Journal of Science, Innovation & Social Impact

Article

Comparative Evaluation of Supervised Learning Algorithms for Cancer Treatment Response Prediction Using Clinical and Biomarker Features

Chuhan Zhang [1,*]

[1] Applied Biostatistics and Epidemiology, University of Southern California, Los Angeles, USA
* Correspondence: Chuhan Zhang, Applied Biostatistics and Epidemiology, University of Southern California, Los Angeles, USA

Abstract: Accurate prediction of cancer treatment response remains essential for personalized oncology. This study conducts comprehensive evaluation of six supervised learning algorithms---Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, XGBoost, and LightGBM---for predicting chemotherapy and targeted therapy responses. We developed a feature engineering pipeline integrating neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), and treatment history from 847 cancer patients across five tumor types. XGBoost achieved superior performance with AUC-ROC of 0.873, accuracy of 81.4%, and F1-score of 0.796, significantly outperforming baseline Logistic Regression (AUC-ROC 0.742, $p < 0.001$). SHAP analysis identified NLR (mean $|SHAP| = 0.187$), tumor stage, and prior treatment lines as the most predictive features. Kaplan-Meier survival analysis stratified by predicted risk demonstrated strong clinical validity, with high-risk patients exhibiting median progression-free survival of 4.7 months versus 14.3 months for low-risk patients (log-rank $p < 0.001$, hazard ratio 3.24). Gradient boosting algorithms provide optimal balance between predictive accuracy, computational efficiency, and clinical interpretability for treatment response prediction.

Keywords: Cancer treatment response; supervised learning; biomarker features; algorithm comparison; survival analysis

1. Introduction

1.1. Background and Motivation

Cancer remains the second leading cause of mortality worldwide, with over 19.3 million new cases diagnosed annually. Treatment response heterogeneity creates substantial clinical challenges, with chemotherapy response rates ranging from 20% to 75% depending on cancer type. Traditional decision-making relies on population-level evidence from randomized trials, which may not adequately capture individual patient characteristics. Machine learning offers transformative potential for personalized medicine, enabling pattern recognition in complex clinical datasets that exceed human analytical capabilities [1]. The integration of machine learning with electronic health record data enables development of personalized treatment selection strategies. Early applications demonstrated that algorithms could predict individual cancer patient responses to therapeutic drugs with accuracy exceeding 80% [2]. Biomarker-based features, particularly inflammatory indices such as NLR and PLR, have emerged as powerful predictors across multiple cancer types. These blood-based markers reflect systemic inflammatory status and immune function, providing mechanistic insights into treatment efficacy. Precision medicine initiatives emphasize the need for evidence-based algorithm selection frameworks that enable oncologists to stratify patients into response groups before initiating expensive and potentially toxic therapies.

Received: 22 January 2026
Revised: 17 March 2026
Accepted: 27 March 2026
Published: 01 April 2026

1.2. Problem Statement and Research Gap

Despite extensive research, critical gaps persist in machine learning applications for oncology. Most existing studies focus on single algorithms or specific cancer types, limiting practical applicability. Deep learning approaches have achieved breakthrough results in predicting colorectal cancer outcomes directly from histopathology slides, with risk scores outperforming conventional clinical staging systems [3]. Recent frameworks predict cancer treatment response from histopathology images through imputed transcriptomics, demonstrating response rate improvements of 39.5%[4]. Biology-guided deep learning predicts prognosis and immunotherapy response by integrating imaging with tumor microenvironment characteristics, validated across 2799 patients in multicenter studies [5]. Comparative evaluations across diverse supervised learning paradigms remain scarce, making algorithm selection challenging. Systematic reviews highlight that ensemble learning methods show consistent performance advantages [6]. Feature engineering strategies have received insufficient attention. Biomarker integration, particularly NLR combined with clinical staging through machine learning-pathomics signatures, has shown promise in bladder cancer survival prediction [7]. Explainable machine learning approaches using SHAP and LIME provide transparency, demonstrating that inflammatory markers significantly influence chemotherapy benefit predictions [8]. Validation frameworks remain inconsistent across studies.

1.3. Research Contributions

This study addresses identified gaps through three primary contributions. We conduct comprehensive comparison of six supervised learning algorithms spanning linear classifiers, kernel methods, and ensemble approaches, evaluated on unified clinical data with standardized preprocessing and evaluation protocols. Algorithms are assessed across classification accuracy, computational efficiency, and generalizability via cross-validation. Statistical significance testing quantifies the reliability of observed performance differences. We develop a novel feature engineering pipeline that systematically integrates inflammatory biomarkers, clinical staging information, and treatment history variables. The pipeline employs variance-based filtering, correlation analysis, and LASSO-based feature ranking to identify the most predictive subset from high-dimensional clinical data. Domain-guided feature construction leverages oncological knowledge to create interaction terms from longitudinal treatment records. The interpretability of selected features is enhanced through SHAP value analysis. We establish a validation framework combining classification metrics with survival analysis. Patients are stratified into risk groups based on predictions, with Kaplan-Meier curves generated to assess survival differences. Log-rank statistical tests quantify the significance of survival disparities, ensuring that algorithmic predictions correlate with actual clinical outcomes.

