Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 9, No. 1, March 2026, pp. 27 – 38
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

CT Radiomics and Ensemble Learning for 5-Year Survival Prediction in Colorectal Liver Metastases

[1*]Widya Astuti, [2]Catur Edi Widodo, [3]Qidir Maulana Binu Soesanto
[1,2,3]Department of Physics, Faculty of Science and Mathematics,
Diponegoro University, Central Java, Indonesia
Email: [1]widyastuti1320@gmail.com,
[2]caturediwidodo@lecturer.undip.ac.id, [3]qidirbinu@fisika.fsm.undip.ac.id

Article Info

Article history:
Received Jan 28th, 2026
Revised Feb 16th, 2026
Accepted Feb 25th, 2026

Keyword:
CT Imaging
Liver Metastasis
Machine Learning
Radiomics
Survival Prediction

ABSTRACT

Colorectal Liver Metastases (CRLM) significantly impact patient survival with high recurrence rates. Traditional prognostic models often overlook tumor heterogeneity, leading to suboptimal risk stratification. To address this, radiomics was employed to quantify sub-visual tumor phenotypes, while ensemble learning was selected to robustly handle high-dimensional feature complexity and improve generalization capability. This retrospective study analyzed 145 CRLM patients from The Cancer Imaging Archive, extracting 1130 radiomics features from preoperative CT scans alongside clinical variables. Data were split into training (n=101) and testing (n=44) sets, with feature selection reducing the input to 16 key features. Three ensemble models (XGBoost, LightGBM, Random Forest) were optimized using Optuna, incorporating SMOTE and isotonic calibration. On the test set, XGBoost achieved ROC-AUC 0.918, sensitivity 0.739, and specificity 0.952. LightGBM yielded ROC-AUC 0.916, sensitivity 0.782, and specificity 0.904. Random Forest recorded ROC-AUC 0.888, sensitivity 0.826, and specificity 0.667. Key features included "progression or recurrence" and wavelet-based texture metrics reflecting tumor heterogeneity. These findings demonstrate the effectiveness of combining CT radiomics with gradient boosting models to capture complex prognostic patterns. This integration enhances 5-year survival prediction in CRLM, offering a non-invasive tool for personalized risk stratification and improved clinical decision-making compared to the currently utilized traditional prognostic models.

Corresponding Author:
Widya Astuti
Department of Physics, Faculty of Science and Mathematics
Diponegoro University
Central Java 50275, Indonesia
Email: widyastuti1320@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v9i1.39071

1. INTRODUCTION

Colorectal Cancer (CRC) is one of the most common malignancies worldwide, and liver metastasis (Colorectal Liver Metastases/ CRLM) occurs in 25–50% of patients during the disease course, serving as the primary cause of mortality [1]. In patients with resectable CRLM, the 5-year overall survival (OS) rate ranges from 20–58%, while the overall prognosis remains poor without aggressive intervention [2]. Surgical resection or thermal ablation offers the best opportunity for long-term survival; however, high recurrence rates (60–80%) and early recurrence (ER) represent independent prognostic factors adversely affecting long-term survival [3]. Traditional clinical prognostic models, such as the Fong score, rely on demographic, pathologic, and clinical factors but often fail to capture tumor heterogeneity and microenvironment dynamics due to limitations in spatial and quantitative information [1]. Radiomics approaches, which extract quantitative features from

medical imaging (such as CT or MRI), have emerged as a non-invasive method for quantifying tumor heterogeneity, spatial patterns, and peritumoral interactions, thereby enhancing prognostic prediction accuracy when integrated with clinical data [4].

In the past 5 years, studies have demonstrated that combining intratumoral and peritumoral radiomics features with clinical variables and processing them with machine learning (ML) algorithms can improve predictions of recurrence, therapy response, and survival in CRLM patients. Tree-based models such as Random Forest (RF) and gradient boosting (including XGBoost and LightGBM) have proven effective in handling high-dimensional and imbalanced data, offering superior performance compared to traditional logistic regression [1]. For instance, an XGBoost model based on MRI radiomics and clinical scores achieved an AUC of 0.772 on external validation for predicting early recurrence after thermal ablation, with evidence that early recurrence strongly correlates with long-term prognosis [1]. Brunetti et al. (2025) developed explainable radiomics-based machine learning models, including ensembles optimized via genetic algorithms, to predict recurrence and overall survival (OS) status in CRLM. These models attained AUCs of 0.78 for recurrence and 0.68–0.71 for OS status on external validation, outperforming conventional clinical models such as the Fong score while highlighting opportunities for further improvement in discriminative capability [4]. Lin et al. (2025) integrated radiomics with machine learning classifiers, such as Random Forest, to predict intrahepatic recurrence within one year post-surgery in CRLM, achieving an AUC of 0.708 and accuracy of 75.86% when combining imaging and clinical data [5]. Yan et al. (2025) proposed a multimodal MRI-based model fusing radiomics, deep learning, and clinical features to predict CRLM occurrence, yielding AUCs of 0.889 in the training set and 0.822 in external validation [6]. A comprehensive review by Kokkinakis et al. (2024) indicated that the majority of prognostic prediction models for CRLM, including those predicting 5-year overall survival using Cox proportional hazards regression and Lasso regression, exhibit AUC values ranging from 0.60 to 0.80. This underscores the limitations of current approaches and the potential of advanced ensemble methods to improve outcomes [2].

Despite recent progress in radiomics for CRLM, a critical gap remains: most studies focus on short-term recurrence or general overall survival status, often utilizing linear methods like Lasso or Cox regression which may fail to capture complex, non-linear interactions within high-dimensional data. In contrast to previous works that utilized MRI or multi-modal fusion, which can be resource-intensive, this study focuses exclusively on preoperative CT-based radiomics, as CT remains the most accessible and standardized modality in clinical practice (Exclusion). Furthermore, while current 5-year survival models reported by Kokkinakis et al. (2024) exhibit modest AUCs (0.60–0.80), this research contributes a high-performance framework by conducting a direct head-to-head comparison of advanced ensemble algorithms (XGBoost, LightGBM, and Random Forest). The primary contribution of this work lies in the implementation of a unified, reproducible pipeline integrating multi-stage feature selection, SMOTE resampling, Optuna optimization, and isotonic probability calibration specifically tailored for binary 5-year survival classification. By addressing these methodological gaps, this study clarifies the relative strengths of gradient boosting versus bagging strategies, providing a robust prognostic tool to support long-term risk stratification in CRLM.

