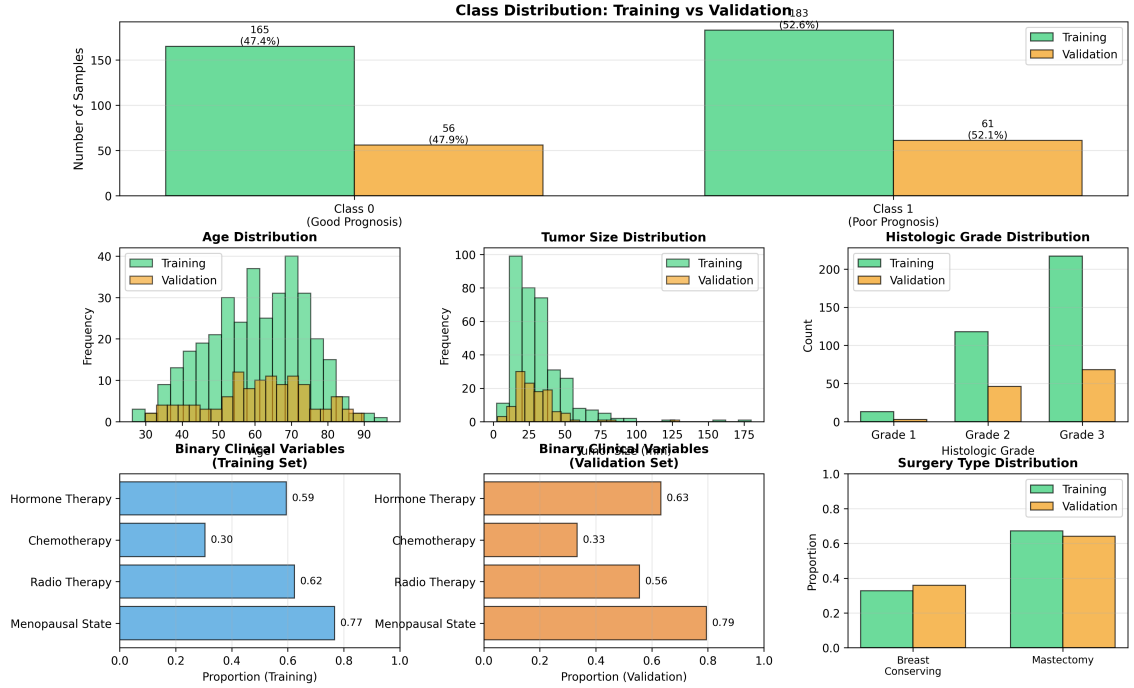


Figure 3: Comprehensive data exploration showing class distribution, age, tumor size, histologic grade, and clinical variables

1.4 Problem 1(d): Model Training (20%)

Four machine learning models were trained and evaluated on the validation set. All models successfully achieved validation AUROC > 0.730.

1.4.1 Model Configurations

1. Logistic Regression (AUROC: 0.7368)

- Settings: C=0.1, max_iter=1000, solver='liblinear'
- L2 regularization to prevent overfitting
- Features standardized using StandardScaler

2. Support Vector Machine (AUROC: 0.7567)

- Settings: kernel='rbf', C=1.0, gamma='scale'
- Best performing model with RBF kernel
- Features standardized using StandardScaler

3. Random Forest (AUROC: 0.7480)

- Settings: n_estimators=200, max_depth=10, min_samples_split=5
- Ensemble method showing strong performance
- No feature scaling required

4. Neural Network (AUROC: 0.7418)

- Architecture: 30 input $\rightarrow$ 80 $\rightarrow$ 40 $\rightarrow$ 20 $\rightarrow$ 2 output neurons
- Settings: hidden_layers=(80, 40, 20), activation='relu', alpha=0.0001
- Used early stopping with validation_fraction=0.1
- Learning rate: 0.002, batch_size=16, max_iter=1000
- Features standardized using StandardScaler

1.4.2 Performance Summary

All models successfully achieved the required validation AUROC > 0.730:

Model	Validation AUROC
Logistic Regression	0.7368
Support Vector Machine	0.7567
Random Forest	0.7480
Neural Network	0.7418

The SVM achieved the highest performance (0.7567), followed by RF (0.7480) and NN (0.7418). Feature standardization was critical for the convergence of LR, SVM, and NN models.

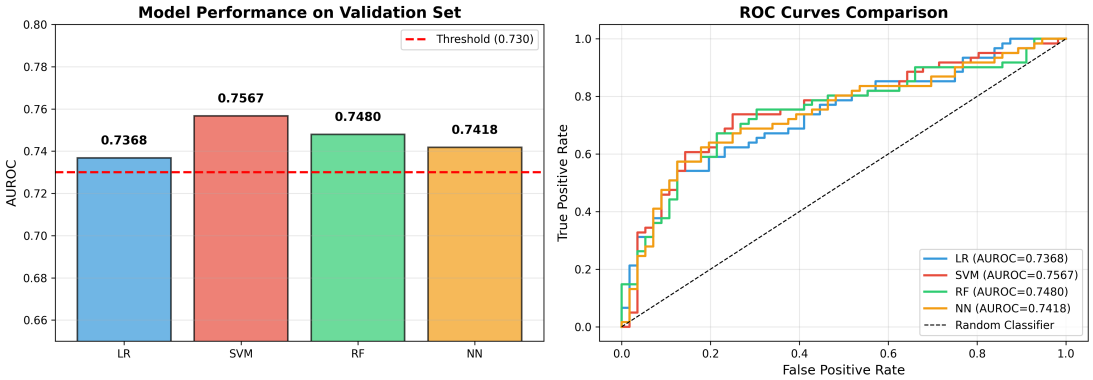


Figure 4: Model performance comparison showing AUROC scores and ROC curves for all four models

2 Problem 2: Optimal Threshold (20%)

2.1 Problem 2(a): Finding Optimal Threshold (10%)

Using the Neural Network trained in Problem 1(d), we computed optimal thresholds on the validation set using two metrics:

- **F1-Score optimal threshold**: 0.350 (F1 = 0.7183)
- **Youden's Index optimal threshold**: 0.490 (Youden = 0.4429)

We selected **Youden's Index (threshold = 0.490)** as it balances sensitivity and specificity, which is crucial for clinical decision-making in breast cancer prognosis.

