Metaheuristic-Driven Machine Learning Pipelines for Radiomics-Based Prediction of Locoregional Recurrence in Head and Neck Cancer

Hasan Shaikh[1], Balu Krishna S[1], Amal Joseph Varghese[1], Rajendra Benny Kuchipudi[1], Sathya A[1], Praveenraj C[1], Ezhil Sindhanai[1], Jino Wilson Victor[1], Simon Pavamani[1], Julia Priyadarshini Rao[1], Vijayshree C[1], Nikhil Raynor Cecil[1], Pon Preetham Emmanuel[1], Rajesh I[1], Manu Mathew[1], and Hannah Mary Thomas T[1]

[1]Quantitative Imaging Research and Artificial Intelligence Lab, Department of Radiation Oncology, Christian Medical College, Vellore, Tamil Nadu, India, 632004

Abstract

Purpose: Cancer recurrence occurs in 15–50% of head and neck cancer (HNC) patients despite advanced treatment strategies suggesting the need for more accurate risk predictors. Existing clinical parameters, radiological staging and tumour stage-based prognostication provide limited predictive accuracy for individual patients. Cancer Radiomics, which refers to the process of converting radiological images into quantifiable descriptions of the tumour has been actively studied to evaluate for imaging biomarkers. However, this analysis faces the classic $p \gg n$ problem, where hundreds of imaging biomarkers must be analyzed in small patient cohorts, leading to feature instability and selection bias. This study systematically compares feature selection methods across multiple classifier families for HNC locoregional recurrence prediction. The aim is to identify stable, generalizable signatures by integrating radiomics and clinical features.

Methods: We prospectively recruited HNC patients treated with chemoradiation in our institution between 2020–2024. 103 Radiomics features were extracted from this tumour volume using Pyradiomics, including first-order intensity, 3D shape and texture families. Eight clinical variables were also collected: age, location of tumour within the head and neck region, TNM/AJCC staging, and HPV status. We compared models using radiomic features alone versus radiomics features combined with clinical features. Five machine learning classifiers (Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest) were tested with seven feature selection methods including LASSO, SelectKBest and metaheuristic algorithms (Particle Swarm Optimization, Whale Optimization, Grey Wolf Optimizer, Genetic Algorithm, Simulated Annealing). Models were trained and evaluated under stratified 5-fold stratified cross-validation and independent test splits, with area under the receiver operating curve (AUC) as the primary performance evaluation metric.

Results: We recruited 367 HNC patients of which 163 HNC patients (55 with locoregional recurrence, 108 disease-free) were included for this study. The optimal model was Logistic Regression with Grey Wolf Optimizer selection using combined features, achieving test AUC 0.81 [0.62, 0.95]. The 10-feature signature included 4 clinical variables (age, AJCC stage, T-stage, location) and 6 radiomics features capturing tumor shape and texture heterogeneity. Clinical variable integration dramatically improved SVM performance from AUC ∼0.35 to ∼0.78.

Conclusion: Simple linear models with carefully selected radiomics + clinical features outperform complex algorithms in high-dimensional, small-sample scenarios. External validation across multiple institutions is the critical next step for clinical translation.

Keywords: Head-and-neck cancer; Loco-regional recurrence; Machine learning; Prognosis; Radiomics.

1 Introduction

Head and neck cancer (HNC) is a major global health problem, with approximately 890,000 new cases and 450,000 deaths reported annually worldwide, making it the seventh most common cancer globally [1, 2]. Despite improvements in radiation delivery and multimodality treatment, locoregional recurrence rates remain high, with 50–60% of patients with advanced disease experiencing recurrence within two years of treatment, and an overall locoregional recurrence rate of approximately 14% [3, 4]. These recurrences severely affect both patient survival and quality of life, with median survival of only 10–15 months for recurrent/metastatic disease and major difficulties in essential functions including speech, swallowing, eating, and social interaction [5, 6, 7]. The ability to predict recurrence risk would allow for personalized surveillance schedules, earlier interventions, and better allocation of resources for patients who would benefit most from intensive monitoring.

Current prognostic methods rely mainly on population-based clinical and pathological parameters including TNM staging, tumor site, and patient performance status [8, 9, 10]. While these factors provide general risk stratification, they are not precise enough for individual patient risk prediction. This limitation comes from their inability to account for the biological heterogeneity that drives treatment response and disease progression. As a result, clinicians lack reliable tools to identify which patients have aggressive disease requiring closer surveillance versus those who can safely undergo standard follow-up.

Cancer radiomics has emerged as a promising approach to address this gap by extracting quantitative imaging features that describe tumor characteristics beyond what is visible to the human eye [11, 12, 13]. By treating medical images as high-dimensional data, radiomics can detect subtle patterns in tumor shape, intensity distribution, and texture that may reflect biological processes such as hypoxia, cellular heterogeneity, and microenvironmental characteristics [11, 14]. When combined with clinical parameters, these imaging biomarkers may provide more personalized risk assessment than traditional staging systems alone.

Several studies have shown the potential of radiomic signatures for predicting locoregional recurrence in HNC [15, 16, 17, 18]. However, important gaps limit their clinical use. First, existing models have been mostly developed in Western populations and show poor performance in South Asian cohorts, which have different tumor subtype distributions and disease characteristics [19]. Second, most published radiomics research uses retrospective data collected under varying imaging protocols, introducing technical variability that undermines model reproducibility and external validity [19, 20]. These limitations point to the need for prospective studies with standardized protocols that address population-specific disease patterns.