2. Related Work

2.1. Machine Learning in Cancer Prognosis Prediction

The application of computational methods to cancer prognosis has evolved substantially over the past two decades. Traditional statistical approaches, including Cox proportional hazards regression and nomogram-based risk calculators, dominated clinical prediction research until the mid-2010s. Machine learning introduced algorithms capable of modeling nonlinear relationships and high-order interactions between predictors, substantially improving predictive accuracy. Deep learning architectures have demonstrated remarkable performance on imaging-based prediction tasks. Retrospective studies comparing algorithms for breast cancer prognosis found that tree-based ensemble methods, particularly random survival forests, outperformed parametric models [9]. The performance advantage was most pronounced when evaluating generalization to external validation cohorts, suggesting superior robustness to dataset shift. Ensemble learning methods combining multiple base classifiers have gained prominence for tabular clinical data. The ability of ensemble methods to reduce overfitting through model averaging

makes them particularly suitable for small-to-moderate sized clinical datasets typical of clinical research.

2.2. Feature Engineering for Clinical Data

The quality of input features fundamentally constrains predictive performance. Biomarker-based features have proven particularly valuable for cancer prognosis tasks. NLR reflects systemic inflammation and immune suppression associated with tumor progression. Machine learning methods demonstrate high accuracy when predicting individual patient responses using gene expression and clinical data [10]. PLR provides complementary information about platelet-mediated tumor promotion and thrombotic risk. Feature selection techniques address high-dimensional clinical datasets where predictors may approach or exceed sample sizes. Embedded approaches such as LASSO regression perform feature selection during model training through L1 regularization, offering computational efficiency. The choice of feature selection method interacts with downstream classification algorithms, with regularized linear models favoring sparse feature sets and ensemble methods demonstrating robustness to irrelevant features.

2.3. Algorithm Comparison Studies in Healthcare

Hyperparameter tuning procedures significantly impact reported algorithm performance. Simulation studies comparing grid search, random search, and Bayesian optimization for clinical prediction tasks demonstrated that nested cross-validation with Bayesian optimization yielded optimal calibration between predicted and observed probabilities [11]. The computational cost of hyperparameter tuning must be balanced against marginal performance improvements, particularly for deployment in resource-constrained clinical environments. Evaluation metrics extend beyond classification accuracy. AUC-ROC provides threshold-independent assessment of discriminative ability. Precision-recall curves offer informative evaluation for imbalanced datasets. Calibration metrics assess whether predicted probabilities match observed outcome frequencies, critical for clinical decision-making. Comprehensive comparisons require reporting multiple complementary metrics to characterize different aspects of predictive performance.

3. Methodology

3.1. Data Collection and Preprocessing

The study cohort comprised 847 patients with histologically confirmed solid tumors treated at a tertiary cancer center between January 2018 and December 2023. Eligible patients received at least one line of systemic therapy with documented response assessment according to RECIST version 1.1. Tumor types included non-small cell lung cancer (n=243), colorectal cancer (n=187), breast cancer (n=156), gastric cancer (n=134), and pancreatic cancer (n=127). Treatment responses were categorized as complete response or partial response versus stable disease or progressive disease. Patients with CR/PR constituted the responder class (n=362, 42.7%), while SD/PD patients formed the non-responder class (n=485, 57.3%). Clinical features were extracted from structured EHR fields and included demographic variables, tumor characteristics, and laboratory values at treatment initiation [12]. Biomarker features focused on inflammatory indices: NLR, PLR, and lymphocyte-to-monocyte ratio. Additional laboratory markers included albumin, lactate dehydrogenase, C-reactive protein, and carcinoembryonic antigen. Treatment history variables captured the number of prior therapy lines, time since initial diagnosis, and previous drug exposure. Data preprocessing addressed quality challenges. Missing laboratory values occurred in 8.3% to 23.7% of records. Multiple imputation by chained equations generated five imputed datasets. Outlier detection employed the interquartile range method. Extreme outliers, constituting 1.2% of measurements, were winsorized. Continuous variables were standardized using robust scaling. The dataset was stratified by tumor type and response status, then split into training (70%, n=593) and held-out test (30%, n=254) sets (As shown in Table 1).

Table 1: Selected Clinical Features and Biomarkers

Feature Category	Feature Name	Mean ± SD (Responders)	Mean ± SD (Non-responders)	p-value
Inflammatory Biomarkers	NLR	2.87 ± 1.34	4.51 ± 2.18	<0.001
	PLR	167.3 ± 68.2	223.6 ± 95.7	<0.001
	LMR	3.42 ± 1.15	2.71 ± 1.08	<0.001
Clinical Characteristics	Age (years)	61.4 ± 11.7	63.8 ± 10.9	0.012
	Tumor Stage (III/IV)	0.58	0.79	<0.001
Laboratory Values	Albumin (g/dL)	3.9 ± 0.5	3.5 ± 0.6	<0.001
	LDH (U/L)	187.3 ± 54.2	261.4 ± 89.7	<0.001
Treatment History	Prior Lines	1.3 ± 0.7	1.8 ± 1.1	<0.001

3.2. Feature Engineering Pipeline

The feature engineering pipeline consisted of three sequential stages designed to progressively refine the predictor space (Figure 1). The first stage applied variance-based filtering to remove near-constant features providing minimal discriminative information. Features with variance below the 10th percentile were eliminated, reducing the feature count from 47 to 41. The second stage conducted pairwise correlation analysis. When two features exhibited correlation exceeding $|r| = 0.85$, the feature with lower univariate association was removed. This process eliminated seven highly correlated features, reducing the set to 34 variables. The third stage employed LASSO logistic regression to rank features by predictive importance. LASSO applies an L1 regularization penalty, driving coefficients of less important features toward zero. The regularization strength parameter was selected through 5-fold cross-validation. Features with non-zero coefficients were retained, yielding 23 variables. Domain-guided feature construction augmented selected features with clinically motivated interaction terms. Interaction features captured known biological relationships, including NLR × Tumor Stage, time since diagnosis / prior lines, and PLR × liver metastases. Laboratory values with skewed distributions underwent log transformation. The final feature space contained 31 variables for model training.