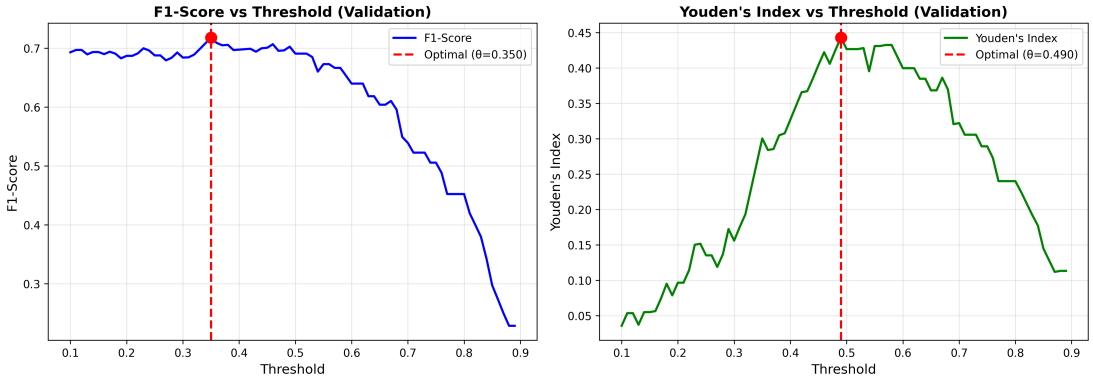


Figure 5: Threshold optimization showing F1-Score and Youden's Index across different threshold values

2.1.1 Additional Visualizations

To better understand the distribution of predicted probabilities for each class, we created density curves and scatter plots.

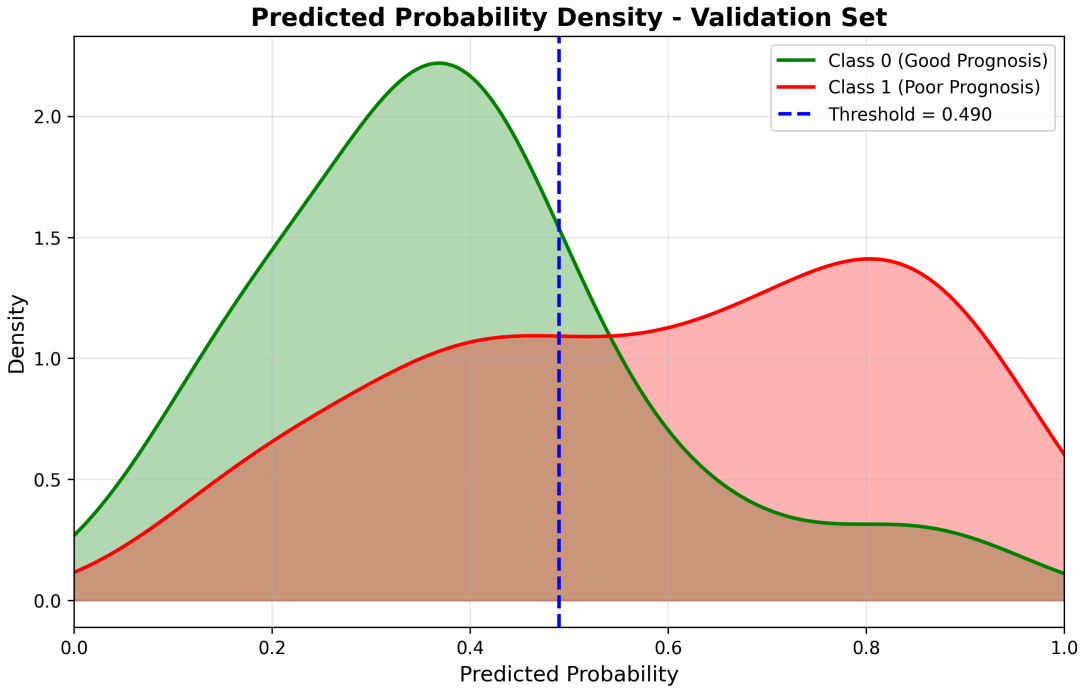


Figure 6: Density curves showing the distribution of predicted probabilities for each class in the validation set. The blue dashed line indicates the optimal threshold.