Beyond data collection issues, radiomics research faces basic methodological problems. Extracting hundreds of imaging features from relatively small patient cohorts creates a classic high-dimensional problem ($p \gg n$), where the risk of overfitting and feature instability is high [20, 21, 22]. Models become very sensitive to feature selection methods and classifier choice, with different analytical approaches often giving contradictory results on the same datasets [21, 23]. Standard feature selection methods often produce unstable signatures that change significantly across data resamples, while complex machine learning algorithms may show impressive training performance but fail on independent test sets [24]. The key question is: which small, stable subset of radiomic and clinical features contains real prognostic information that works in unseen patients across different clinical settings?

This study addresses both the data quality and methodological challenges in radiomics-based recurrence prediction for HNC. We built a prospective, protocol-driven cohort at a major tertiary cancer center in India, ensuring standardized CT imaging and systematic clinical follow-up for recurrence assessment. To address the feature selection challenge, we systematically compared conventional methods (LASSO, SelectKBest) with five metaheuristic optimization algorithms (Particle Swarm Optimization, Whale Optimization, Grey Wolf Optimizer, Genetic Algorithm, Simulated Annealing) [23, 25, 26] across multiple classifier types. We also introduce a hybrid approach that combines Bootstrap-LASSO stability pre-filtering with metaheuristic subset optimization to find robust feature signatures [23, 24]. Through evaluation of 35 different machine learning pipelines across four experimental settings, we identify the best modeling strategies that balance predictive performance with clinical interpretability.

Contributions:

- **C1:** Establishment of a large prospective protocol-driven dataset of head and neck cancer patients (2020–2024) under a standardized imaging protocol and follow-up for recurrence prediction and addressing the need for high-quality data representing South Asian tumor subtype distributions.
- **C2:** Systematic comparison of 35 machine learning pipelines across four experimental settings, showing that simple linear models with carefully selected features outperform complex ensemble methods in high-dimensional, small-sample scenarios.
- **C3:** Identification of clinically interpretable radiomic-clinical signatures linked to recurrence risk, showing that sparse feature sets combining established clinical factors with tumor heterogeneity biomarkers achieve good generalization performance and support individualized follow-up strategies.

2 Related Work

Radiomics has emerged as a powerful approach for prognostic modeling in head and neck cancer, with consistent findings demonstrating the superior performance of combined clinical-radiomic models. Gangil et al. reported that clinico-radiomic models achieved 72% test accuracy for locoregional recurrence prediction in 311 HNC patients, markedly exceeding the performance of clinical-only or radiomics-only approaches. Their systematic comparison of Random Forest, SVM, and XGBoost found that SVM performed best for clinico-radiomic feature sets [15]. Building on this integration approach, Bruixola et al. demonstrated that combining CT-derived radiomic features with TNM stage and clinical factors significantly outperformed TNM-8 staging alone for progression risk stratification in locally advanced HNC [16].

Extending beyond single-modality CT imaging, Hu et al. explored multi-modal approaches using paired PET/CT scans combined with clinical variables, achieving an average AUC of 0.82 for HNSCC recurrence prediction. Their work also incorporated data augmentation strategies, demonstrating improved model robustness through Gaussian noise upsampling that enhanced both sensitivity and specificity [17]. Advanced optimization techniques have also been developed, with Zhang et al. proposing multi-objective approaches that simultaneously optimize sensitivity, specificity, and feature sparsity, addressing the challenge of balancing predictive performance with model interpretability [18].

Despite promising single-institution results, model generalizability remains a critical challenge in radiomics research. Varghese et al. directly addressed this limitation through a multi-center analysis of 562 patients from four institutions for 2-year locoregional recurrence prediction. Their systematic evaluation compared feature selection methods (LASSO vs. univariate filtering), classifiers (logistic regression vs. SVM), and batch effect correction techniques, revealing substantial performance variability with AUC values ranging from 0.56 to 0.68 depending on pipeline configuration and data pooling strategies. While combining multi-institutional data improved SVM performance to approximately 0.68, significant variability persisted across different methodological choices, underscoring the critical need for standardized imaging protocols and robust feature selection methods for successful clinical translation [27].

These studies collectively demonstrate both the promise and limitations of current radiomics approaches for HNC recurrence prediction. While clinical-radiomic integration consistently outperforms single-modality approaches, the substantial inter-institutional variability underscores the critical importance of robust feature selection methods and standardized imaging protocols for successful clinical translation.

3 Methods

3.1 Patient Cohort and Study Design

A prospective study was conducted between 2020 and 2024 at Christian Medical College Vellore under a standardized imaging and clinical protocol designed to minimize technical variability. Inclusion criteria comprised patients aged 18–70 years with Eastern Cooperative Oncology Group (ECOG) performance status ≤ 2, locally advanced disease (stage >T2 and/or N1+), high-quality baseline contrast-enhanced CT imaging acquired prior to treatment, and complete one-year clinical follow-up with documented locoregional recurrence status [28]. Patients with prior head and neck malignancy, distant metastatic disease, previous radiotherapy exposure, imaging artifacts compromising tumor segmentation, missing primary gross tumor volume (GTVp) delineation, absence of planning CT scans, and incomplete treatment courses were excluded. The study protocol received institutional review board approval, and all patient data were anonymized prior to analysis. Figure 1 illustrates the patient selection process and final dataset composition.

