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

Anonymous Submission

Abstract

Purpose: Predicting risk of recurrence remains a major challenge in head and neck cancer (HNC), current clinical practice fails to account for patient-specific variability in tumour biology. Radiomics, by converting CT scans into quantitative descriptors, provides a high-dimensional substrate for computational risk modelling. This study uses a prospectively collected HNC cohort (2020–2024) to carefully test different machine learning (ML) pipelines with metaheuristic feature selection for predicting LRR, with the goal of building reproducible and clinically relevant risk-adaptive tools.

Methods: A total of 1466 patients were enrolled prospectively under a standardized radiomics protocol, with 367 primary radiation patients selected for analysis based on inclusion/exclusion criteria. Contrast-enhanced planning CT scans were used to extract first-order, shape, and texture radiomic features. We constructed 42 ML pipelines by combining six classifiers (Logistic Regression, Naive Bayes, Linear SVM, RBF SVM, Decision Tree, Random Forest) with seven feature selection methods: SelectKBest, LASSO, and five population-based metaheuristic algorithms [Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), Genetic Algorithm (GA), Simulated Annealing (SA)]. Hybrid pipelines were also developed, where Bootstrap-LASSO pre-filtered stable features before metaheuristic selection to overcome overfitting. Models were trained and evaluated under stratified 5-fold cross-validation and independent test splits, with ROC AUC and accuracy as primary metrics.

Results: The Bootstrap-LASSO + PSO + Random Forest pipeline achieved the highest test performance (ROC AUC = 0.80, accuracy = 0.84). Comparable results were obtained with Bootstrap-LASSO + GWO and Bootstrap-LASSO + GA. Recurrent predictive features included shape descriptors (e.g., maximum 2D diameters, sphericity) and texture heterogeneity measures (e.g., GLCM information measures, GLRLM run-length non-uniformity), suggesting that tumour geometry and intra-tumoural texture patterns are key markers of recurrence risk.

Conclusions: This study demonstrates that prospectively collected, protocol-driven datasets combined with metaheuristic-enhanced ML pipelines provide a robust and interpretable strategy for recurrence prediction in HNC. By uniting data standardization with algorithmic benchmarking, this work establishes a foundation for risk-adaptive follow-up strategies and sets the stage for future multicentre validation.

Keywords— Head and Neck Cancer, Deep Learning, Radiomics

1 Introduction

Locoregional recurrence (LRR) in head and neck cancer (HNC) remains one of the most pressing challenges in oncology. Despite advances in radiation delivery and multimodality care, recurrence rates remain high, eroding long-term survival and quality of life. The challenge lies not only in treating recurrence, but in anticipating it early enough to enable adaptive follow-up and timely intervention.

Over the past decade, radiomics has emerged as a promising approach to address this gap by converting routine CT imaging into quantitative descriptors of tumour phenotype. These features capture aspects of tumour shape, intensity, and texture that may reflect biological heterogeneity invisible to human observers. However, progress in radiomics-based recurrence prediction has been limited by three persistent issues: (i) Most prior studies rely on retrospective cohorts collected under heterogeneous imaging protocols, introducing uncontrolled variability and undermining reproducibility; and (ii) High dimensionality of radiomic features ($p \gg n$) makes models highly sensitive to the choice of feature selection and classifier, leading to instability and inconsistent results.

Our work addresses these gaps by combining prospective, protocol-driven data collection with rigorous benchmarking of machine learning (ML) pipelines. Between 2020 and 2024, a cohort of more than 1466 HNC patients was enrolled under a standardized radiomics protocol, from which 367 primary radiation patients were selected for analysis. Our dataset includes high-quality imaging, clear rules for selecting patients, and reliable follow-up, resulting in a higher recurrence rate (43% within one year) that makes it strong for building predictive models. On the algorithmic side, we systematically evaluate the interaction between feature selection and classification, incorporating not only conventional approaches such as SelectKBest and LASSO, but also a family of population-based metaheuristic algorithms that can navigate complex feature spaces. To further enhance stability, we introduce a hybrid feature selection framework that combines Bootstrap-LASSO pre-filtering with metaheuristic search.

The key contributions of this study are as follows:

- **C1:** Establishment of a large prospective protocol-driven dataset of head and neck cancer patients (2020–2024) under a standardized radiomics imaging protocol and follow-up, with 367 primary radiation patients selected for recurrence prediction.