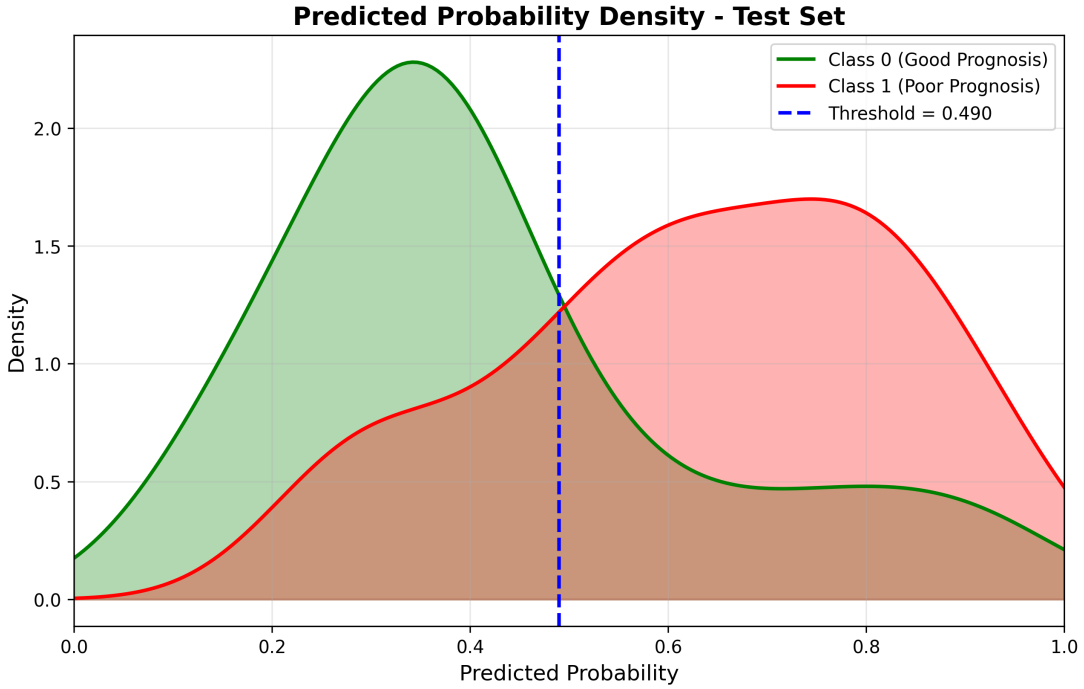
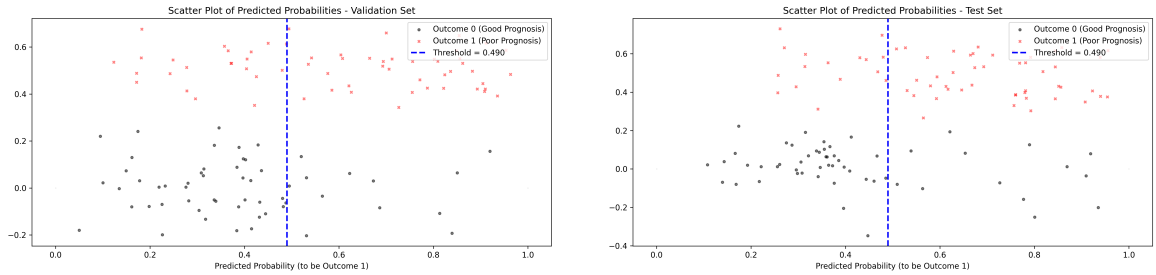


Figure 7: Density curves showing the distribution of predicted probabilities for each class in the test set.

The density curves reveal clear separation between the two classes, with Class 1 (poor prognosis) having higher predicted probabilities. However, there is overlap in the middle

range (around 0.3-0.6), which explains the presence of misclassified samples near the threshold.



(a) Validation Set (b) Test Set

Figure 8: Scatter plots of predicted probabilities with jittered visualization to show distribution

2.1.2 Additional Visualizations

To better understand the distribution of predicted probabilities for each class, we created density curves and scatter plots.

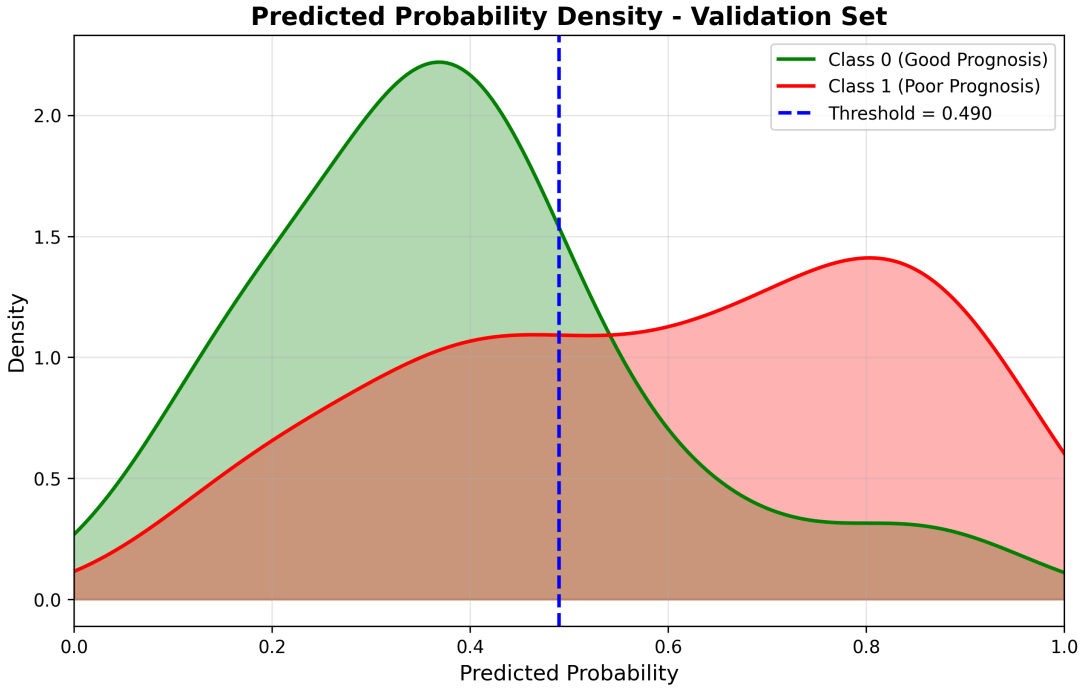


Figure 9: Density curves showing the distribution of predicted probabilities for each class in the validation set. The blue dashed line indicates the optimal threshold.