2. RESEARCH METHOD

2.1. Dataset

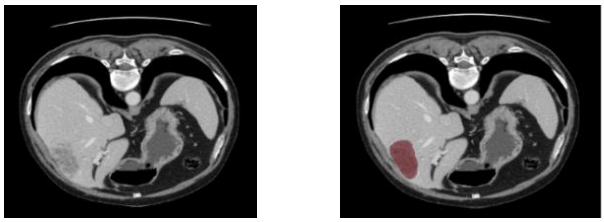
This retrospective study utilized data from The Cancer Imaging Archive (TCIA), specifically the Colorectal Liver Metastases (CRLM) collection, comprising 197 patients with a diagnosis of colorectal liver metastases [7]. Inclusion criteria included: (a) histopathologically confirmed CRLM that had undergone resection, (b) availability of pathologic analysis data from non-tumoral and tumoral liver tissue, and (c) preoperative portal venous phase contrast-enhanced multi-detector computed tomography (MDCT) scans acquired within 6 weeks prior to liver resection. Patients were excluded if they experienced death within 90 days postoperatively, had follow-up of less than 24 months, received hepatic arterial infusion (HAI) chemotherapy, underwent local tumor ablation, had more than three wedge resections, or had no visible tumor on preoperative imaging, to ensure accurate assessment of tumor burden. The dataset included semi-automatic segmentations of the liver, tumors, and vessels performed using Scout Liver by expert radiologists. For consistency in analysis and to reduce bias, only the largest lesion from each patient was analyzed.

2.2. Radiomics Pipeline

In the TCIA dataset, CT images in DICOM format and the corresponding multi-label segmentations (liver, remaining liver, portal vein, hepatic vein, and tumor) were converted to NIfTI format (.nii.gz) to ensure data uniformity for radiomics analysis. This conversion step is essential for compatibility with PyRadiomics, which requires NIfTI format for standardized feature extraction [8]. Images were loaded as three-dimensional volumes, and the segmentations were resampled to match the dimensions of the CT volumes to ensure spatial correspondence between imaging data and regions of interest.

All NIfTI images were subsequently resampled to an isotropic voxel spacing of 1 × 1 × 1 mm using B-spline interpolation to standardize spatial resolution across patients and improve the robustness of feature extraction [9]. This resampling strategy follows the Image Biomarker Standardization Initiative (IBSI) guidelines, which recommend isotropic voxel spacing to reduce inter-scanner variability and enhance feature reproducibility across different acquisition protocols [9]. B-spline interpolation was chosen over linear or nearest-neighbor methods because it provides smoother interpolation while preserving edge information, which is critical for accurate texture feature calculation [10]. The resampling grid was aligned by matching the corner of the origin voxel to maintain consistency with PyRadiomics implementation standards.

Radiomics features were extracted exclusively from the tumor-labeled regions (label value = 1 in the segmentation mask), as shown in Figure 1a–b. Only the largest tumor lesion per patient was analyzed to ensure consistency and reduce potential confounding from multifocal disease heterogeneity, consistent with prior CRLM radiomics studies [4]. The expert-annotated segmentations from TCIA were used without modification, having been previously validated by board-certified radiologists using Scout Liver software.



(a) (b)

Figure 1. (a) CT image, (b) CT image with tumor marked.

Feature extraction from the tumor volumes was performed using PyRadiomics (version 3.0.1). The extracted features included first-order statistics, shape-based features, and texture features derived from the Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighboring Gray Tone Difference Matrix (NGTDM). Additional features were obtained through logarithmic transformations (LoG) with sigma values of 1.0–5.0 mm and wavelet decompositions (e.g., HLH, HLL, LHH). Prior to texture feature extraction, image intensities were normalized using z-score scaling and discretized with a fixed bin width of 25 HU to ensure consistency and compliance with IBSI guidelines [9]. This process generated a total of 1130 radiomics features per patient, representing a comprehensive standard set from PyRadiomics to capture multifaceted tumor heterogeneity, as widely adopted in oncology radiomics studies [8], [11].

2.3. Dataset Splitting

The initial dataset comprised 197 patients with colorectal liver metastases (CRLM) who underwent liver resection, with clinical data and radiomics features extracted from preoperative CT/MRI imaging. Patients were classified into the "Survive" group if overall survival (OS) was ≥60 months with vital status alive (vital_status=0), and into the "Non-survive" group if OS was <60 months with vital status deceased (vital_status=1). Patients with censored data or those not meeting these strict criteria were excluded to avoid bias, resulting in 145 usable patients (75 Survive, 70 Non-survive). The dataset was then split into a training set (70%) and a test set (30%), stratified by 5-year survival status to maintain balanced class proportions. This was achieved through iterative random splitting until homogeneous distributions of clinical and numerical variables were obtained across sets (using Chi-square or Fisher's exact test for categorical variables and Mann-Whitney U test for numerical variables), yielding a training set of 101 patients (52 Survive, 49 Non-survive) and a testing set of 44 patients (23 Survive, 21 Non-survive). Baseline characteristics for the overall cohort, training set, and testing set were summarized in a table to verify distribution balance.

2.4. Feature Selection

Feature extraction yielded 1130 radiomics features, which, when combined with clinical variables, resulted in a total of 1143 features. To address the challenges of high dimensionality, multicollinearity, and potential overfitting inherent in radiomics datasets, a stepwise feature selection pipeline was employed [12]. This approach is widely adopted in recent radiomics studies on colorectal liver metastases (CRLM) to enhance model stability, reduce computational burden, and improve generalizability while retaining prognostic value [13].

First, redundant features with an absolute Pearson correlation coefficient >0.85 were removed, retaining the one with the highest variance to mitigate multicollinearity, a common issue in texture and wavelet features derived from filtered images [14]. Second, low-variance features (threshold <0.015) were discarded

to eliminate non-informative or near-constant features that contribute little to class discrimination. Third, if the remaining features exceeded 25, the top 25 were pre-selected based on mutual information with the target class. Mutual information was preferred because it effectively captures non-linear dependencies, offering clear advantages over linear univariate methods (e.g., ANOVA) in high-dimensional radiomics data [15]. This pre-filtering step to 20–30 features before wrapper methods is a standard practice to minimize overfitting risk and computational cost prior to more intensive selection [13].

Finally, Recursive Feature Elimination with Cross-Validation (RFECV) using a Random Forest estimator was applied to identify the optimal subset (targeting 10–20 features), with minor tuning for improved generalization on CRLM-specific cohorts [1]. RFECV was chosen for its robustness as a wrapper method that iteratively removes the least contributory features while monitoring cross-validated performance [13]. This stepwise pipeline ultimately yielded 16 robust, reproducible, and clinically interpretable features (radiomic + clinical), in line with current best-practice guidelines for radiomics-based prognostic modeling in liver malignancies [12], [13].

2.5. Machine Learning Models

Three tree-based ensemble models, Random Forest (RF), XGBoost, and LightGBM, were employed to predict 5-year survival in patients with colorectal liver metastases (CRLM). These algorithms were selected due to their demonstrated superiority in handling high-dimensional, noisy, and potentially imbalanced radiomics data, frequently outperforming conventional logistic regression in medical imaging tasks [16].