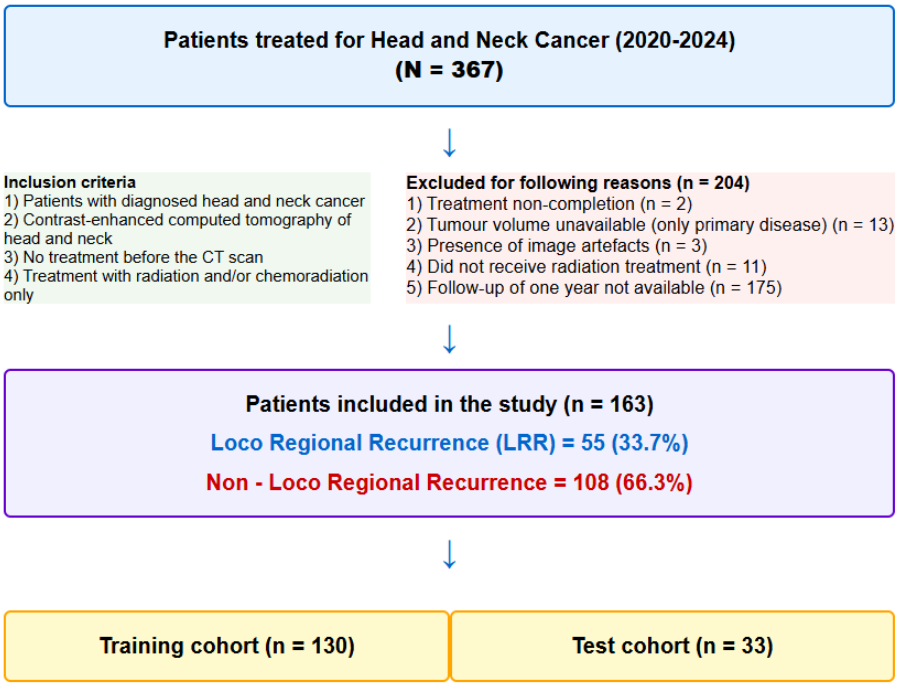

Figure 1: Patient selection and dataset splitting flowchart

3.2 Image Acquisition and Preprocessing

All CT images were contrast-enhanced and acquired using institutional scanners (Siemens Biograph 6, SOMATOM Definition AS, GE Discovery CT750 HD) with an energy range of 100.0–130.0 kVp, an exposure range of 5.0–350.0 mAs, and 512×512 matrix with slice thickness of 2.5–3.0 mm. In-plane resolution ranged from 0.78125×0.78125 mm^2 to 1.367188×1.367188 mm^2. Following image acquisition, gross tumour volumes were delineated by radiation oncologists. The tumor volumes as visible on the planning CT were manually delineated following standard HNC contouring guidelines and labelled as the gross tumour volume (GTV). The tumour outlines (masks) were exported in DICOM RTSTRUCT format.

3.3 Feature Extraction and Integration

Image preprocessing was performed prior to radiomic feature extraction to ensure standardization across different scanners. All preprocessing steps followed the Image Biomarker Standardization Initiative (IBSI) guidelines [29]. The CT images and their masks were resampled to $1 \times 1 \times 1$ mm^3 isotropic voxels using B-spline interpolation to standardize spatial resolution across different scanners. Intensity discretization was performed using a fixed bin width of 25 Hounsfield Units to optimize texture feature stability while maintaining sensitivity to clinically relevant tissue variations.

Following preprocessing, radiomic features were extracted using PyRadiomics v3.1.0 [30], yielding 103 features per GTV: first-order intensity statistics (18 features) capturing distribution properties including mean, variance, skewness, and kurtosis; shape descriptors (14 features) quantifying morphological properties such as volume, surface area, sphericity, and maximum 2D/3D diameters; and texture features (71 features) derived from grey-level co-occurrence matrices (GLCM, 24 features), run-length matrices (GLRLM, 16 features), size-zone matrices (GLSZM, 16 features), dependence matrices (GLDM, 14 features), and neighbouring grey-tone difference matrices (NGTDM, 5 features).

In addition to radiomics features, clinical variables were incorporated into the analysis. The clinical variables were transformed from discrete to numerical format: age, weight, primary tumor location, T/N/M staging, AJCC stage (7th ed), and HPV/p16 status. Missing values were handled using median replacement for continuous variables and conservative assumptions for staging parameters. This resulted in a total of 8 clinical features, which could be combined with the 103 radiomics features for integrated analysis.

3.4 Feature Selection Strategies

To address the high-dimensional challenge ($p \gg n$), seven feature selection methods were implemented to reduce dimensionality and mitigate overfitting. All feature selection and hyperparameter tuning were performed exclusively within the training set using stratified 5-fold cross-validation to prevent information leakage. Three feature selection approaches were evaluated: direct, metaheuristic, and hybrid. Direct approaches apply a single feature selection method to identify optimal feature subsets in one step. Metaheuristic approaches use population-based optimization algorithms to directly search for optimal feature combinations. Hybrid approaches combine multiple methods sequentially, using an initial stability selection method to pre-filter features before applying a metaheuristic algorithm for final feature subset optimization.

Direct Selection Approach: A single-feature selection method was applied within each training fold to identify the optimal feature subsets. Bootstrap-LASSO implemented stability selection through 1,000 stratified bootstrap resamples per fold, applying L1-penalized logistic regression across a log-spaced range of regularization parameter C to identify features with consistent selection patterns. SelectKBest employed univariate filtering using ANOVA F-statistics, evaluating various k values (5 to 20) to select the number of top-ranked features yielding optimal inner-CV ROC-AUC performance.

Metaheuristic Selection Approach: Five population-based algorithms conducted direct feature subset optimization, each enforcing exactly 6 features to balance model complexity with interpretability. Particle Swarm Optimization utilizes swarm intelligence with particles navigating the search space based on personal and global best solutions, balancing population size and iterations for the exploration-exploitation trade-off. Whale Optimization Algorithm mimicked humpback whales' bubble-net hunting behavior, alternating between encircling prey and spiral motion phases following original WOA parameter settings. Grey Wolf Optimizer simulated wolf pack social hierarchy with alpha/beta/delta leadership structure, employing standard update equations without additional hyperparameters. Genetic Algorithm implemented evolutionary search using binary chromosome encoding, with crossover rates (~0.8) and mutation rates (~0.02) optimized for feature selection tasks. Simulated Annealing applied probabilistic optimization inspired by metallurgical annealing processes, utilizing cooling schedules designed to balance global exploration with convergence stability.