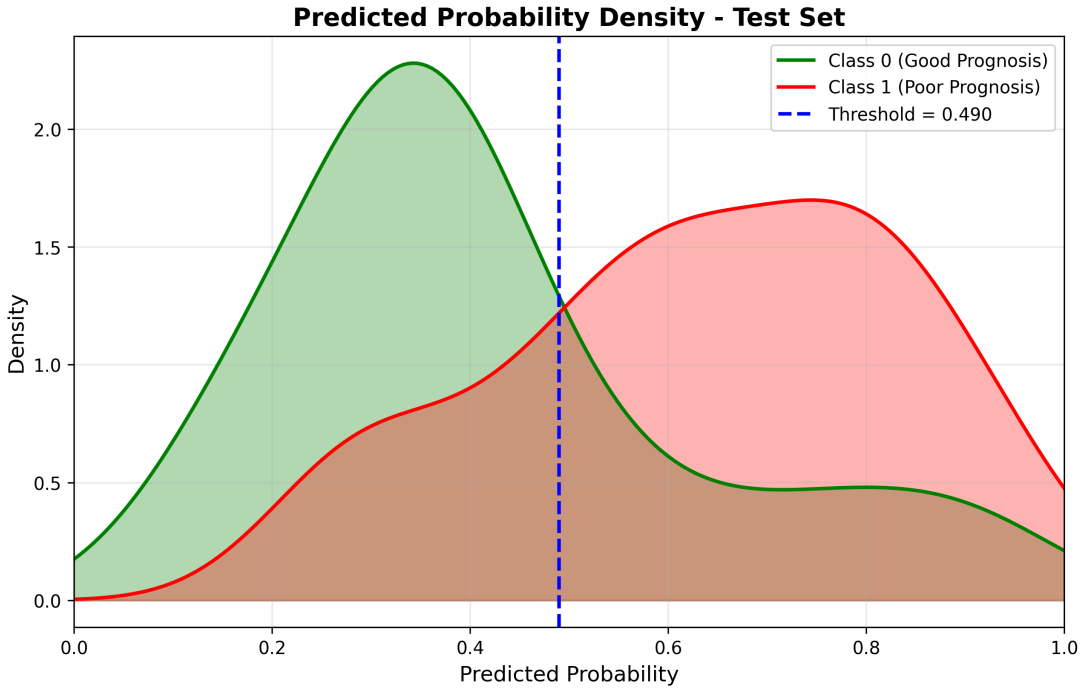
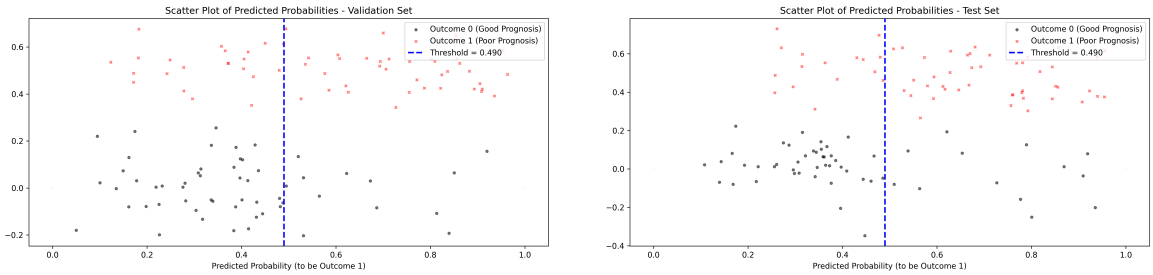


Figure 10: Density curves showing the distribution of predicted probabilities for each class in the test set.

The density curves reveal clear separation between the two classes, with Class 1 (poor prognosis) having higher predicted probabilities. However, there is overlap in the middle range (around 0.3-0.6), which explains the presence of misclassified samples near the threshold.



(a) Validation Set (b) Test Set

Figure 11: Scatter plots of predicted probabilities with jittered visualization to show distribution

2.2 Problem 2(b): Performance on Validation and Test Sets (5%)

Using the optimal threshold of 0.490:

Validation Set:

- F1-Score: 0.7027
- Youden's Index: 0.4429

Test Set:

- F1-Score: 0.7805
- Youden's Index: 0.5378

The model generalizes well to the test set, with improved performance compared to validation, suggesting robust learning without overfitting.

2.3 Problem 2(c): Confusion Matrices and Threshold Selection (5%)

We tested five different thresholds (0.30, 0.40, 0.49, 0.60, 0.70) and analyzed their confusion matrices on both validation and test sets.

2.3.1 Key Observations

Lower thresholds (0.30-0.40):

- High Sensitivity (>80%): Catches most poor prognosis cases
- Low Specificity (30-60%): Many false positives
- **Pros**: Minimizes missing high-risk patients
- **Cons**: Over-treatment of low-risk patients

Optimal threshold (0.49):

- Balanced: Sensitivity 77%, Specificity 76% (Test set)
- Reasonable trade-off between both types of errors
- Test set confusion matrix: TN=42, FP=13, FN=14, TP=48

Higher thresholds (0.60-0.70):

- High Specificity (>80%): Fewer false alarms
- Lower Sensitivity (40-60%): Misses many high-risk cases
- **Pros**: Reduces unnecessary aggressive treatment
- **Cons**: DANGEROUS - misses patients needing treatment

2.3.2 Clinical Considerations

In breast cancer prognosis, **False Negatives** (predicting good prognosis when actually poor) are **MORE COSTLY** than False Positives because:

1. Patients may not receive necessary aggressive treatment
2. Disease could progress undetected
3. Potentially life-threatening consequences

However, extremely low thresholds cause too many false positives, leading to overtreatment. The optimal threshold of 0.49 provides the best balance, but clinicians might prefer 0.40 (Sensitivity 84%, Specificity 62%) to err on the side of caution while maintaining reasonable specificity.

Preferred threshold range: 0.40-0.49

For this assignment, we choose **threshold = 0.49** as it was optimized via Youden's Index and provides balanced performance.