Hyperparameter optimization was performed using Optuna (100 trials) with Tree-structured Parzen Estimator (TPESampler) and MedianPruner to maximize stratified cross-validated ROC-AUC. Optuna's Bayesian optimization strategy was chosen over grid or random search because it converges faster and has become the preferred framework in recent radiomics and survival-prediction studies [16]. To address mild class imbalance in the training set, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic minority-class samples, thereby improving model sensitivity without introducing substantial bias [17]. Finally, probability calibration was performed using CalibratedClassifierCV with isotonic regression and 3-fold cross-validation. The use of 3-fold stratified cross-validation during isotonic calibration was implemented to obtain more robust and Isotonic calibration was preferred over parametric (sigmoid/Platt) scaling for its non-parametric flexibility, better handling of non-monotonic distortions, and superior reliability on heterogeneous, small-sample radiomics datasets typical of CRLM cohorts [18]. Optimized hyperparameters for each model are provided in Table 1. Model performance was comprehensively evaluated using ROC-AUC, PRC-AUC, accuracy, F1-score, precision, recall, and specificity on both training and independent test sets.

Table 1. Hyperparameter: Search Range dan Best Value

Hyperparameter	XGBoost Search Range	XGBoost Best Value	LightGBM Search Range	LightGBM Best Value	Random Forest Search Range	Random Forest Best Value
n_estimators	100 – 300	300	100 – 300	123	100 – 300	124
learning_rate	0.01 – 0.2 (log)	0.0112	0.01 – 0.2 (log)	0.0238	–	–
max_depth	3 – 6	5	3 – 7	5	3 – 15	9
min_child_weight	1 – 10	2	–	–	–	–
min_child_samples	–	–	10 – 30	11	–	–
min_samples_split	–	–	–	–	10 – 30	18
min_samples_leaf	–	–	–	–	3 – 15	3
subsample	0.7 – 0.9	0.7681	0.5 – 0.9	0.6673	–	–
colsample_bytree	0.7 – 0.9	0.8814	0.5 – 0.9	0.8380	–	–
gamma	0 – 2	1.8637	–	–	–	–
reg_alpha	0 – 1	0.9145	0 – 1	0.5491	–	–
reg_lambda	0.5 – 2	1.3354	0 – 1	0.6043	–	–
num_leaves	–	–	15 – 30	24	–	–
min_split_gain	–	–	0 – 0.1	0.0194	–	–
max_features	–	–	–	–	'sqrt', 'log2', 0.5	0.5
max_samples	–	–	–	–	0.5 – 0.9	0.8798
min_impurity_decrease	–	–	–	–	0.0 – 0.01	0.0073

3. RESULTS AND DISCUSSION

3.1. Patient Characteristics

A total of 145 patients with colorectal liver metastases (CRLM) who underwent curative liver resection were included in this study, with the data randomly divided into a training set (n=101, 69.7%) and a testing set (n=44, 30.3%). Patient demographics and clinical characteristics are summarized in Table 2. More than half of the patients were aged over 60 years (53.79%), male (57.24%), had major comorbidities (60.69%), and presented with synchronous CRLM (57.93%). Multiple metastases and bilobar disease were observed in

56.55% and 44.14% of cases, respectively, while preoperative chemotherapy and preoperative portal vein embolization were administered to 63.45% and 11.72% of patients. The 5-year survival outcomes were balanced (51.72% survived vs. 48.28% did not survive). No statistically significant differences ($p > 0.05$) were observed in demographic or clinical variables between the training and testing sets, and continuous variables also showed no significant differences. This balanced distribution indicates a representative, unbiased training-test split, supporting the reliability of subsequent model performance evaluations.

Table 2. Patient Characteristics

Characteristic	All (n=145)	Train (n=101)	Test (n=44)	p-value
Age				1
≤60 years	67 (46.21%)	47 (46.53%)	20 (45.45%)	
>60 years	78 (53.79%)	54 (53.47%)	24 (54.55%)	
Sex				0.40
Female (0)	62 (42.76%)	46 (45.54%)	16 (36.36%)	
Male (1)	83 (57.24%)	55 (54.46%)	28 (63.64%)	
Major comorbidity				0.66
No (0)	57 (39.31%)	38 (37.62%)	19 (43.18%)	
Yes (1)	88 (60.69%)	63 (62.38%)	25 (56.82%)	
Node-positive primary				0.09
No (0)	89 (61.38%)	67 (66.34%)	22 (50.00%)	
Yes (1)	56 (38.62%)	34 (33.66%)	22 (50.00%)	
Synchronous CRLM				0.72
No (0)	61 (42.07%)	41 (40.59%)	20 (45.45%)	
Yes (1)	84 (57.93%)	60 (59.41%)	24 (54.55%)	
Multiple metastases				0.82
No (0)	63 (43.45%)	45 (44.55%)	18 (40.91%)	
Yes (1)	82 (56.55%)	56 (55.45%)	26 (59.09%)	
Bilobar disease				0.45
No (0)	81 (55.86%)	59 (58.42%)	22 (50.00%)	
Yes (1)	64 (44.14%)	42 (41.58%)	22 (50.00%)	
Chemotherapy before liver resection				0.83
No (0)	53 (36.55%)	38 (37.62%)	15 (34.09%)	
Yes (1)	92 (63.45%)	63 (62.38%)	29 (65.91%)	
Preoperative PVE				0.06
No (0)	128 (88.28%)	93 (92.08%)	35 (79.55%)	
Yes (1)	17 (11.72%)	8 (7.92%)	9 (20.45%)	
Presence of sinusoidal dilatation				1
No (0)	127 (87.59%)	88 (87.13%)	39 (88.64%)	
Yes (1)	18 (12.41%)	13 (12.87%)	5 (11.36%)	
Progression or recurrence				1
No (0)	56 (38.62%)	39 (38.61%)	17 (38.64%)	
Yes (1)	89 (61.38%)	62 (61.39%)	27 (61.36%)	
5-year survival				1
Non-survive	70 (48.28%)	49 (48.51%)	21 (47.73%)	
Survive	75 (51.72%)	52 (51.49%)	23 (52.27%)	
BMI (kg/m²), median (IQR)	26.50 (23.50–30.10)	26.30 (23.50–30.10)	26.80 (22.85–29.88)	0.87†
Overall survival (months), median (IQR)	61.10 (32.90–102.43)	60.80 (34.40–104.43)	62.73 (28.23–96.32)	0.33†
Maximum tumor size (cm), median (IQR)	2.90 (2.00–4.50)	3.00 (2.00–5.00)	2.60 (1.45–4.00)	0.17†

Note: PVE: Portal Vein Embolization, BMI: Body Mass Index, †: Mann-Whitney U test is used for continuous variables, Chi-square test is used for categorical variables.