Hybrid Selection Approach: This two-stage process first applied Bootstrap-LASSO stability selection to narrow the feature pool to 10–20 stable candidates using the frequency-threshold procedure. Subsequently, metaheuristic algorithms (PSO/WOA/GWO/GA/SA) refined this reduced feature set to identify optimal 6-feature subsets, using mean inner-CV ROC-AUC as the fitness criterion. Table 1 provides comprehensive parameter configurations for all feature selection methods.

3.5 Model Building and Evaluation

To evaluate prediction performance, four modeling configurations were created by combining two factors. A modeling configuration refers to a specific combination of feature input type and feature selection strategy. The two feature input types were: radiomics-only (103 features) and combined radiomics-clinical (111 features). The two feature selection strategies were: direct selection (applying Bootstrap-LASSO, SelectKBest, or metaheuristic algorithms in a single step) and hybrid selection (two-stage Bootstrap-LASSO followed by metaheuristic refinement). This resulted in four configurations: (i) radiomic features with direct selection; (ii) radiomic + clinical features with direct selection; (iii) radiomic features with hybrid selection; and (iv) radiomic + clinical features with hybrid selection.

Five classifiers were evaluated with comprehensive hyperparameter optimization via GridSearchCV using 5-fold stratified cross-validation. All random aspects (e.g., data shuffling, model initialization) were controlled with fixed random seeds (e.g.,

Table 1: Feature Selection Methods and Parameter Configurations

Method	Key Parameters	Rationale & Implementation Details
Bootstrap-LASSO	1,000 bootstrap samples Frequency threshold: 70% Adaptive threshold: 0.7→0.2	High sample size ensures stability; frequency threshold balances stability vs sensitivity; adaptive threshold maintains 10-20 features preventing over/under-selection
SelectKBest (ANOVA)	F-test ranking k optimization: 5-20 Cross-validation	Univariate statistical ranking with data-driven k selection; no classifier bias in selection process (purely statistical approach)
Particle Swarm Optimization	20 particles, 50 iterations Inertia weight: 0.7 c_1=1.4, c_2=1.4	Balances exploration vs computational efficiency; moderate momentum preservation; balanced personal vs global learning; sigmoid transfer function for binary decisions
Whale Optimization Algorithm	20 whales, 50 iterations Spiral probability: 0.5 Shape parameter: 1.0	Adequate population for solution diversity; equal exploitation/exploration balance; standard spiral tightness; alternates between encircling prey ($\|A\| < 1$) and random search ($\|A\| \geq 1$)
Grey Wolf Optimizer	12 wolves, 20 iterations Hierarchical structure Parameter a: 2→0	Smaller population with faster convergence; alpha/beta/delta leadership mimics natural pack behavior; linear parameter reduction decreases exploration over time
Genetic Algorithm	Population: 40, Gen: 60 Crossover rate: 0.8 Mutation rate: 0.02 Elitism: 10%	Sufficient genetic diversity; tournament selection (size=3) provides moderate selection pressure; high recombination with low disruption; preserves best solutions
Simulated Annealing	1,000 iterations Initial temp: 100 Cooling rate: 0.95	Thorough search space exploration; high initial acceptance probability; gradual temperature reduction; exp(Δf/T) allows escaping local optima early in search

`random_state=42`) to ensure reproducibility across all experiments. The final chosen hyperparameters for each model correspond to those yielding the best cross-validated AUC in the training data. Table 2 details the complete hyperparameter search spaces for each classifier.

Classifier performance was evaluated using ROC-AUC with 95% confidence intervals (CI) computed via 1000 bootstrap resamples of the test set. Models were trained using 5-fold stratified cross-validation on the training cohort (n = 130), with optimal hyperparameters selected based on cross-validation performance. Test set results are reported as AUC [95% CI] to assess model generalization capability on unseen data.

Table 2: Hyperparameter Configurations for Machine Learning Classifiers

Classifier	Parameters and Search Values
Logistic Regression	C: [1e-3, 3e-3, 1e-2, 3e-2, 0.1, 0.3, 1, 3, 10]; penalty: ['l1', 'l2']; solver: ['liblinear']; class_weight: [None, 'balanced']; max_iter: [1000]
Gaussian Naive Bayes	No tunable parameters
Support Vector Machine	C: [1e-3, 1e-2, 0.1, 1, 10]; kernel: ['linear', 'rbf']; gamma (RBF only): ['scale', 'auto', 1e-3, 1e-2, 1e-1]; probability: [True]; class_weight: [None, 'balanced']
Decision Tree	criterion: ['gini', 'entropy']; max_depth: [2, 3, 4, 5, 7]; min_samples_split: [5, 10, 15]; min_samples_leaf: [2, 4, 6]; max_features: ['sqrt', 'log2', None]; class_weight: [None, 'balanced']
Random Forest	n_estimators: [100, 200, 400]; max_depth: [3, 5, 7, 10]; min_samples_split: [5, 10]; min_samples_leaf: [2, 4]; max_features: ['sqrt', 'log2']; bootstrap: [True]; class_weight: [None, 'balanced', 'balanced_subsample']

4 Results

We analyzed 163 prospectively recruited HNC patients (55 with LRR and 108 disease-free) treated with chemoradiation between 2020 and 2024. Most patients were male (143, 88%) with a median age of 62 years (range 22–85 years). Most patients had a history of tobacco use (93 patients, 57%), while 73 patients (45%) were non-smokers. Chewable tobacco use was reported in 60 patients (37%).