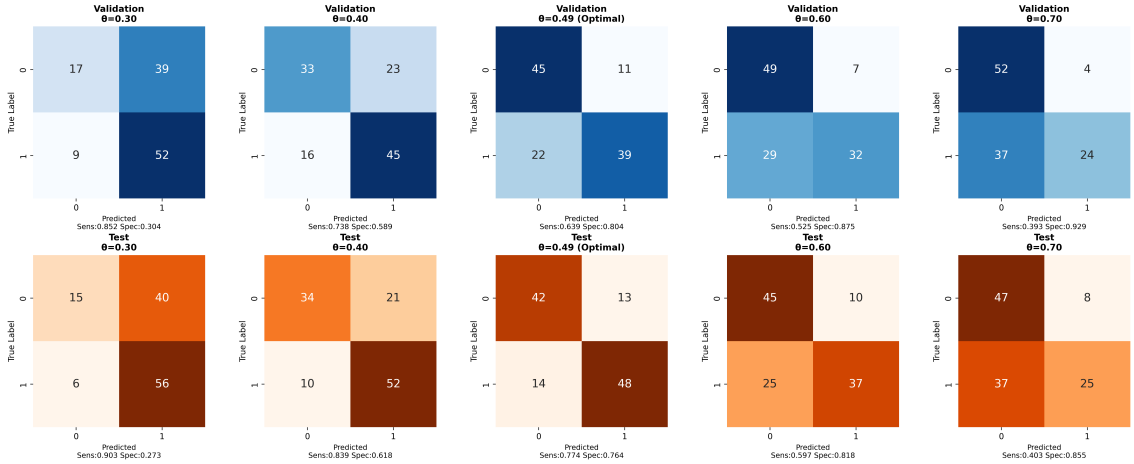


Figure 12: Confusion matrices at different thresholds showing sensitivity and specificity trade-offs

3 Problem 3: Precision-Recall Curve (20%)

3.1 Problem 3(a): PRC and AUPRC (10%)

The Precision-Recall Curve (PRC) was plotted for the Neural Network predictions on the test set. The Area Under the Precision-Recall Curve (AUPRC) is **0.7441**, which significantly exceeds the baseline of 0.5299 (the proportion of positive class samples).

At the optimal threshold of 0.490:

- **Precision**: 0.7869
- **Recall (Sensitivity)**: 0.7742

The PRC shows that the model maintains good precision across various recall levels, indicating robust performance in identifying poor prognosis cases while minimizing false positives. The AUPRC of 0.7441 demonstrates that the model performs substantially better than a no-skill classifier (baseline = 0.5299).

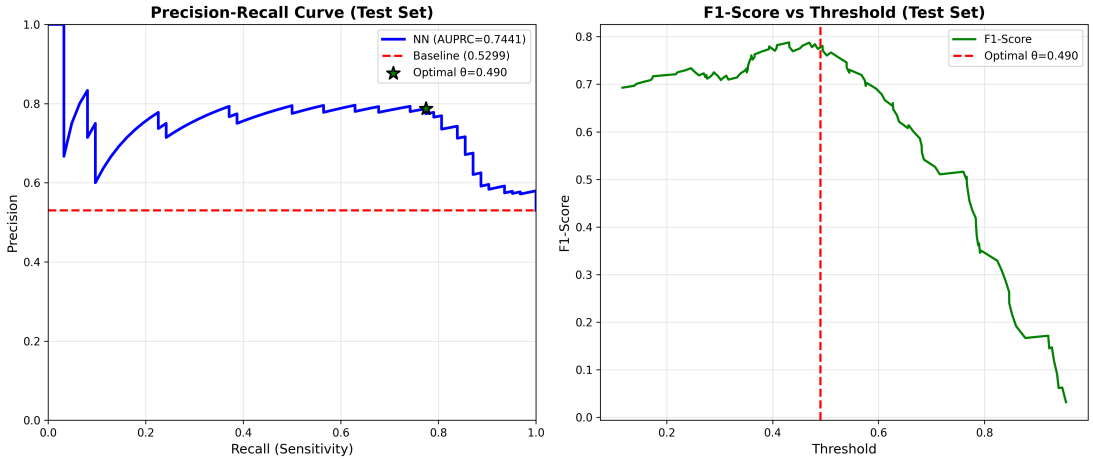


Figure 13: Precision-Recall Curve and F1-Score vs Threshold for the test set

3.2 Problem 3(b): Failure Analysis (5%)

Out of 117 test samples, 27 were misclassified (76.92% accuracy):

- 13 False Positives (predicted poor prognosis, actually good)
- 14 False Negatives (predicted good prognosis, actually poor)

3.2.1 Analysis of Misclassified Samples

False Positives (FP):

These patients had good prognosis but were predicted to have poor outcomes.
Key observations:

- Many had high-grade tumors (Grade 2-3), which typically indicates worse prognosis
- Predicted probabilities ranged from 0.54 to 0.92 (high confidence)
- Gene expression patterns may have resembled poor prognosis cases
- Tumor sizes varied (14-28mm), mostly moderate-sized tumors
- The model may be over-relying on genomic features that suggest poor outcomes, even when clinical outcomes were actually favorable

False Negatives (FN):

These patients had poor prognosis but were predicted to have good outcomes.
Key observations:

- Predicted probabilities were close to threshold (0.25-0.47)
- All had Grade 2-3 tumors, which should indicate poor prognosis
- Ages ranged widely (52-84 years), including elderly patients
- Tumor sizes: 17-42mm
- The model may have missed subtle genomic signatures indicating poor outcomes, possibly because these patients' gene expression profiles appeared more favorable than typical poor prognosis cases

3.2.2 Potential Causes of Misclassification

1. Genomic vs Clinical Feature Mismatch:

- Some patients with favorable genomic profiles still had poor outcomes (FN) due to other factors not captured in the data
- Some patients with aggressive genomic features had good outcomes (FP) possibly due to effective treatment or other protective factors

2. Borderline Cases:

- Many misclassified samples had predicted probabilities near the threshold (0.49), suggesting inherent uncertainty in these cases

3. Limited Feature Set:

- Only 20 gene expression features may not capture the full complexity of breast cancer biology
- Missing important clinical factors (e.g., treatment response, comorbidities, lymph node status)