3.2. Feature Importance and Clinical Interpretation

3.2.1. Feature Importance Analysis

Feature importance analysis across the three machine learning models revealed that the "progression or recurrence" feature was the most dominant, achieving the highest importance scores of 0.408 in Random Forest, 0.174 in XGBoost, and 0.159 in LightGBM. Wavelet- and log-sigma-based radiomics features demonstrated significant contributions to model predictions, particularly "Tumor log-sigma-3-0-mm-3D glszm SmallAreaHighGrayLevelEmphasis," which consistently ranked as an important feature in all three models with importance values of 0.169 (Random Forest), 0.079 (XGBoost), and 0.157 (LightGBM). Other wavelet-based texture features, such as "Tumor wavelet-LHL glszm SmallAreaLowGrayLevelEmphasis" and "Tumor wavelet-LHH glcm SumEntropy," also provided substantial contributions to classification, reflecting the ability of radiomics to capture tumor heterogeneity not visible on visual inspection [19]. These findings align with recent studies demonstrating that texture features and tumor heterogeneity are strong predictors of prognosis and recurrence in cancer [20].

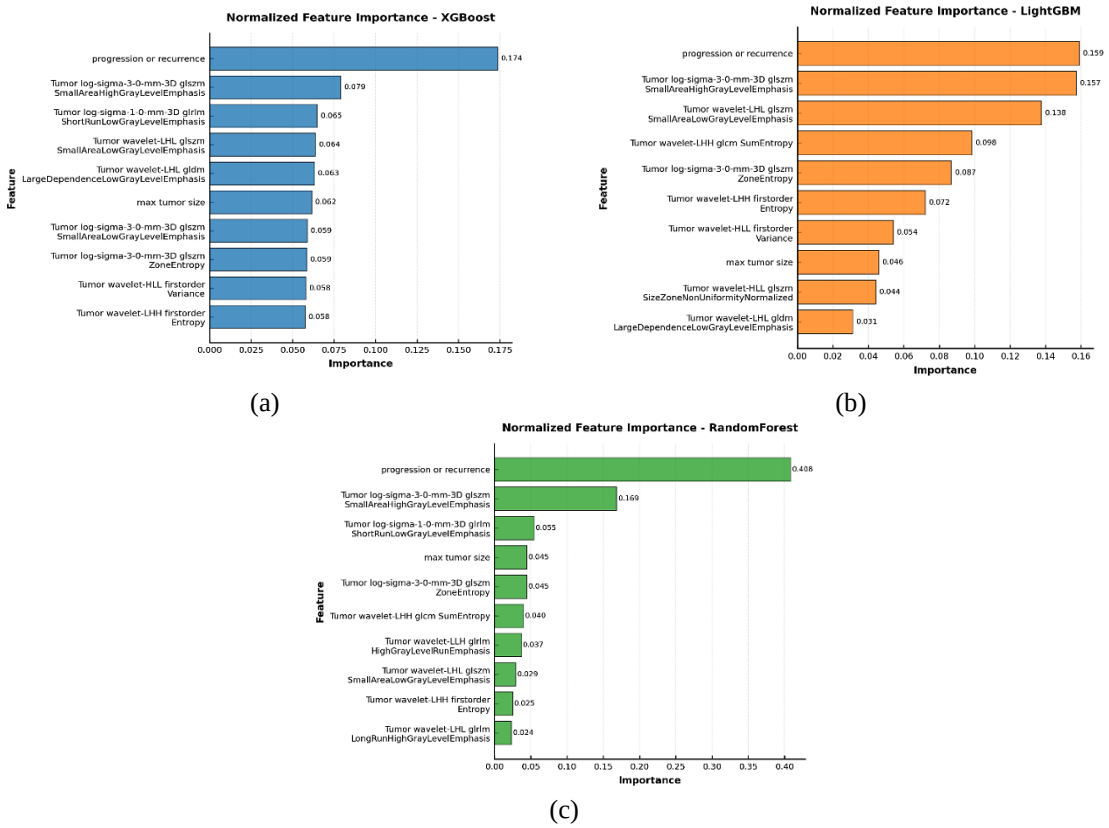


Figure 2. Normalized top 10 feature importance values in models (a) XGBoost, (b) LightGBM, and (c) Random Forest

Differences in the distribution of feature importance among the models reflect the distinct algorithmic characteristics of each machine learning method. Random Forest exhibited a markedly high concentration of importance on the "progression or recurrence" feature (40.8%), indicating a more selective approach heavily reliant on a single dominant feature [21]. In contrast, XGBoost and LightGBM displayed more evenly distributed importance across various radiomics features, with maximum importance values of 17.4% for XGBoost and 15.9% for LightGBM on the same top feature. The more balanced distribution in gradient boosting methods (XGBoost and LightGBM) suggests superior capability in capturing complex patterns through combinations of multiple radiomics features, consistent with LightGBM's leaf-wise growth mechanism and XGBoost's level-wise growth [22]. Recent studies indicate that LightGBM tends to produce models with better generalization due to reduced reliance on individual features, whereas Random Forest shows a propensity for overfitting on high-variance features [23]. This pattern explains why gradient boosting-based ensemble models such as XGBoost and LightGBM frequently outperform Random Forest in radiomics studies for predicting clinical outcomes in cancer [24].

3.2.2. Clinical Interpretation

SHapley Additive exPlanations (SHAP) analysis was performed to elucidate the clinical relevance of the selected features in the biological context of colorectal liver metastasis. All three models, LightGBM, XGBoost, and Random Forest (Figure 3a–c), consistently ranked progression_or_recurrence as the top contributor and Tumor_log-sigma-3-0-mm-3D_glszm_SmallAreaHighGrayLevelEmphasis as the second most influential radiomic feature, whereby low feature values (blue) positively influenced the prediction output while high values (red) contributed negatively, with the widest SHAP range observed in LightGBM (−1.0 to >+2.0), followed by XGBoost and Random Forest despite differences in scale. Divergence emerged at the third rank: LightGBM prioritized Tumor_wavelet-LHL_glszm_SmallAreaLowGrayLevelEmphasis, capturing small low-intensity regions in the wavelet domain; XGBoost ranked Tumor_log-sigma-3-0-mm-3D_glszm_ZoneEntropy, reflecting the disorder of gray-level zone distribution as an indicator of intratumoral heterogeneity; whereas Random Forest placed max_tumor_size at the third position, indicating that this ensemble-based model still assigns considerable weight to conventional morphological clinical parameters consistent with the findings of Wang et al., who reported that GLSZM features and tumor size jointly contribute to the prediction of metachronous liver metastasis in colorectal cancer [25]. Entropy-based features, namely