Most patients had primary tumor location as larynx (62 patients, 38%), followed by the hypopharynx (32 patients, 20%) and the nasopharynx (25 patients, 15%). According to TNM staging, T3 tumors were most common (63 patients, 39%), followed by T4a (36 patients, 22%) and T2 (36 patients, 22%). Nodal involvement was present in 98 patients (60%), while 65 patients (40%) had N0 disease. Overall, AJCC staging showed Stage III disease in 53 patients (33%), Stage IVA in 41 patients (25%),

and Stage IVB in 32 patients (20%). HPV/p16 status was available for 58 patients, with 10 (17%) testing positive. Patients were randomly divided into training (130 patients, 80%) and independent test (33 patients, 20%) sets.

Using radiomic features exclusively (103 features), performance varied across classifiers and feature selection methods (Table 3). Decision Tree with SelectKBest achieved the highest test AUC of 0.73 [0.53, 0.88]. SVM performance was highly variable, ranging from 0.32 [0.14, 0.54] with Grey Wolf Optimizer to 0.64 [0.39, 0.85] with SelectKBest.

Table 3: Performance of Machine Learning Models Using Radiomic Features Only

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.64 [0.42, 0.83]	0.65 [0.40, 0.86]	0.65 [0.41, 0.85]	0.64 [0.41, 0.85]	0.69 [0.47, 0.88]	0.62 [0.40, 0.83]	0.65 [0.41, 0.87]
Naive Bayes	0.68 [0.47, 0.86]	0.65 [0.40, 0.87]	0.69 [0.48, 0.88]	0.66 [0.47, 0.84]	0.71 [0.50, 0.89]	0.67 [0.47, 0.84]	0.65 [0.43, 0.83]
SVM	0.36 [0.17, 0.59]	0.64 [0.39, 0.85]	0.35 [0.15, 0.60]	0.36 [0.15, 0.60]	0.32 [0.14, 0.54]	0.63 [0.39, 0.85]	0.36 [0.15, 0.60]
Decision Tree	0.68 [0.48, 0.85]	0.73 [0.53, 0.88]	0.63 [0.42, 0.82]	0.55 [0.35, 0.73]	0.66 [0.48, 0.83]	0.65 [0.46, 0.82]	0.53 [0.30, 0.76]
Random Forest	0.67 [0.45, 0.84]	0.63 [0.40, 0.84]	0.66 [0.44, 0.84]	0.62 [0.41, 0.80]	0.68 [0.48, 0.86]	0.62 [0.41, 0.80]	0.65 [0.42, 0.84]

Adding 8 clinical variables to the radiomic feature set (111 total features) altered performance patterns across classifiers (Table 4). Logistic Regression with Grey Wolf Optimizer achieved the highest test AUC of 0.81 [0.62, 0.95]. SVM performance continued to remain highly variable, ranging from 0.34 [0.28, 0.56] with Whale Optimization Algorithm to 0.79 [0.62, 0.93] with Grey Wolf Optimizer.

Table 4: Performance of Machine Learning Models Using Radiomics and Clinical Features

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.78 [0.58, 0.94]	0.66 [0.42, 0.88]	0.74 [0.52, 0.92]	0.62 [0.49, 0.83]	**0.81 [0.62, 0.95]**	0.75 [0.54, 0.92]	0.74 [0.52, 0.92]
Naive Bayes	0.71 [0.50, 0.91]	0.63 [0.38, 0.86]	0.67 [0.43, 0.89]	0.68 [0.57, 0.83]	0.79 [0.60, 0.93]	0.75 [0.54, 0.91]	0.73 [0.52, 0.90]
SVM	0.78 [0.57, 0.94]	0.35 [0.13, 0.58]	0.72 [0.50, 0.90]	0.34 [0.28, 0.56]	0.79 [0.62, 0.93]	0.74 [0.52, 0.93]	0.74 [0.53, 0.91]
Decision Tree	0.59 [0.38, 0.76]	0.68 [0.46, 0.87]	0.45 [0.25, 0.64]	0.58 [0.50, 0.70]	0.60 [0.40, 0.80]	0.62 [0.41, 0.80]	0.70 [0.49, 0.88]
Random Forest	0.68 [0.45, 0.88]	0.64 [0.40, 0.85]	0.67 [0.43, 0.86]	0.64 [0.57, 0.71]	0.73 [0.55, 0.89]	0.74 [0.53, 0.92]	0.72 [0.51, 0.89]

Bootstrap-LASSO combined with metaheuristic optimization produced different performance patterns compared to direct selection (Tables 5–6). For radiomic-only features, the highest test AUC was 0.72 [0.51, 0.89] achieved by Naive Bayes with Grey Wolf Optimizer and Simulated Annealing. For combined features, the highest performance was 0.77 [0.58, 0.92] and 0.77 [0.57, 0.93] achieved by Naive Bayes with Whale Optimization Algorithm and Grey Wolf Optimizer respectively.