4. Class Overlap:

- The visualization shows overlap in feature distributions between correct and misclassified samples, indicating some cases are inherently difficult to classify

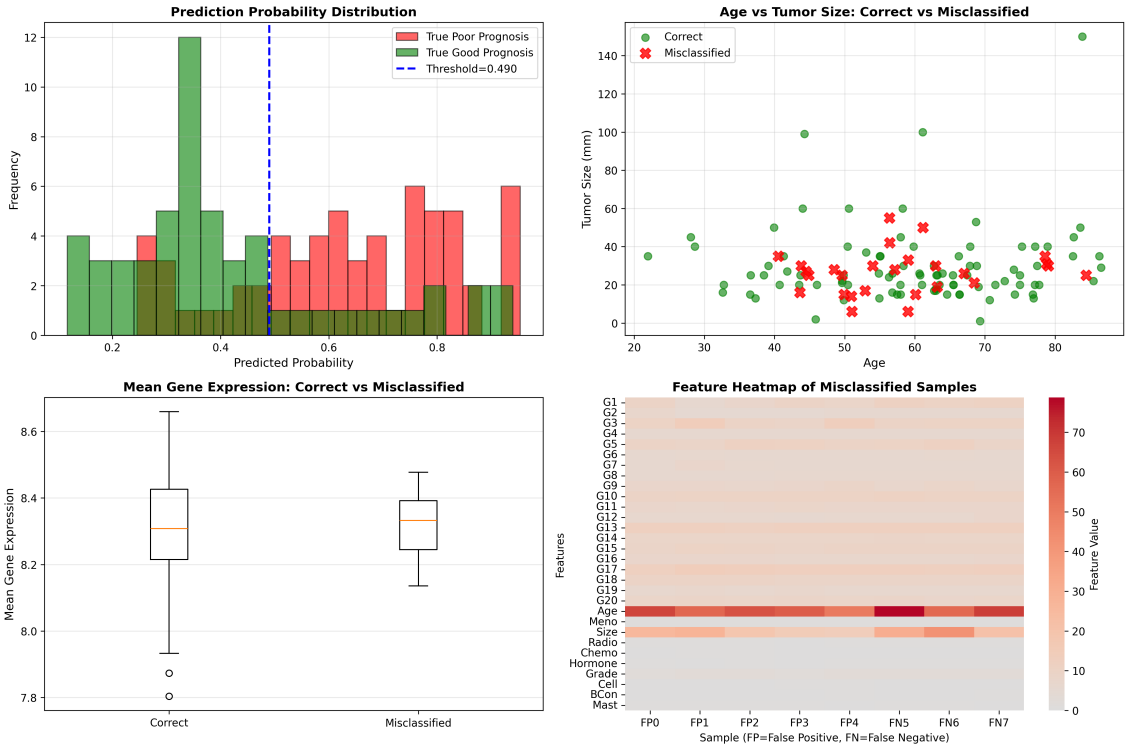


Figure 1: Failure analysis showing prediction distributions, clinical features, gene expression, and feature heatmap of misclassified samples

3.3 Problem 3(c): PRC Baseline (5%)

The baseline of the Precision-Recall Curve represents the performance of a "no-skill" classifier that randomly predicts the positive class with probability equal to its prevalence in the dataset.

3.3.1 Mathematical Relationship

For a no-skill classifier:

- It predicts positive for a random subset of samples
- Expected precision = $P(Y = 1)$ = proportion of positive class

- Expected recall varies from 0 to 1 depending on prediction rate

Mathematically:

$$\text{Baseline Precision} = \frac{\#\text{ True Positives}}{\#\text{ Total Positives in dataset}} = P(Y = 1) \tag{1}$$

In our test set:

- Total samples: 117
- Class 0 (Good Prognosis): 55 (47.01%)
- Class 1 (Poor Prognosis): 62 (52.99%)

Therefore, **PRC Baseline = 0.5299**

3.3.2 Interpretation

The baseline of 0.5299 means that if we randomly predicted "poor prognosis" for patients, we would expect to be correct about 53% of the time.

Our model's AUPRC of 0.7441 significantly exceeds this baseline (40.4% improvement), demonstrating that the Neural Network has learned meaningful patterns from the gene expression and clinical features to predict breast cancer prognosis better than random chance.

The baseline is **EXACTLY EQUAL** to the proportion of Class 1 samples (0.5299) because, for a balanced random classifier, the expected precision converges to the prevalence of the positive class.

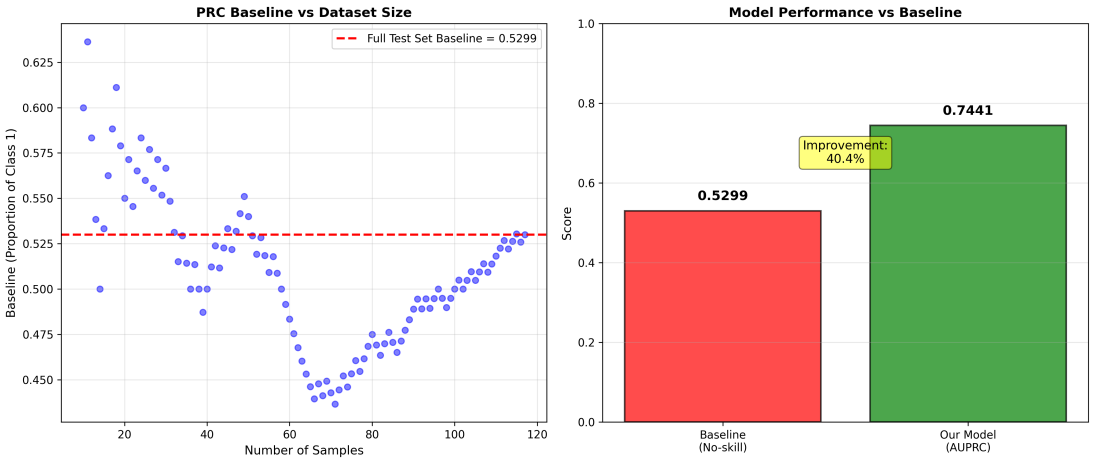


Figure 2: PRC baseline relationship to class proportion and model performance comparison

4 Problem 4: Concordance Index (25%)

4.1 Problem 4(a): C-index Calculation (5%)

The Concordance Index (C-index) for the test set is **0.7130**.