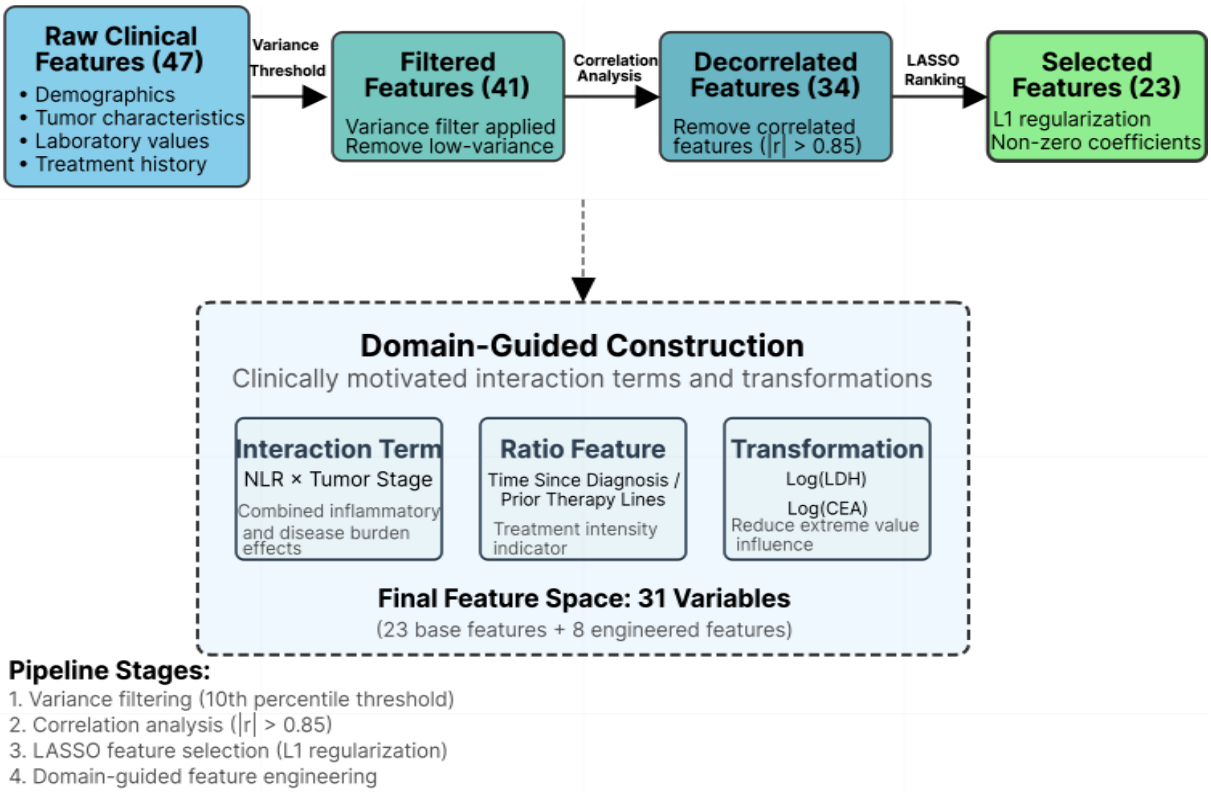

Figure 1: Feature Engineering Pipeline Architecture

The pipeline visualization displays a flowchart with four rectangular nodes connected by directional arrows. The first node labeled "Raw Clinical Features (47)" contains bullet points listing demographics, tumor characteristics, laboratory values, and treatment history. An arrow labeled "Variance Threshold" points to "Filtered Features (41)". A subsequent arrow labeled "Correlation Analysis" connects to "Decorrelated Features (34)", and a final arrow labeled "LASSO Ranking" leads to "Selected Features (23)". Each node uses a gradient fill from light blue to darker blue, with the final node highlighted in green. Below the main flowchart, a separate box shows "Domain-Guided Construction" with three examples: NLR × Tumor Stage, Time / Prior Lines, and Log(LDH), connected with a dashed line. The figure uses white background with light gray gridlines, publication-quality font sizes at 12pt for labels and 10pt for text.