Table 5: Performance of Machine Learning Models Using Hybrid Feature Selection with Radiomic Features Only

Classifier	PSO	WOA	GWO	GA	SA
Logistic Regression	0.66 [0.45, 0.86]	0.65 [0.43, 0.85]	0.66 [0.44, 0.86]	0.65 [0.41, 0.85]	0.68 [0.45, 0.87]
Naive Bayes	0.71 [0.50, 0.89]	0.71 [0.50, 0.88]	0.72 [0.51, 0.89]	0.68 [0.46, 0.87]	0.72 [0.51, 0.88]
SVM	0.34 [0.15, 0.56]	0.35 [0.15, 0.57]	0.34 [0.15, 0.55]	0.36 [0.15, 0.59]	0.32 [0.15, 0.55]
Decision Tree	0.64 [0.44, 0.81]	0.67 [0.46, 0.84]	0.53 [0.34, 0.71]	0.59 [0.37, 0.78]	0.62 [0.43, 0.79]
Random Forest	0.72 [0.54, 0.88]	0.70 [0.50, 0.88]	0.69 [0.47, 0.87]	0.67 [0.45, 0.85]	0.69 [0.48, 0.87]

Across all experimental configurations, Logistic Regression with Grey Wolf Optimizer feature selection using combined radiomic and clinical features achieved the highest test performance (AUC 0.81 [0.62, 0.95]). This model selected a 10-feature signature comprising 4 clinical variables (Age, AJCC_Stage, T_Stage, Location) and 6 radiomic features (Maximum2DDiameterSlice, MinorAxisLength, LargeDependenceEmphasis, RunLengthNonUniformityNormalized, Idm, Imc1).

5 Discussion

This study systematically evaluated machine learning pipelines for locoregional recurrence prediction in head and neck cancer using prospectively collected data. Our most salient finding is that a simple 10-feature clinical-radiomic signature using Logistic Regression with Grey Wolf Optimizer achieved the best held-out performance (test AUC 0.81), outperforming complex ensemble methods and hybrid feature selection approaches.

Our finding that combined clinical-radiomic models outperform radiomics-alone approaches aligns with other reported findings. Gangil et al. reported that clinico-radiomic models achieved 72% accuracy for LRR prediction in 311 HNC patients, markedly exceeding single-modality approaches [15]. Similarly, Bruixola et al. demonstrated that combining CT-derived

Table 6: Performance of Machine Learning Models Using Hybrid Feature Selection with Radiomics and Clinical Features

Classifier	PSO	WOA	GWO	GA	SA
Logistic Regression	0.75 [0.54, 0.92]	0.70 [0.47, 0.88]	0.75 [0.53, 0.93]	0.74 [0.53, 0.91]	0.74 [0.51, 0.93]
Naive Bayes	0.71 [0.49, 0.90]	0.77 [0.58, 0.92]	0.77 [0.57, 0.93]	0.69 [0.46, 0.89]	0.74 [0.52, 0.91]
SVM	0.68 [0.45, 0.88]	0.74 [0.53, 0.90]	0.69 [0.46, 0.90]	0.69 [0.45, 0.89]	0.67 [0.42, 0.88]
Decision Tree	0.68 [0.44, 0.90]	0.61 [0.38, 0.80]	0.71 [0.52, 0.87]	0.63 [0.42, 0.83]	0.54 [0.32, 0.75]
Random Forest	0.68 [0.44, 0.88]	0.75 [0.55, 0.92]	0.72 [0.49, 0.91]	0.69 [0.45, 0.90]	0.75 [0.53, 0.94]

radiomic features with TNM staging significantly outperformed TNM-8 staging alone for progression risk stratification [16]. Our optimal performance (test AUC 0.81) is also in the range reported by prior CT-radiomics studies focused on LRR after chemoradiotherapy (typically ~0.7–0.8), such as Keek et al., who developed a (peri)tumoral CT radiomic signature for LRR/DM risk stratification [31]. However, our results reveal important nuances not emphasized in prior work. While Gangil et al. found SVM performed best for combined features [15], our data show SVM response to clinical integration is highly method dependent. Some feature selection combinations achieved strong performance (LASSO: 0.78, GWO: 0.79), while others remained poor (SelectKBest: 0.35, WOA: 0.34), indicating that SVM's reported superiority may reflect specific methodological choices rather than consistent algorithmic advantages. In contrast, Logistic Regression demonstrated stable improvement across all seven feature selection methods (range 0.62–0.81), suggesting greater methodological robustness for clinical deployment.

Systematic overfitting was observed in Random Forest and Decision Tree models, despite ensemble design, highlighting fundamental challenges in high-dimensional, small-sample radiomics. Varghese et al.'s multi-institutional study similarly found substantial performance variability (AUC 0.56–0.68) depending on pipeline complexity, supporting our observation that sophisticated methods don't guarantee superior performance [27]. Their finding that even multi-center data pooling achieved only modest improvements (SVM AUC ~0.68) underscores the persistent challenge of model generalization in radiomics research. Our observation that simple linear models outperformed complex ensembles contrasts with some radiomics literature but aligns with emerging evidence about small-sample limitations [24]. The persistent overfitting we documented—with Random Forest achieving training AUC 0.90–0.98 but test performance dropping to 0.62–0.74—demonstrates that theoretical overfitting resistance fails in extreme high-dimensional scenarios [24]. This pattern remained consistent across all feature selection methods and configurations, indicating a fundamental rather than methodological limitation. High-dimensional data with small sample sizes is particularly prone to biased machine learning performance estimates, with models learning noise rather than underlying patterns [24].

Although two-stage selection (Bootstrap-LASSO prefilter → metaheuristic subset search) is appealing for stability, it did not outperform direct selection here (best hybrid test AUC 0.77 vs best direct 0.81). Most studies assume sophisticated feature selection improves results, but our systematic comparison suggests this assumption requires empirical validation. Zhang et al.'s multi-objective optimization achieved good performance but required complex computational frameworks [18], whereas our simpler Grey Wolf Optimizer approach achieved comparable results with greater interpretability. The 4-point AUC difference between hybrid and direct methods, while seemingly modest, represents meaningful clinical discrimination. This finding suggests that two-stage filtering processes, while conceptually appealing for stability, may inadvertently remove informative features during pre-filtering stages. Future radiomics methodology research should empirically validate rather than assume the benefits of increased selection complexity.