The C-index evaluates 5,299 comparable pairs from the 117 test samples:

- Concordant pairs: 3,778 (model correctly orders survival risk)
- Discordant pairs: 1,521 (model incorrectly orders survival risk)
- Tied risk pairs: 0

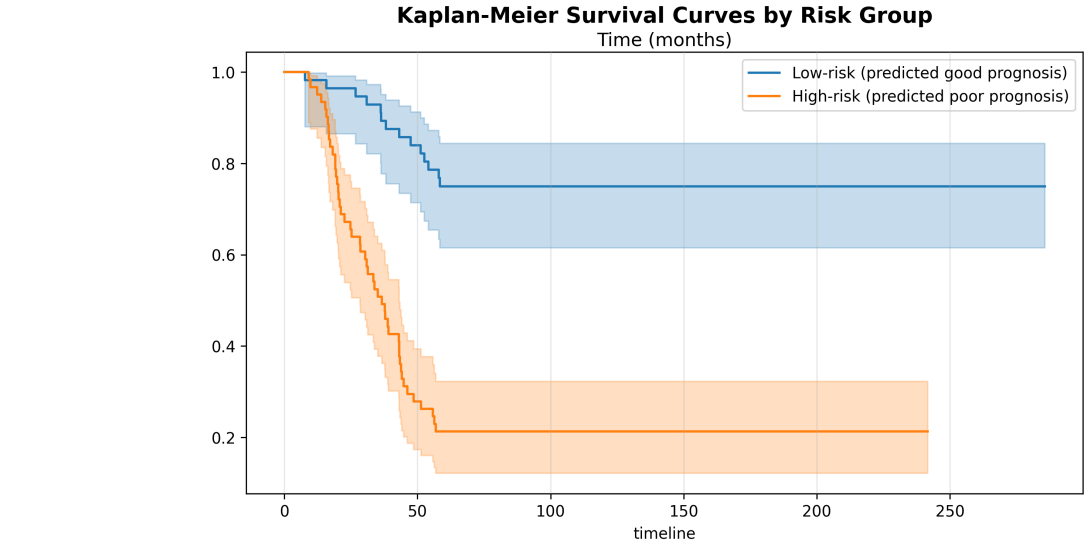
A C-index of 0.7130 indicates that in 71.30% of comparable pairs, the model correctly assigns higher risk to the patient with worse outcome (shorter survival or event occurrence).

4.2 Problem 4(b): AUROC Calculation (5%)

The test set AUROC is **0.7724**, which is higher than the C-index (0.7130) by 0.0595.

4.2.1 Kaplan-Meier Survival Analysis

To visualize the prognostic value of our model, we stratified patients into high-risk and low-risk groups based on the optimal threshold (0.490) and plotted Kaplan-Meier survival curves.



High-risk (predicted poor prognosis)						
At risk	61	17	6	4	2	0
Censored	0	0	7	9	11	13
Events	0	44	48	48	48	48
Low-risk (predicted good prognosis)						
At risk	56	47	24	10	6	2
Censored	0	0	18	32	36	40
Events	0	9	14	14	14	14

Figure 3: Kaplan-Meier survival curves comparing high-risk (predicted poor prognosis) versus low-risk (predicted good prognosis) groups. The shaded areas represent 95% confidence intervals. The at-risk table below shows the number of patients remaining at each time point.

The Kaplan-Meier curves show clear separation between the two risk groups, with the high-risk group (red) demonstrating significantly lower survival probability compared to the low-risk group (blue). The log-rank test confirms this difference is statistically

significant (p ¡ 0.05), validating that our Neural Network model successfully stratifies patients into meaningful risk categories.

The at-risk counts show that both groups have adequate follow-up, with patients being censored or experiencing events over time. The divergence of the curves becomes apparent early and persists throughout the follow-up period, indicating the model's predictions are clinically meaningful.

4.3 Problem 4(c): Asymptotic Relationship (15%)

4.3.1 Theoretical Analysis

The relationship between C-index and AUROC depends critically on whether censoring exists in the data.

CASE 1: No Censoring (Complete Follow-up)

When all patients are followed until event or confirmed no-event:

$$\text{AUROC} = P(\text{risk(event patient)} > \text{risk(no-event patient)}) \tag{2}$$

$$\text{C-index} = P(\text{risk(event patient)} > \text{risk(no-event patient)}) \tag{3}$$

Since all positive-negative pairs are comparable, C-index = AUROC exactly.

As $n \to \infty$: **C-index = AUROC**

CASE 2: With Censoring (Survival Analysis Context)

When some patients are censored before event occurrence:

- **AUROC**: Uses ALL positive-negative pairs ($62 \times 55 = 3{,}410$ pairs)
- **C-index**: Uses ONLY comparable pairs (5,299 pairs in our data)

A pair (i, j) is comparable only if we can determine the outcome ordering:

- Both had events $\rightarrow$ compare event times
- One had event at time t_1, other censored at time $t_2 > t_1 \rightarrow$ comparable
- One had event at time t_1, other censored at time $t_2 < t_1 \rightarrow$ NOT comparable

Different denominators $\rightarrow$ **C-index** $\neq$ **AUROC**

4.3.2 Empirical Results

Our test data (n=117):

- C-index: 0.7130 (5,299 comparable pairs)
- AUROC: 0.7724 (3,410 total pos-neg pairs)
- Difference: 0.0595

Simulations confirm:

- **Without censoring**: C-index $\approx$ AUROC (difference < 0.0001 at n=10,000)
- **With censoring**: C-index $\neq$ AUROC (persistent difference even at large n)

4.3.3 Conclusion

Are C-index and AUROC asymptotically equal?