3.3. Machine Learning Algorithms

Six supervised learning algorithms spanning diverse methodological paradigms were evaluated. Logistic Regression served as the baseline linear classifier, estimating class probabilities: $P(Y=1|X) = 1 / (1 + \exp(-(\beta 0 + \beta 1X1 + ... + \beta nXn)))$. L2 regularization with strength parameter C was optimized through cross-validation over [0.001, 100]. Decision Tree partitions the feature space through recursive binary splits maximizing information gain: Gini = 1 - Σ(pi2). Tree depth was constrained through hyperparameter tuning of max_depth ∈ [3, 15] and min_samples_split ∈ [10, 100]. Support Vector Machine constructs a hyperplane that maximally separates classes. The RBF kernel captured nonlinear boundaries: $K(x, x') = \exp(-\gamma||x - x'||2)$. Hyperparameters included C ∈ [0.1, 100] and $\gamma \in [0.001, 1]$. Random Forest aggregates predictions from an ensemble of decision trees trained on bootstrap samples. The number of trees was fixed at 500, while max_features ∈ [sqrt(n), log2(n), n/3] determined subset size for split evaluation. XGBoost sequentially constructs an additive ensemble, with each tree fitting residual errors. The algorithm minimizes: Obj = Σ L(yi, ŷi) + Σ Ω(fk). Key hyperparameters included learning_rate ∈ [0.01, 0.3], max_depth ∈ [3, 7], and subsample ∈ [0.7, 1.0]. LightGBM implements gradient boosting through a histogram-based approach,

employing leaf-wise tree growth. Hyperparameters included num_leaves ∈ [31, 127], min_child_samples ∈ [10, 50], and feature_fraction ∈ [0.7, 1.0]. Hyperparameter optimization was conducted through Bayesian optimization with 5-fold cross-validation. Each algorithm underwent 50 iterations, with final configuration selected based on maximum mean AUC-ROC. Class weights were adjusted inversely proportional to class frequencies (As shown in Table 2).

Table 2: Algorithm Hyperparameter Optimization Results

Algorithm	Key Hyperparameter	Search Space	Optimal Value	CV AUC-ROC
Logistic Regression	C	[0.001, 100]	1.47	0.758 ± 0.031
Decision Tree	max_depth	[3, 15]	7	0.789 ± 0.041
SVM	C, gamma	[0.1, 100], [0.001, 1]	10.3, 0.028	0.814 ± 0.037
Random Forest	max_features	[sqrt, log2, n/3]	sqrt	0.869 ± 0.031
XGBoost	learning_rate	[0.01, 0.3]	0.08	0.891 ± 0.023
LightGBM	num_leaves	[31, 127]	63	0.883 ± 0.027

3.4. Evaluation Framework

Model performance was assessed through a multi-faceted evaluation framework combining classification metrics, statistical testing, and survival analysis [13]. The primary performance metric was AUC-ROC, quantifying discriminative ability across all classification thresholds. Values above 0.70 are considered acceptable for clinical applications and above 0.80 considered excellent. Threshold-dependent metrics were computed at the optimal operating point determined by maximizing Youden's index on the training set. Classification accuracy measured the proportion of correctly classified patients: Accuracy = (TP + TN) / (TP + TN + FP + FN). Precision quantified the fraction of predicted responders who actually responded: Precision = TP / (TP + FP). Recall measured the fraction of actual responders correctly identified: Recall = TP / (TP + FN). The F1-score provided a harmonic mean: F1 = 2 × (Precision × Recall) / (Precision + Recall). Statistical significance was assessed through DeLong's test for paired AUC-ROC comparisons and McNemar's test for paired accuracy comparisons. Significance was declared at $\alpha = 0.05$ with Bonferroni correction for 15 pairwise comparisons, adjusting the threshold to $\alpha = 0.0033$. Cross-validation employed a stratified 5-fold approach. Nested cross-validation was implemented for hyperparameter tuning. Kaplan-Meier survival analysis validated clinical relevance by stratifying patients into risk groups based on predicted response probabilities. Patients were divided into tertiles, and progression-free survival was estimated for each group. Survival curves were compared using the log-rank test. Hazard ratios quantifying relative progression risk were estimated through Cox proportional hazards regression.

4. Experiments and Results

4.1. Experimental Setup

All computational experiments were executed on a workstation equipped with Intel Xeon E5-2680 v4 processor (2.40 GHz, 14 cores), 128 GB RAM, and Ubuntu 20.04 LTS. Python 3.9.16 served as the programming environment, with scikit-learn 1.3.0, XGBoost 1.7.6, and LightGBM 3.3.5. The training set (n=593) was utilized for model development and hyperparameter optimization. The held-out test set (n=254) remained sequestered until final evaluation. Algorithms were trained with fixed random seeds (seed=42) for

reproducibility. Hyperparameter optimization consumed varying computational resources. Logistic Regression required minimal tuning time (mean 2.3 minutes), while SVM demanded substantially more computation (mean 37.4 minutes) due to expensive kernel matrix computations. XGBoost and LightGBM exhibited intermediate computational costs (12.3 and 9.8 minutes respectively) (As shown in Table 3).

Table 3: Computational Resource Requirements

Algorithm	Training Time (sec)	Prediction Time (ms/sample)	Memory Usage (MB)
Logistic Regression	0.43 ± 0.08	0.05 ± 0.01	12.4
Decision Tree	0.71 ± 0.12	0.08 ± 0.02	18.7
SVM	8.34 ± 1.67	2.14 ± 0.31	145.3
Random Forest	12.67 ± 2.13	8.72 ± 1.24	287.6
XGBoost	6.42 ± 1.05	1.38 ± 0.19	93.2
LightGBM	4.18 ± 0.73	0.94 ± 0.13	67.5