- **C2:** Development and benchmarking of 42 ML pipelines combining six classifiers with seven feature selection strategies, including a novel hybrid framework (Bootstrap-LASSO + metaheuristics), demonstrating improved robustness compared to conventional methods.

- **C3:** Identification of reproducible radiomic signatures (shape and texture heterogeneity features) linked to recurrence risk, highlighting their potential translational relevance for individualized follow-up strategies.

2 Related Work

3 Methods

This prospective study was conducted between 2020 and 2024 at Christian Medical College under a standardized radiomics protocol. A total of 1466 patients with histopathologically confirmed head and neck squamous cell carcinoma (HNSCC) were recruited, of which 367 patients treated with primary radiation were selected for analysis. Eligibility required patients to be 18–70 years of age, an ECOG performance score ≤ 2, stage >T2 and/or N1+ disease, baseline contrast-enhanced CT before treatment, and at least one year of structured follow-up or documented locoregional recurrence (LRR). Patients with prior head and neck cancer, distant metastasis at baseline, previous radiotherapy, imaging artifacts interfering with segmentation, or incomplete treatment were excluded. Gross tumour volumes (GTVs) were delineated by expert radiation oncologists according to institutional guidelines and exported in DICOM RTSTRUCT format. All CT images were acquired using standard radiotherapy planning CT protocols with a 512×512 matrix and 3 mm slice thickness. Ethical approval was obtained, and data were anonymized before analysis.

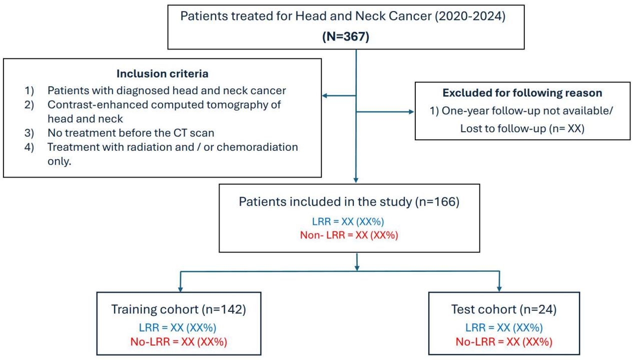
Patients treated for Head and Neck Cancer (2020-2024)
(N=367)
Inclusion criteria
1) Patients with diagnosed head and neck cancer
2) Contrast-enhanced computed tomography of head and neck
3) No treatment before the CT scan
4) Treatment with radiation and / or chemoradiation only.
Excluded for following reason
1) One-year follow-up not available/ Lost to follow-up (n= XX)
Patients included in the study (n=166)
LRR = XX (XX%)
Non- LRR = XX (XX%)
Training cohort (n=142)
LRR = XX (XX%)
No-LRR = XX (XX%)
Test cohort (n=24)
LRR = XX (XX%)
No-LRR = XX (XX%)

Figure 1: Patient selection and dataset splitting flowchart

All scans were preprocessed in accordance with the Image Biomarker Standardisation Initiative (IBSI) guidelines. Images were resampled to $1 \times 1 \times 1$ mm^3 isotropic voxels using B-spline interpolation and discretized with a fixed bin width of 25 Hounsfield Units. Radiomic features were extracted using `PyRadiomics v3.1.0`, yielding a total of 105 original features per patient. These included first-order intensity statistics (e.g., mean, variance, skewness, kurtosis), shape descriptors (e.g., volume, surface area, sphericity, maximum 2D/3D diameters), and texture features derived from grey-level co-occurrence (GLCM), run-length (GLRLM), size-zone (GLSZM), dependence (GLDM), and neighbouring grey-tone difference (NGTDM) matrices. No wavelet or Laplacian-of-Gaussian filtering was applied to maintain interpretability and reproducibility.