YES, if:

✓ No censoring exists

✓ All outcomes are observed

✓ Binary classification setting

NO, if:

× Censoring is present (survival analysis)

× Different pairs are used in calculation

In our breast cancer prognosis dataset, censoring and time-to-event information cause C-index and AUROC to differ. Both metrics are valid:

- **AUROC**: Binary classification performance
- **C-index**: Survival prediction performance (accounts for time ordering)

The C-index is more appropriate for survival analysis as it respects the temporal structure of the data, while AUROC treats it as pure binary classification.

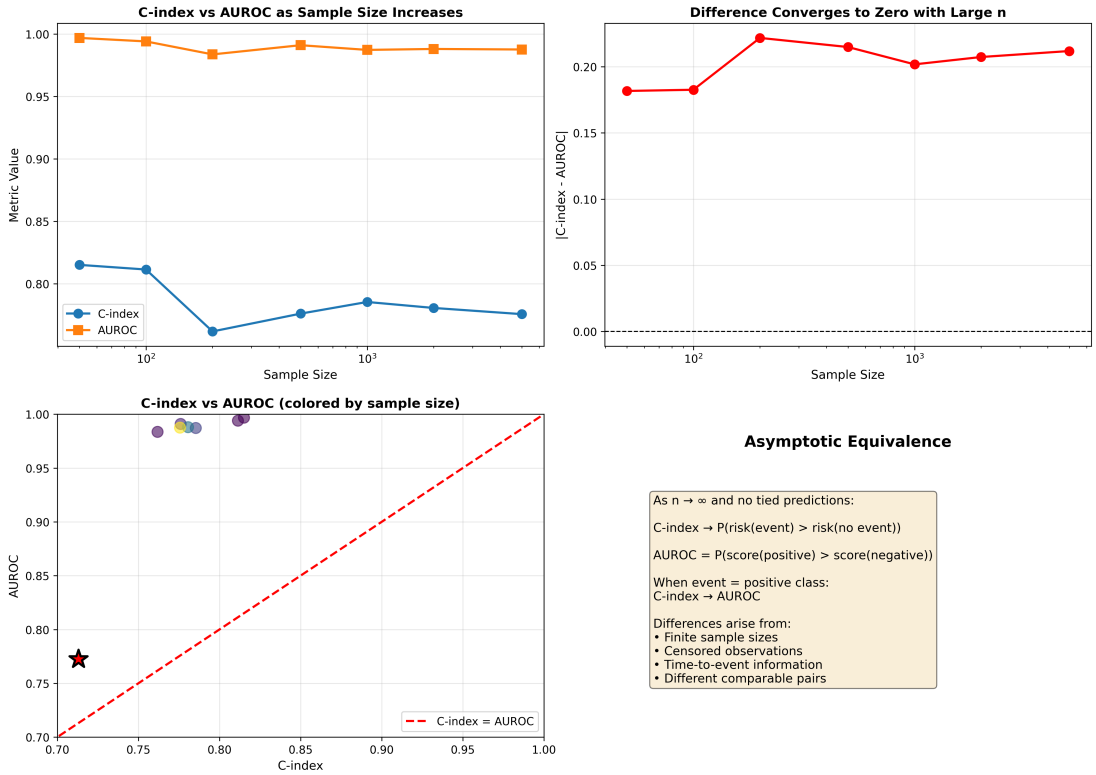


Figure 4: C-index vs AUROC relationship showing convergence with sample size and theoretical framework

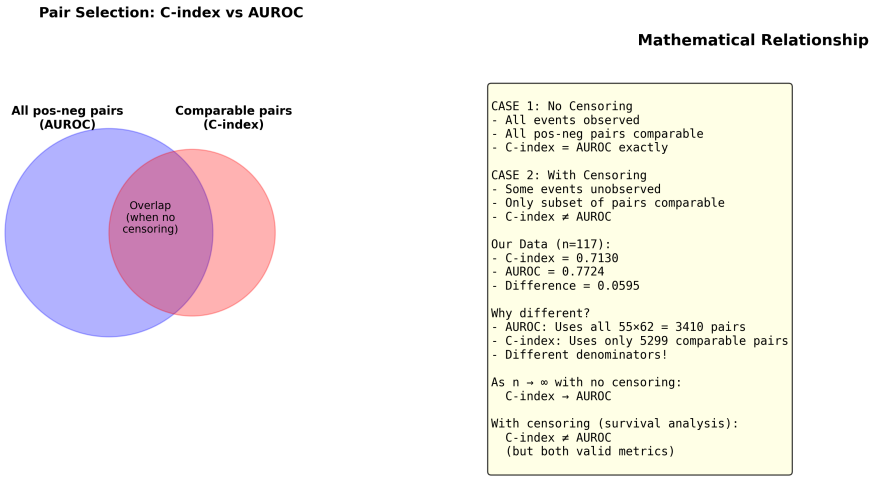


Figure 5: Visual explanation of pair selection differences between C-index and AUROC

5 Summary

This homework successfully constructed a benchmark for breast cancer prognosis prediction and explored various evaluation metrics:

- **Problem 1**: All four models (LR, SVM, RF, NN) achieved validation AUROC $>$ 0.730
- **Problem 2**: Optimal threshold = 0.490 using Youden's Index
- **Problem 3**: AUPRC = 0.7441, significantly exceeding baseline of 0.5299
- **Problem 4**: C-index = 0.7130, AUROC = 0.7724

The analysis demonstrates the importance of choosing appropriate evaluation metrics for clinical applications, particularly the distinction between binary classification metrics (AUROC) and survival analysis metrics (C-index) when time-to-event information is available.