The clinical variables in our optimal signature (Age, AJCC_Stage, T_Stage, Location) are well-established prognostic factors extensively validated in HNC literature [8, 9, 10]. Age has been consistently demonstrated as an independent prognostic factor with hazard ratios of approximately 1.04 per year [10], while AJCC staging system incorporates tumor and nodal characteristics that fundamentally determine treatment approach and prognosis [8, 9]. The selected radiomic features align with biological mechanisms of treatment resistance. Shape features (Maximum2DDiameterSlice, MinorAxisLength) capture tumor geometry, consistent with evidence that larger, irregular tumors have worse outcomes [32, 33]. Texture features (LargeDependenceEmphasis, RunLengthNonUniformityNormalized, Idm, Imc1) quantify intratumoral heterogeneity, which correlates with hypoxia, necrosis, and cellular density variations—factors known to influence radiosensitivity [33, 34, 35]. Radiomic texture features measure the spatial distribution relationship of voxel intensities and provide quantitative assessment of tumor heterogeneity that reflects underlying biological processes including metabolic activity, oxygenation levels, and vascularization [34, 35]. Our sparse 6-feature radiomic component contrasts with studies using larger feature sets, which may contribute to our model's superior generalization. The 4:6 clinical-to-radiomic ratio suggests imaging biomarkers provide substantial but not overwhelming prognostic value relative to established clinical parameters, supporting the complementary rather than replacement role of radiomics in clinical assessment.

Our single-institution performance (test AUC 0.81) exceeds the multi-institutional results reported by Varghese et al. (0.56–0.68) [27], but this comparison must be interpreted cautiously. Single-institution studies benefit from protocol homogeneity but risk overfitting to local characteristics, while multi-center studies better represent real-world heterogeneity despite reduced performance. The wide confidence intervals in our optimal model [0.62, 0.95] reflect this uncertainty, with

the upper bound potentially representing overfitted local patterns rather than generalizable performance. Their systematic evaluation of batch effect correction and feature harmonization techniques [27] highlights challenges we did not address in our single-institution design. Their finding that model AUCs varied substantially with harmonization approaches suggests our results may not directly translate to external institutions without similar preprocessing considerations.

Several critical limitations must be acknowledged. Our small test set (n = 33) produces wide confidence intervals [0.62, 0.95] that substantially exceed the uncertainty reported in larger studies like Varghese et al.'s 562-patient cohort [27]. This uncertainty indicates our model's true performance could range from marginally useful to highly discriminative, emphasizing the essential need for external validation before clinical consideration. The single-institution design limits generalizability across different imaging protocols, scanner vendors, and patient populations. Multi-institutional validation studies incorporating standardized protocols and harmonized feature extraction pipelines are required to establish broader applicability and assess model robustness across diverse clinical settings. The prospective data collection under standardized protocols is a strength, but retrospective analysis of this prospectively collected data limits assessment of real-world clinical utility and decision-making impact.

6 Conclusion

This study demonstrates that combined radiomic-clinical models significantly outperform radiomic-only approaches for HNC locoregional recurrence prediction. Key contributions include: (1) systematic evaluation showing clinical variables are essential for effective radiomic model performance; (2) demonstration that simple linear models outperform complex ensembles in high-dimensional, small-sample scenarios; (3) evidence that sophisticated hybrid feature selection provides no advantage over direct methods; and (4) identification of an optimal feature signature combining clinical parameters with tumor morphology and texture features using routine imaging and clinical data. Future work requires external validation with multi-institutional datasets, prospective clinical trials to assess real-world impact, and investigation of additional imaging modalities and molecular biomarkers for enhanced predictive accuracy.

References

[1] David I. Conway, M. Purkayastha, and Ivor G. Chestnutt. The changing epidemiology of oral cancer: definitions, trends, and risk factors. *British Dental Journal*, 225:867–873, 2018.

[2] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[3] Jer-Hwa Chang, Chia-Che Wu, Kevin Sheng-Po Yuan, Alexander T. H. Wu, and Szu-Yuan Wu. Locoregionally recurrent head and neck squamous cell carcinoma: incidence, survival, prognostic factors, and treatment outcomes. *Oncotarget*, 8(33):55600–55612, 2017.

[4] Gemma Gatta, Laura Botta, Maria José Sánchez, Lesley A. Anderson, Daniela Pierannunzio, Lisa Licitra, and EUROCARE Working Group. Prognoses and improvement for head and neck cancers diagnosed in europe in early 2000s: The eurocare-5 population-based study. *European Journal of Cancer*, 51(15):2130–2143, 2015.

[5] Carly T. Haring, Shamus M. Dermody, Prakash Yalamanchi, Steven Y. Kang, Michelle Shepard, Alexandra R. Bukatko, Kara Wooten, William R. Ryan, Andrew C. Birkeland, J. Chad Brenner, Mark E. P. Prince, Andrew G. Shuman, Matthew E. Spector, and Keith A. Casper. Patterns of recurrence in head and neck squamous cell carcinoma to inform personalized surveillance protocols. *Cancer*, 129(17):2722–2732, 2023.

[6] Scott N. Rogers, S. Gwanne, D. Lowe, Gerry Humphris, Bevan Yueh, and Ernest A. Weymuller, Jr. The addition of mood and anxiety domains to the university of washington quality of life scale. *Head & Neck*, 24(6):521–529, 2002.