SumEntropy, ZoneEntropy, and firstorder_Entropy appeared consistently across all three models albeit in slightly different orders, in accordance with Zhang et al., who demonstrated that ZoneEntropy in the wavelet domain reflects greater intratumoral heterogeneity and is associated with higher recurrence risk in colorectal liver metastasis [26]; meanwhile, ShortRunLowGrayLevelEmphasis appeared in XGBoost and Random Forest but was absent in LightGBM, and LongRunHighGrayLevelEmphasis was exclusively identified in Random Forest, suggesting that each algorithm captures distinct GLRLM texture dimensions according to its internal learning mechanism. Overall, the consistency of the two top-ranked features and the dominance of GLSZM-entropy features across all three algorithms reinforce the validity of these radiomic descriptors as robust imaging biomarkers, as further corroborated by Granata et al., who applied SHAP to CT-based radiomics machine learning models and confirmed that wavelet-GLSZM and first-order features dominated the prediction of recurrence and overall survival in colorectal liver metastasis patients [4].

3.3. Model Performance Evaluation

3.3.1. Discriminative Performance

Evaluation of model performance on the testing data demonstrated the high effectiveness of integrating CT-based radiomics with ensemble machine learning for 5-year survival prediction in CRLM patients. All three models achieved excellent classification performance, as illustrated in Figure 4 (ROC curves) and detailed in Table 3. XGBoost exhibited the best overall performance with a Test ROC-AUC of 0.9182, PRC-AUC of 0.9332, accuracy of 84 %, and F1-score of 82.9 %. LightGBM followed closely (ROC-AUC 0.9161, PRC-AUC 0.9370, accuracy 84.09 %, highest F1-score 83.7 %), while Random Forest recorded a slightly lower ROC-AUC of 0.8882 and accuracy of 75 %.

The relatively small performance gap between XGBoost and LightGBM highlights the strength of gradient-boosting ensemble models when combined with high-dimensional radiomics features: both algorithms effectively captured tumor heterogeneity through wavelet and GLSZM texture metrics, delivering superior discrimination (AUC > 0.91) compared to conventional clinical scores and most previous CRLM models (0.60–0.80) [2], [27]. XGBoost offered the highest specificity (95.2 %), making it particularly advantageous for conservative clinical decision-making, whereas LightGBM provided better balance in handling class imbalance (sensitivity 78.2 %, specificity 90.4 %). In contrast, Random Forest's lower specificity (66.7 %) despite high recall reveals a limitation of bagging-based ensembles in this radiomics setting, where high feature-to-sample ratio can amplify overfitting risk on subtle textural patterns. Overall, these results confirm that CT radiomics paired with optimized ensemble learning provides a clear advantage over traditional prognostic tools, achieving "excellent discrimination" (AUC > 0.88) while remaining fully non-invasive and preoperative [27].

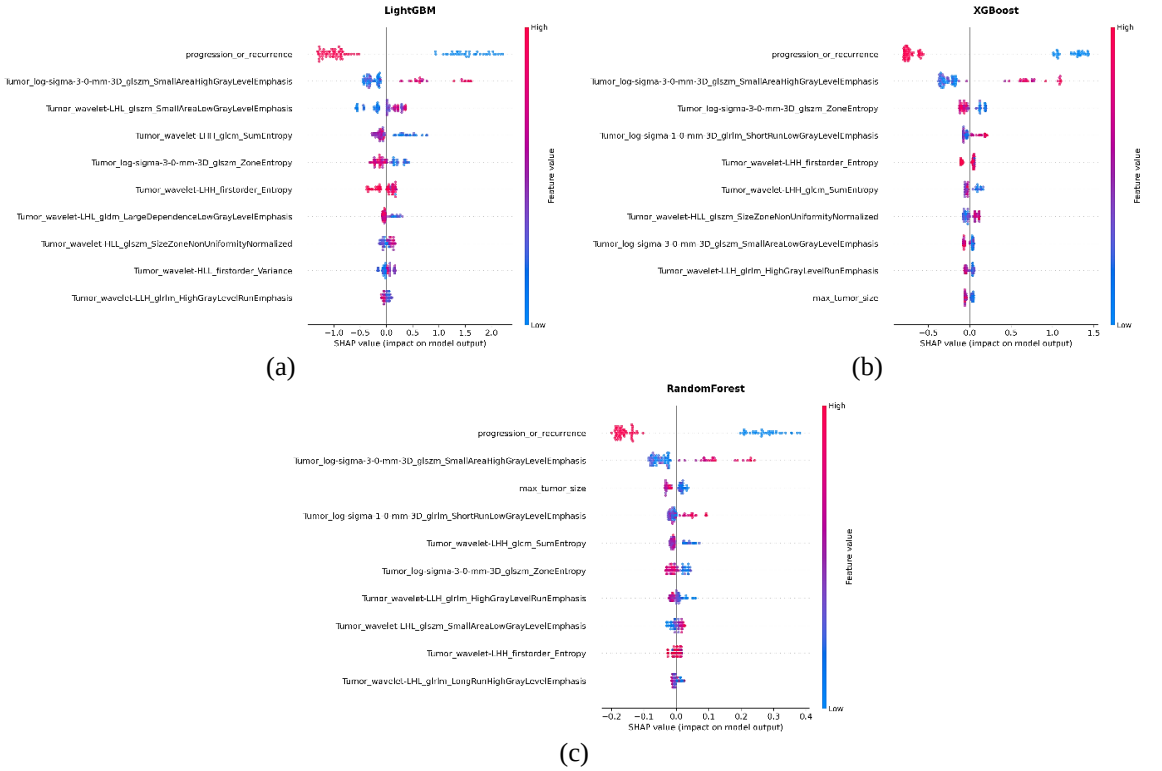


Figure 3. SHAP summary beeswarm plots: (a) LightGBM, (b) XGBoost, (c) Random Forest.