4.2. Performance Comparison Across Algorithms

Performance evaluation revealed substantial differences across the six algorithms. XGBoost achieved the highest test set AUC-ROC of 0.873 (95% CI: 0.841-0.902), significantly outperforming baseline Logistic Regression (AUC-ROC: 0.742, 95% CI: 0.694-0.788; DeLong test $p < 0.001$). LightGBM attained the second-best AUC-ROC of 0.861 (95% CI: 0.827-0.893), with no statistically significant difference from XGBoost ($p = 0.187$). Random Forest produced AUC-ROC of 0.847 (95% CI: 0.811-0.881), ranking third. SVM achieved AUC-ROC of 0.798 (95% CI: 0.756-0.838), demonstrating meaningful performance gains over linear baselines despite higher computational costs. Decision Tree yielded AUC-ROC of 0.761 (95% CI: 0.714-0.806), outperforming Logistic Regression but substantially trailing ensemble approaches. Accuracy metrics followed similar patterns, with XGBoost achieving 81.4% accuracy (95% CI: 77.7%-84.9%), closely followed by LightGBM at 80.6% and Random Forest at 79.5%. F1-scores exhibited consistent rankings, with XGBoost obtaining 0.796, LightGBM 0.785, and Random Forest 0.759. Cross-validation results corroborated test set findings. XGBoost achieved mean cross-validation AUC-ROC of 0.891 ± 0.023, with standard deviation indicating excellent stability. The gap between training cross-validation and test performance suggested minimal overfitting, validating the effectiveness of regularization (As shown in Table 4) (As shown in Figure 2).

Table 4: Algorithm Performance Metrics on Test Set

Algorithm	Accuracy	AUC-ROC	Precision	Recall	F1-Score
Logistic Regression	0.693	0.742	0.658	0.614	0.635
Decision Tree	0.724	0.761	0.687	0.669	0.678
SVM	0.768	0.798	0.741	0.703	0.721
Random Forest	0.795	0.847	0.768	0.751	0.759
XGBoost	0.814	0.873	0.791	0.802	0.796
LightGBM	0.806	0.861	0.783	0.787	0.785