[7] Marc A. List, Laura L. D'Antonio, David F. Cella, Amy Siston, Patti Mumby, Daniel Haraf, and Everett Vokes. The performance status scale for head and neck cancer patients and the functional assessment of cancer therapy-head and neck scale: A study of utility and validity. *Cancer*, 77(11):2294–2301, 1996.

[8] Dennis K. Zanoni, Snehal G. Patel, and Jatin P. Shah. Changes in the 8th edition of the american joint committee on cancer (ajcc) staging of head and neck cancer: Rationale and implications. *Current Oncology Reports*, 21(6):52, 2019.

[9] Margaret Brandwein-Gensler and Richard V. Smith. Prognostic indicators in head and neck oncology including the new 7th edition of the ajcc staging system. *Head and Neck Pathology*, 4(1):53–61, 2010.

[10] Gabriella Cadoni et al. Prognostic factors in head and neck cancer: a 10-year retrospective analysis in a single-institution in italy. *Acta Otorhinolaryngologica Italica*, 37(6):458–466, 2017.

[11] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, 2012.

[12] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5:4006, 2014.

[13] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

[14] Chintan Parmar et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific Reports*, 5:11044, 2015.

[15] Tanmay Gangil, Kundan Sharan, B. Dinesh Rao, Karthick Palanisamy, Bidyut Chakrabarti, and Rajagopal Kadavigere. Utility of adding radiomics to clinical features in predicting the outcomes of radiotherapy for head and neck cancer using machine learning. *PLoS ONE*, 17(12):e0277168, 2022.

[16] Gisela Bruixola, David Dualde-Beltrán, Ana Jimenez-Pastor, Albert Nogué, Francisco Bellvís, Alejandro Fuster-Matanzo, Cristina Alfaro-Cervelló, Nuria Grimalt, Nadia Salhab-Ibáñez, Valentín Escorihuela, M. Emilia Iglesias, Mercedes Maroñas, Ángel Alberich-Bayarri, Andrés Cervantes, and Noemí Tarazona. Ct-based clinical-radiomics model to predict progression and drive clinical applicability in locally advanced head and neck cancer. *European Radiology*, 35(7):4277–4288, 2025.

[17] Yuliang Hu, Karla Taing, Jiawei Wang, David Sher, and Michael Dohopolski. Enhancing prediction of primary site recurrence in head and neck cancer using radiomics and uncertainty estimation. *Frontiers in Artificial Intelligence*, 8:1623393, 2025.

[18] Qiongwen Zhang, Kai Wang, Zhiguo Zhou, et al. Predicting local persistence/recurrence after radiation therapy for head and neck cancer from pet/ct using a multi-objective, multi-classifier radiomics model. *Frontiers in Oncology*, 12:955712, 2022.

[19] Devadhas Devakumar et al. Framework for machine learning of ct and pet radiomics to predict local failure after radiotherapy in locally advanced head and neck cancers. *Journal of Medical Physics*, 46(3):181–188, 2021.

[20] Abdalla Ibrahim, Sergey Primakov, Manon Beuque, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods*, 188:20–29, 2021.

[21] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo J. W. L. Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific Reports*, 5:13087, 2015.

[22] Janita E. van Timmeren, Ralph T. H. Leijenaar, Wouter van Elmpt, et al. Test-retest data for radiomics feature stability analysis: A phantom study. *Radiotherapy and Oncology*, 121(3):440–446, 2016.

[23] Xinzhi Pan et al. A multi-objective based radiomics feature selection method for response prediction following radiotherapy. *Physics in Medicine & Biology*, 68(5), 2023.

[24] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. Machine learning algorithm validation with a limited sample size. *PLoS One*, 14(11):e0224365, 2019.

[25] Fan Yang et al. A hybrid feature selection algorithm combining information gain and grouping particle swarm optimization for cancer diagnosis. *PLoS One*, 19(3):e0290332, 2024.

[26] Simone A. Ludwig. Guided particle swarm optimization for feature selection: Application to cancer genome data. *Algorithms*, 18(4):220, 2025.

[27] Abby J. Varghese et al. Multi-centre radiomics for prediction of recurrence following radical radiotherapy for head and neck cancers: Consequences of feature selection, machine learning classifiers and batch-effect harmonization. *Physics and Imaging in Radiation Oncology*, 26:100450, 2023.

[28] Martin M. Oken, Richard H. Creech, Douglass C. Tormey, et al. Toxicity and response criteria of the eastern cooperative oncology group. *American Journal of Clinical Oncology*, 5(6):649–655, 1982.

[29] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020.

[30] Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.

[31] Simon A. Keek, Frank W. R. Wesseling, Henry C. Woodruff, et al. A prospectively validated prognostic model for patients with locally advanced squamous cell carcinoma of the head and neck based on radiomics of computed tomography images. *Cancers*, 13(13):3271, 2021.

[32] Khoi V. Huynh, Grace Chen, Muhammad M. Qureshi, et al. Head and neck cancer treatment outcome prediction: a comparison between machine learning with conventional radiomics features and deep learning radiomics. *Frontiers in Medicine*, 10:1217037, 2023.

[33] Marta Bogowicz, Stephanie Tanadini-Lang, Matthias Guckenberger, and Oliver Riesterer. Combined ct radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Scientific Reports*, 9:15198, 2019.

[34] Camil Ciprian Mirestean et al. Radiomic machine learning and texture analysis - new horizons for head and neck oncology. *Maedica*, 14(2):126–130, 2019.

[35] Lei Wang, Tao Dong, Bo Xin, et al. Application of radiomics and machine learning in head and neck cancers. *International Journal of Biological Sciences*, 17(2):475–486, 2021.