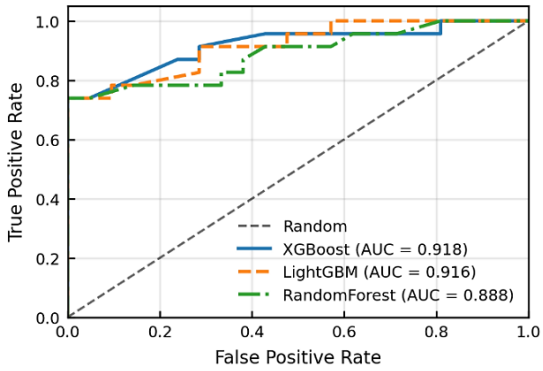


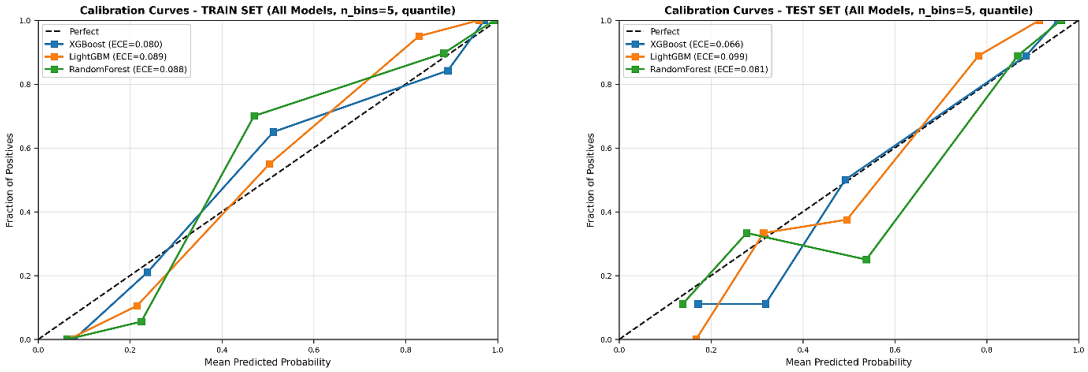
Figure 4. Comparison of ROC Curves for Machine Learning Models

Table 3. Performance Comparison of Machine Learning Models

Model		XGBoost	LightGBM	Random Forest
ROC AUC	Train	0.948	0.967	0.963
	Test	0.918	0.916	0.888
PRC AUC	Train	0.942	0.970	0.964
	Test	0.933	0.937	0.922
Accuracy	Train	0.841	0.881	0.861
	Test	0.840	0.840	0.750
F1	Train	0.836	0.880	0.862
	Test	0.829	0.837	0.775
Precision	Train	0.891	0.916	0.880
	Test	0.944	0.900	0.730
Sensitivity	Train	0.788	0.846	0.846
	Test	0.739	0.782	0.826
Specificity	Train	0.898	0.918	0.877
	Test	0.952	0.904	0.667

3.3.2. Probabilistic Calibration

Model calibration measures the degree of agreement between a model's predicted probabilities and the actual observed outcomes. Two complementary metrics were used: the Expected Calibration Error (ECE), which measures the mean absolute gap between predicted probabilities and observed event frequencies across probability bins (ideal = 0), and the Brier Score, which captures overall probabilistic accuracy as the mean squared difference between predicted probabilities and binary outcomes (ideal = 0, with values below 0.25 generally considered acceptable for imbalanced clinical datasets) [28]. On the training set, XGBoost achieved the lowest ECE (0.080) with a Brier Score of 0.097, followed by Random Forest (ECE = 0.088, Brier = 0.084) and LightGBM (ECE = 0.089, Brier = 0.080), indicating comparable moderate calibration across all three models (Figure 5a). The calibration curves of all three models exhibited a characteristic non-monotonic dip in the low-to-mid probability range (0.2–0.35), where the fraction of positives dropped sharply below the diagonal before recovering a pattern attributable to sparse bin occupancy under quantile-based binning on small datasets, where individual bins contain very few samples, causing local fraction-of-positives estimates to fluctuate substantially rather than reflecting true systematic miscalibration [29].



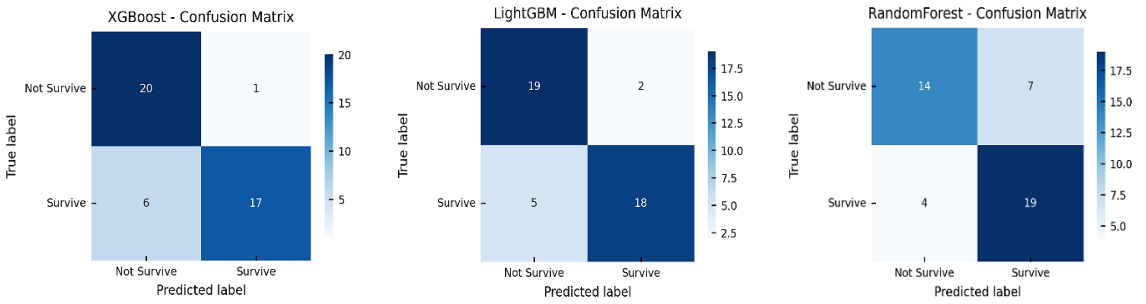
(a) (b)

Figure 5. Calibration curves for XGBoost, LightGBM, and Random Forest models on the (a) training set and (b) test set, generated using quantile-based binning (n_bins = 5).

On the test set, XGBoost demonstrated the best calibration with the lowest ECE (0.066) and the smallest train-to-test Brier Score increase (0.097), confirming its superior probability stability across data splits (Figure 5b). Random Forest achieved a moderate (ECE = 0.081, Brier = 0.126), while LightGBM showed the highest miscalibration (ECE = 0.099, Brier = 0.125) with the largest Brier Score deterioration. The non-monotonic dip pattern persisted on the test set and was most prominent in XGBoost, whose curve remained flat at a fraction of positives of approximately 0.11 across the 0.15–0.32 probability range before rising sharply and in LightGBM, which dropped to near zero at the 0.20 bin before recovering; these irregularities are consistent with the known instability of isotonic calibration on small test partitions, where sparse bin occupancy amplifies local estimation variance [29]. Isotonic calibration was selected over Platt scaling due to its non-parametric flexibility in correcting any monotonic distortion without assuming a sigmoidal mapping, though its susceptibility to overfitting on small cohorts means that the observed ECE range of 0.066–0.099 while acceptable should be interpreted cautiously pending external validation [17].

3.3.4. Confusion Matrix Analysis

Confusion matrix analysis on the testing dataset (n=44) demonstrated the practical effectiveness of the radiomics-ensemble approach in classifying 5-year survival, as shown in Figure 6. XGBoost achieved the highest true negative rate (20 TN, only 1 FP), correctly identifying nearly all non-survivors with excellent specificity (95.2 %), making it the most conservative and clinically safe model for ruling out low-risk patients (Figure 6a). LightGBM provided the best balance (19 TN, 18 TP; Figure 6b), while Random Forest showed the highest sensitivity (82.6 %) but sacrificed specificity (66.7 %), resulting in 7 false positives that could lead to unnecessary aggressive adjuvant therapy (Figure 6c).