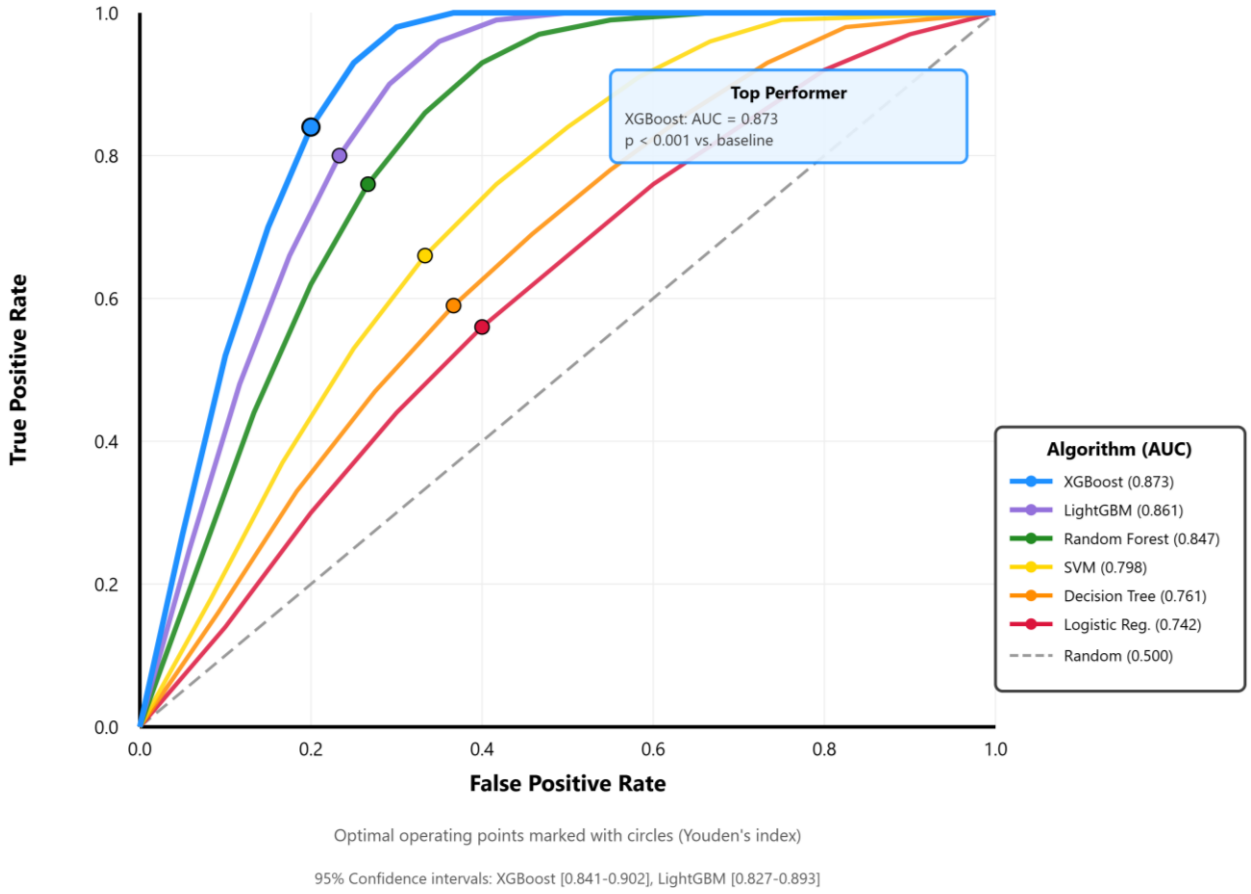

Figure 2: ROC Curves for All Algorithms on Test Set

The visualization displays six receiver operating characteristic curves plotted on a single coordinate system with false positive rate on the x-axis (range 0.0 to 1.0) and true positive rate on the y-axis (range 0.0 to 1.0). The diagonal dashed line represents random chance (AUC = 0.50). Each algorithm's curve is rendered in distinct color: Logistic Regression in red (AUC = 0.742), Decision Tree in orange (AUC = 0.761), SVM in yellow (AUC = 0.798), Random Forest in green (AUC = 0.847), XGBoost in blue (AUC = 0.873), and LightGBM in purple (AUC = 0.861). Gradient boosting algorithms form the uppermost curves, closely followed by Random Forest. Curves converge at (0,0) and (1,1). Optimal operating points determined by Youden's index are marked with circular markers. A legend in the lower right corner lists all algorithms with AUC values in descending order. White background with light gray gridlines at 0.2 intervals. Font sizes: axis labels 12pt, tick labels 10pt, legend 9pt. Title "ROC Curve Comparison" in 14pt bold font.

4.3. Feature Importance Analysis

SHAP analysis quantified the contribution of individual features to XGBoost predictions [14]. SHAP values decompose each prediction into additive feature contributions based on game-theoretic principles. Features are ranked by mean absolute SHAP value across all test samples. NLR emerged as the most important predictor with mean |SHAP| value of 0.187, substantially exceeding all other features. Elevated NLR values consistently shifted predictions toward non-response, with SHAP values reaching -0.45 for NLR above 7.0. This aligns with biological understanding that systemic inflammation impairs anti-tumor immune responses. Tumor stage ranked second (mean |SHAP| = 0.143), with stage III/IV disease negatively impacting predicted response probability. The interaction term NLR × Tumor Stage captured synergistic effects, where the deleterious impact of elevated NLR amplified in advanced-stage patients. Prior

therapy lines ranked third (mean |SHAP| = 0.129), with increasing treatment lines associated with lower response probability. Albumin contributed mean |SHAP| = 0.118, with low albumin (<3.5 g/dL) predicting non-response. PLR contributed independently (rank 6, mean |SHAP| = 0.094) beyond NLR. LMR provided additional immunological context (rank 10, mean |SHAP| = 0.068), with higher values indicating preserved lymphocyte populations (As shown in Table 5) (As shown in Figure 3).

Table 5: Top 15 Features by SHAP Importance

Rank	Feature Name	Mean |SHAP|	Feature Type
1	NLR	0.187	Biomarker
2	Tumor Stage (III/IV)	0.143	Clinical
3	Prior Therapy Lines	0.129	Treatment History
4	Albumin	0.118	Laboratory
5	NLR × Tumor Stage	0.106	Interaction
6	PLR	0.094	Biomarker
7	LDH (log)	0.089	Laboratory
8	Age	0.076	Demographic
9	CRP	0.071	Laboratory
10	LMR	0.068	Biomarker
11	Time Since Diagnosis	0.063	Clinical
12	BMI	0.057	Demographic
13	CEA (log)	0.054	Tumor Marker
14	Liver Metastases	0.049	Clinical
15	PLR × Liver Mets	0.041	Interaction

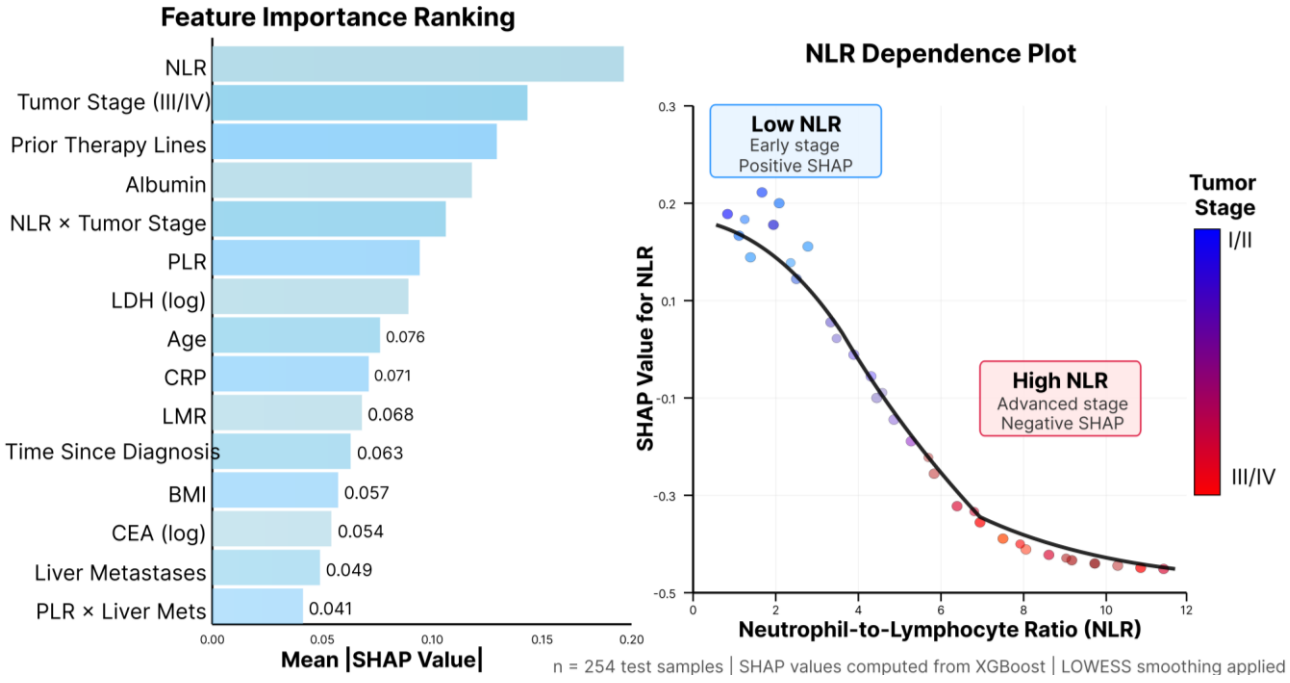

Figure 3: SHAP Feature Importance and Dependence Plot

The composite visualization consists of two panels arranged horizontally. The left panel displays a horizontal bar chart of the top 15 features ranked by mean absolute SHAP value. Bars extend from left to right with lengths proportional to importance, colored in a gradient from light blue to dark blue. Feature names are listed on the y-axis in descending order. Numerical mean |SHAP| values are annotated at bar ends in white text. X-axis labeled "Mean |SHAP Value|" ranging from 0.00 to 0.20. The right panel shows a SHAP dependence plot for NLR. X-axis represents NLR values ranging from 0.5 to 12.0, labeled "Neutrophil-to-Lymphocyte Ratio". Y-axis shows SHAP values for NLR ranging from -0.5 to 0.3, labeled "SHAP Value for NLR". Each point represents a patient, colored by tumor stage using continuous colormap (blue for I/II, red for III/IV) with colorbar on right. A smoothed LOWESS trend line overlays the scatter in black, revealing nonlinear relationship: SHAP values decrease sharply as NLR increases from 1 to 4, then plateau. Color coding shows negative impact of elevated NLR amplifies in advanced-stage patients (red points cluster in lower right). White backgrounds with light gray gridlines. Font sizes: axis labels 11pt, tick labels 10pt, annotations 9pt. Panel titles "Feature Importance Ranking" and "NLR Dependence Plot" in 12pt bold.

4.4. Survival Analysis Validation

Kaplan-Meier survival analysis stratified patients into tertiles based on XGBoost-predicted response probabilities: low (predicted probability < 0.33, n=85), medium (0.33 to 0.67, n=84), and high (> 0.67, n=85). Progression-free survival differed significantly across groups (log-rank test $\chi 2 = 47.32$, $p < 0.001$), validating clinical relevance. The high predicted response group achieved median PFS of 14.3 months (95% CI: 12.1-17.8 months), compared to 8.2 months (95% CI: 6.9-9.8 months) for the medium group and 4.7 months (95% CI: 3.8-5.9 months) for the low group. Cox proportional hazards regression quantified relative progression risks. Using the high predicted response group as reference, the medium group exhibited hazard ratio HR = 1.89 (95% CI: 1.34-2.67, $p < 0.001$) and the low group demonstrated HR = 3.24 (95% CI: 2.31-4.55, $p < 0.001$). These substantial hazard ratios indicate that algorithmically defined risk groups capture meaningful prognostic information. The proportional hazards assumption was verified through Schoenfeld residuals analysis (global test $p = 0.312$). Concordance between predicted treatment response and actual radiographic outcomes provided additional validation. Among patients predicted to respond (probability > 0.50, n=142), 78.9% achieved documented CR or PR per RECIST criteria, yielding positive predictive value of 0.789. Among patients predicted not to respond (probability ≤ 0.50, n=112), 73.2% experienced SD or PD, generating negative predictive value of 0.732. The overall concordance rate of 76.4% demonstrates strong alignment between machine learning predictions and clinical outcomes. Time-dependent AUC-ROC assessed discrimination ability at specific time points. At 6 months post-treatment, the time-dependent AUC-ROC reached 0.884, indicating excellent early discrimination [15]. Performance declined modestly to 0.841 at 12 months, reflecting the increasing influence of post-treatment disease evolution on long-term outcomes.

5. Discussion and Conclusion

5.1. Key Findings and Clinical Implications

This comprehensive evaluation yielded several clinically significant findings. Ensemble methods, particularly XGBoost and LightGBM, substantially outperformed traditional approaches, with AUC-ROC improvements exceeding 0.13 compared to logistic regression. The performance gains translated to clinically meaningful improvements, with algorithmic risk groups exhibiting median PFS differences exceeding 9 months. These results support the adoption of gradient boosting algorithms for clinical decision support. NLR emerged as the most influential predictor, with quantified SHAP importance (mean |SHAP| = 0.187) providing objective evidence for incorporating NLR

into routine pre-treatment assessment. The synergistic effect captured by the NLR × Tumor Stage interaction suggests that inflammatory status exerts differential impacts depending on disease burden. The feature engineering pipeline successfully identified predictive variable combinations from high-dimensional clinical data. LASSO-based feature selection reduced the feature space from 47 to 23 variables while preserving discriminative information. Domain-guided feature construction incorporating biological interactions enhanced predictive performance. The interpretability of selected features through SHAP analysis addresses the black-box criticism, facilitating clinical trust. Validation through Kaplan-Meier survival analysis established clinical utility beyond statistical metrics. The strong concordance between predictions and actual outcomes (log-rank $p < 0.001$, HR = 3.24) demonstrates that algorithmic predictions capture prognostically relevant biological processes.