(a) (b) (c)

Figure 6. Confusion matrix for models (a) XGBoost, (b) LightGBM, and (c) Random Forest

These differences highlight the advantages and limitations of the proposed method. The superior performance of gradient boosting models (XGBoost and LightGBM) in reducing false positives stems from their sequential error-correction mechanism, which effectively handles the heterogeneous nature of radiomics features and class imbalance after SMOTE demonstrating clear methodological strength for preoperative risk stratification in CRLM [30]. However, the 5–6 false negatives across models and Random Forest's low specificity reveal the remaining limitation of the current pipeline: with only 145 patients and 1130 initial features, even rigorous selection cannot fully eliminate overfitting risk on subtle radiomic patterns in small lesions. This pattern mirrors previous radiomics studies on tumors and underscores the need for external validation to ensure reliable clinical translation [31].

3.3.5. Methodological Strengths and Limitations

The integration of CT-based radiomics with ensemble machine learning models demonstrated clear methodological strengths for 5-year survival prediction in CRLM. By extracting 1130 quantitative features from preoperative portal venous phase CT and applying stepwise selection (correlation filter, low-variance filter, mutual information, and RFECV) to retain only 16 robust predictors, the pipeline effectively captured intratumoral heterogeneity that is invisible on conventional imaging particularly wavelet-GLSZM texture metrics and entropy features that consistently ranked highest in both feature importance and SHAP analyses across all three models [19], [26]. XGBoost and LightGBM achieved excellent discriminative performance (test ROC-AUC 0.918 and 0.916, PRC-AUC up to 0.937) with balanced or high specificity (up to 95.2 %), significantly outperforming Random Forest and previous CRLM prognostic models that typically report AUC values of 0.60–0.80 [27]. Optuna hyperparameter optimization, SMOTE for imbalance handling, and isotonic calibration further enhanced reliability (lowest test ECE 0.066 for XGBoost), delivering a non-invasive,

preoperative, and clinically interpretable tool that provides actionable risk stratification superior to traditional scores [28], [29].

Nevertheless, several limitations must be acknowledged. The relatively small cohort (n=145, test set n=44) from a single public TCIA repository increases the risk of overfitting, as evidenced by non-monotonic calibration curves caused by sparse probability bins and Random Forest's markedly lower specificity (66.7 %) in this high-dimensional setting [21], [31]. The absence of external validation, peritumoral radiomics features, or multimodal imaging (MRI/PET-CT) limits generalizability, reflecting well-documented challenges in radiomics reproducibility for CRLM [4], [11]. Although rigorous feature selection mitigated multicollinearity, the initial 1130-feature space still carries the inherent "curse of dimensionality" typical of radiomics studies with modest sample sizes. These constraints highlight the necessity for future multicenter prospective validation and imaging harmonization to strengthen clinical translation.

4. CONCLUSION

This study demonstrated the high effectiveness of integrating CT-based radiomics with ensemble machine learning for predicting 5-year overall survival in colorectal liver metastases (CRLM). XGBoost and LightGBM achieved excellent discriminative performance (test ROC-AUC 0.918 and 0.916, respectively) with strong specificity and balanced sensitivity, clearly outperforming Random Forest (AUC 0.888) and most existing prognostic models in CRLM (AUC 0.60–0.80). The integration of radiomics features, particularly progression or recurrence status and wavelet-based texture metrics that capture intratumoral heterogeneity, was further validated through SHAP analysis, confirming their robustness as imaging biomarkers and providing a non-invasive preoperative tool that enhances risk stratification and supports personalized treatment strategies beyond conventional clinical factors. Despite promising results, limitations include a relatively small single-institution sample (n=145) and a lack of external validation. Future studies should prioritize multicenter external validation, radiomics harmonization across scanners, expansion of SHAP-based explainability in larger cohorts, and prospective clinical evaluation to strengthen generalizability and facilitate real-world implementation.

DATA AVAILABILITY

The data are openly available in a public repository that publishes datasets with a DOI. The data supporting the findings of this study are openly available in [TCIA] at https://doi.org/10.7937/QXK2-QG03, under the [CC BY 4.0] license.

REFERENCES

[1] Y. Kong, L. Wan, X. Yue, F. Tang, and X. Zhou, "An interpretable machine learning model based on MRI radiomics and GAME score for predicting early recurrence after thermal ablation in colorectal liver metastases," *Int. J. Colorectal Dis.*, vol. 41, no. 1, p. 29, 2026, doi: 10.1007/s00384-025-05079-2.

[2] S. Kokkinakis, I. A. Ziogas, J. D. Llaque Salazar, D. P. Moris, and G. Tsoulfas, "Clinical Prediction Models for Prognosis of Colorectal Liver Metastases: A Comprehensive Review of Regression-Based and Machine Learning Models," *Cancers (Basel).*, vol. 16, no. 9, 2024, doi: 10.3390/cancers16091645.

[3] Y. Kong, X. Huang, X. Cao, F. Tang, and X. Zhou, "Early Recurrence of Colorectal Liver Metastasis (Number ≤ 5 and Largest Diameter ≤ 3 cm) after Resection or Thermal Ablation: a Multi-center Study of Patterns, Safety, Survival and Risk Factors," *J. Gastrointest. Cancer*, vol. 56, no. 1, p. 77, 2025, doi: 10.1007/s12029-025-01200-4.

[4] A. Brunetti, G. M. Zaccaria, E. Sibilano, S. Marzi, A. Vidiri, and V. Bevilacqua, "Development and independent validation of explainable radiomics-based machine learning models for prognosis in colorectal liver metastases," *Front. Digit. Health*, vol. Volume 7-2025, 2026, doi: 10.3389/fdgth.2025.1752699.

[5] Y. Lin, Y. Huang, Z. Liu, X. Feng, and C. Yang, "Predicting early recurrence of colorectal cancer liver metastases: an integrative approach using radiomics and machine learning," *Front. Oncol.*, vol. 15, p. 1613093, 2025, doi: 10.3389/fonc.2025.1613093.

[6] X. Yan *et al.*, "A Multimodal MRI-based model for colorectal liver metastasis prediction: integrating radiomics, deep learning, and clinical features with SHAP interpretation," *Current Oncology*, vol. 32, no. 8, p. 431, 2025, doi: 10.3390/curroncol32080431.

[7] A. L. Simpson *et al.*, "Preoperative CT and survival data for patients undergoing resection of colorectal liver metastases," *Sci. Data*, vol. 11, no. 1, p. 172, 2024, doi: 10.1038/s41597-024-02981-2.

[8] J. J. M. Van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[9] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, Mar. 2020, doi: 10.1148/radiol.2020191145.

[10] D. Mackin *et al.*, "Measuring computed tomography scanner variability of radiomics features," *Invest. Radiol.*, vol. 50, no. 11, pp. 757–765, 2015, doi: 10.1097/RLI.0000000000000180.

[11] P. Lambin *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017, doi: 10.1038/nrclinonc.2017.141.