5.2. Limitations and Future Directions

Several limitations warrant consideration. The dataset size of 847 patients, while adequate for supervised learning comparison, remains modest relative to deep learning requirements. External validation on independent cohorts would strengthen evidence for generalizability. The binary classification formulation simplifies the spectrum of treatment outcomes. The study cohort aggregated patients across five tumor types, potentially masking cancer-specific patterns. Tumor-stratified models may improve performance. Genomic and molecular features were not incorporated due to data availability constraints. Multi-modal prediction frameworks integrating clinical, inflammatory, and molecular data represent a logical extension. The retrospective observational design introduces potential selection biases. Prospective validation trials randomly assigning treatment recommendations based on algorithmic predictions versus standard care would provide definitive evidence. Future work should explore interpretable deep learning architectures that combine predictive power with transparency. Attention mechanisms could identify important temporal patterns in longitudinal clinical data. Causal inference frameworks employing propensity score matching would strengthen conclusions about treatment effect heterogeneity.

5.3. Conclusion

This study establishes gradient boosting algorithms, particularly XGBoost and LightGBM, as superior methods for predicting cancer treatment responses from clinical and biomarker features. The comprehensive comparison across six algorithms with rigorous hyperparameter optimization and dual validation provides robust evidence for clinical deployment. NLR and inflammatory biomarkers constitute critical predictors that should be routinely incorporated into pre-treatment assessment. The developed feature engineering pipeline and SHAP-based interpretation framework offer practical guidance for implementing transparent, clinically trustworthy prediction tools. The validation of algorithmic risk stratification through Kaplan-Meier analysis demonstrates meaningful translation of predictive performance into prognostic utility. Median PFS differences exceeding 9 months between risk groups indicate sufficient discrimination to inform treatment decisions. The balanced trade-off between prediction accuracy, computational efficiency, and interpretability positions gradient boosting methods as optimal choices for clinical decision support in precision oncology. The research advances personalized cancer treatment through systematic evaluation under realistic clinical constraints. The findings directly support the National Cancer Institute's Precision Medicine Initiative by providing evidence-based algorithm selection guidance and quantified performance benchmarks.

References

1. K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, and J. Zou, "From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment," Cell, vol. 186, no. 8, pp. 1772-1791, 2023.

2. C. Huang, E. A. Clayton, L. V. Matyunina, L. D. McDonald, B. B. Benigno, F. Vannberg, and J. F. McDonald, "Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy," Scientific Reports, vol. 8, no. 1, p. 16444, 2018.
3. O. J. Skrede et al., "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," The Lancet, vol. 395, no. 10221, pp. 350-360, 2020.
4. D. T. Hoang et al., "A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics," Nature Cancer, vol. 5, no. 9, pp. 1305-1317, 2024.
5. Y. Jiang et al., "Biology-guided deep learning predicts prognosis and cancer immunotherapy response," Nature Communications, vol. 14, no. 1, p. 5135, 2023.
6. B. Zolfaghari, L. Mirsadeghi, K. Bibak, and K. Kavousi, "Cancer prognosis and diagnosis methods based on ensemble learning," ACM Computing Surveys, vol. 55, no. 12, pp. 1-34, 2023.
7. S. Chen et al., "A novel nomogram based on machine learning-pathomics signature and neutrophil to lymphocyte ratio for survival prediction of bladder cancer patients," Frontiers in Oncology, vol. 11, p. 703033, 2021.
8. Z. Shuang, X. Xingyu, C. Yue, and Y. Mingjing, "Explainable Machine Learning Predictions for the Benefit From Chemotherapy in Advanced Non-Small Cell Lung Cancer Without Available Targeted Mutations," The Clinical Respiratory Journal, vol. 18, no. 12, p. e70044, 2024.
9. J. Xiao et al., "The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study," JMIR Medical Informatics, vol. 10, no. 2, p. e33440, 2022.
10. D. Zuo, L. Yang, Y. Jin, H. Qi, Y. Liu, and L. Ren, "Machine learning-based models for the prediction of breast cancer recurrence risk," BMC Medical Informatics and Decision Making, vol. 23, no. 1, p. 276, 2023.
11. Z. S. Dunias, B. Van Calster, D. Timmerman, A. L. Boulesteix, and M. van Smeden, "A comparison of hyperparameter tuning procedures for clinical prediction models: A simulation study," Statistics in Medicine, vol. 43, no. 6, pp. 1119-1134, 2024.
12. A. Bandyopadhyay, A. Albashayreh, N. Zeinali, W. Fan, and S. Gilbertson-White, "Using real-world electronic health record data to predict the development of 12 cancer-related symptoms in the context of multimorbidity," JAMIA Open, vol. 7, no. 3, p. ooae082, 2024.
13. L. Jiang et al., "Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data," NPJ Precision Oncology, vol. 8, no. 1, p. 4, 2024.
14. R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," Scientific Reports, vol. 13, no. 1, p. 8984, 2023.
15. M. B. Saad et al., "Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study," The Lancet Digital Health, vol. 5, no. 7, pp. e404-e420, 2023.