[12] H. Hu, J. C. Chi, B. Zhai, and J. H. Guo, "CT-based radiomics analysis to predict local progression of recurrent colorectal liver metastases after microwave ablation," *Medicine*, vol. 102, no. 52, 2023, doi: 10.1097/MD.0000000000036586.
[13] Y. Yan *et al.*, "Multiphase MRI-Based Radiomics for Predicting Histological Grade of Hepatocellular Carcinoma," *Journal of Magnetic Resonance Imaging*, vol. 60, no. 5, pp. 2117–2127, Nov. 2024, doi: https://doi.org/10.1002/jmri.29289.
[14] C. Marzi *et al.*, "Collinearity and Dimensionality Reduction in Radiomics: Effect of Preprocessing Parameters in Hypertrophic Cardiomyopathy Magnetic Resonance T1 and T2 Mapping," *Bioengineering*, vol. 10, no. 1, 2023, doi: 10.3390/bioengineering10010080.
[15] O. O. Oladimeji, H. Ayaz, I. McLoughlin, and S. Unnikrishnan, "Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis," *Results in Engineering*, vol. 24, p. 103071, 2024, doi: https://doi.org/10.1016/j.rineng.2024.103071.
[16] M. Karabacak, S. Patil, R. Feng, R. K. Shrivastava, and K. Margetis, "A large scale multi institutional study for radiomics driven machine learning for meningioma grading," *Sci. Rep.*, vol. 14, no. 1, p. 26191, 2024, doi: 10.1038/s41598-024-78311-8.
[17] A. Demircioğlu, "The effect of data resampling methods in radiomics," *Sci. Rep.*, vol. 14, no. 1, p. 2858, 2024, doi: 10.1038/s41598-024-53491-5.
[18] G. Mehri-kakavand, S. Mdletshe, M. Amini, and A. Wang, "Multimodal radiomics fusion for predicting postoperative recurrence in NSCLC patients," *J. Cancer Res. Clin. Oncol.*, vol. 151, no. 10, p. 261, 2025, doi: 10.1007/s00432-025-06311-w.
[19] L. Zedda, A. Loddo, and C. Di Ruberto, "Advancements in radiomics: A comprehensive survey of feature types and their correlation on modalities and regions," *Neurocomputing*, p. 131192, 2025, doi: 10.1016/j.neucom.2025.131192.
[20] P. Lin, Y. Lin, R. Gao, W. Wan, Y. He, and H. Yang, "Integrative radiomics and transcriptomics analyses reveal subtype characterization of non-small cell lung cancer," *Eur. Radiol.*, vol. 33, no. 9, pp. 6414–6425, 2023, doi: 10.1007/s00330-023-09503-5.
[21] X. Li, C. Li, H. Wang, L. Jiang, and M. Chen, "Comparison of radiomics-based machine-learning classifiers for the pretreatment prediction of pathologic complete response to neoadjuvant therapy in breast cancer," *PeerJ*, vol. 12, p. e17683, 2024, doi:10.7717/peerj.17683.
[22] Y. M. Indah, R. Aristawidya, A. Fitrianto, E. Erfiani, and L. M. R. D. Jumansyah, "Comparison of Random Forest, XGBoost, and LightGBM Methods for the Human Development Index Classification," *Jambura Journal of Mathematics*, vol. 7, no. 1, pp. 14–18, 2025, doi: 10.37905/jjom.v7i1.28290.
[23] H. Moradmand *et al.*, "Graph feature selection for enhancing radiomic stability and reproducibility across multiple institutions in head and neck cancer," *Sci. Rep.*, vol. 15, no. 1, p. 27995, 2025, doi: 10.1038/s41598-025-12161-w.
[24] J. Camps, A. Jiménez-Franco, R. García-Pablo, J. Joven, and M. Arenas, "Artificial intelligence-driven integration of multi-omics and radiomics: A new hope for precision cancer diagnosis and prognosis," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1871, no. 6, p. 167841, 2025, doi: 10.1016/j.bbadis.2025.167841.
[25] J.-P. Wang *et al.*, "Machine learning-based radiomics models for the prediction of metachronous liver metastases in patients with colorectal cancer: A multimodal study," *Oncol. Lett.*, vol. 30, no. 2, p. 394, 2025, doi: 10.3892/ol.2025.15140.
[26] D. Zhang, P. Li, Y. Wei, M. Xue, F. Guo, and C. Li, "Predicting the recurrence risk of liver metastasis from colorectal cancer: a study based on preoperative CT intratumoral and peritumoral radiomics features," *Front. Oncol.*, vol. Volume 15-2025, 2025, doi: 10.3389/fonc.2025.1662354.
[27] A. Z. Paredes *et al.*, "A novel machine-learning approach to predict recurrence after resection of colorectal liver metastases," *Ann. Surg. Oncol.*, vol. 27, no. 13, pp. 5139–5147, 2020, doi: 10.1245/s10434-020-08991-9.
[28] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, "A tutorial on calibration measurements and calibration models for clinical prediction models," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 621–633, Apr. 2020, doi: 10.1093/jamia/ocz228.
[29] F. M. Ojeda *et al.*, "Calibrating machine learning approaches for probability estimation: A comprehensive comparison," *Stat. Med.*, vol. 42, no. 29, pp. 5451–5478, Dec. 2023, doi: 10.1002/sim.9921.
[30] S. Buzdugan, M. Mazher, and D. Puig, "Radiogenomics for Glioblastoma Survival Prediction: Integrating Radiomics, Clinical, and Genomic Features Using Artificial Intelligence," *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01692-3.
[31] X. Li *et al.*, "Machine learning for grading prediction and survival analysis in high grade glioma," *Sci. Rep.*, vol. 15, no. 1, p. 16955, 2025, doi: 10.1038/s41598-025-01413-4.

BIBLIOGRAPHY OF AUTHORS

Widya Astuti is a Master's student in Medical Physics at the Department of Physics, Universitas Diponegoro, Indonesia. Her research focuses on medical imaging and computational oncology, particularly using machine learning to improve prognostic accuracy. In this study, she was the primary investigator, responsible for conceptualization, software implementation, data analysis, and drafting the manuscript. Contact: widyastuti1320@gmail.com

Catur Edi Widodo is a Professor at the Department of Physics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia. His expertise lies in Instrumentation Physics and Electronics, with research interests in biomedical instrumentation, intelligent diagnostic systems, and computational signal processing. He served as principal supervisor, providing strategic guidance and methodological validation. Contact: caturediwidodo@lecturer.undip.ac.id

Qidir Maulana Binu Soesanto is a lecturer and researcher at the Department of Physics, Universitas Diponegoro, Indonesia, holding a Ph.D. from Kanazawa University, Japan. His expertise is in Theoretical Physics and mathematical modeling, with a strong publication record in international journals. In this study, he acted as co-supervisor, contributing theoretical insights and ensuring mathematical rigor. Contact: qidirbinu@fisika.fsm.undip.ac.id