To address the high dimensionality of radiomics data and the risk of model overfitting due to the limited sample size ($p \gg n$), we implemented seven feature selection methods: `SelectKBest`, `LASSO`, and five population-based metaheuristic algorithms [Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), and Simulated Annealing (SA)]. LASSO was formulated as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \qquad (1)$$

where X is the feature matrix, y the recurrence labels, and λ the penalty parameter. Metaheuristics were adapted to binary feature selection, where each candidate subset $S \subseteq \{1, \ldots, p\}$ was evaluated using a fitness function defined by classifier performance:

$$Fitness(S) = AUC(f(X_S), y) \qquad (2)$$

with X_S denoting the submatrix of selected features and f the classifier. To improve stability, we implemented a hybrid pipeline in which Bootstrap-LASSO was first applied across 1000 resampled training sets to identify consistently selected features; the resulting reduced pool was then refined by one of the metaheuristics.

The dataset was partitioned into 80% training and 20% test sets with stratified sampling to preserve class balance. Stratified 5-fold cross-validation within the training set was used for both feature selection and hyperparameter tuning to prevent information leakage. For selected pipelines, bootstrap resampling ($n = 1000$) was used to estimate 95% confidence intervals for ROC AUC. Final model performance was reported using ROC AUC and accuracy.

4 Results

A total of 44 machine learning models were constructed by combining six classifiers with seven feature selection methods, along with two hybrid pipelines using Bootstrap-LASSO followed by metaheuristic optimization. All models were evaluated on an 80/20 stratified split, with ROC AUC and accuracy as the primary metrics.

4.1 Radiomics Features

Table 1 summarizes the performance of classifiers when trained solely on radiomic features. Models were evaluated across all seven feature selection techniques, with results reported as ROC-AUC (Train / Validation / Test) along with 95% confidence intervals.

4.2 Clinical Features

Table 2 summarizes the performance of classifiers when trained solely on clinical features. The same feature selection methods were applied, and performance metrics were reported consistently as in Table 1.

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Naive Bayes	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Linear SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
RBF SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Decision Tree	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Random Forest	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx

Table 1: Performance comparison of machine learning models using radiomic features. ROC-AUC with 95% CI values is reported for (Train/Validation/Test)

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Naive Bayes	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Linear SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
RBF SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Decision Tree	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Random Forest	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx

Table 2: Performance comparison of machine learning models using clinical features. ROC-AUC with 95% CI values is reported for Train/Validation/Test

4.3 Combined Radiomics and Clinical Features

Table 3 shows results for models trained on combined radiomic and clinical features. Performance was again reported across all classifiers and feature selection methods.

4.4 Hybrid Feature Selection (Radiomics)

To enhance robustness, hybrid pipelines were developed by combining Bootstrap-LASSO filtering with metaheuristic selection. Table 4 reports model performance using radiomic features under this hybrid pipeline setup.

4.5 Hybrid Feature Selection (Radiomics + Clinical)

Finally, Table 5 summarizes the results of the hybrid feature selection pipelines applied to the combined radiomic and clinical feature space.

5 Discussion

This study demonstrates that recurrence prediction in head and neck cancer depends strongly on how features are selected and integrated. Radiomics alone provided informative patterns, but combining them with clinical variables produced more consistent models, showing that both data types add value. Metaheuristic algorithms outperformed conventional selectors, and the hybrid Bootstrap-LASSO approach gave more stable results across resamples. Stability in feature selection is critical, as unstable models cannot be trusted in practice, and this framework directly addresses that challenge.

A key strength of this work is the use of a prospective cohort collected under a standardized imaging protocol between 2020 and 2024, reducing the inconsistencies that often limit radiomics studies. Focusing on patients treated with primary radiation makes the findings directly relevant to clinical decision-making at treatment planning. The analysis remains limited by its single-institution scope, and external testing will be essential before clinical use. Still, the results indicate that well-designed machine learning pipelines can generate reliable risk estimates from data already available in routine care, providing a path toward individualized follow-up in head and neck oncology.

6 Conclusion

This study presented a systematic benchmarking framework for predicting locoregional recurrence in head and neck cancer using a prospective, protocol-driven dataset. By evaluating multiple classifiers and feature selection strategies, and introducing hybrid metaheuristic approaches, we established a modelling pipeline that emphasizes stability, reproducibility, and practical relevance. The combined use of radiomic and clinical features further strengthened performance, underscoring the importance of integrating complementary data sources.

Future work will focus on external validation across multi-institutional cohorts, the development of explainability tools to build clinician trust, and the creation of a user-friendly interface that allows recurrence risk to be estimated directly at treatment planning. These steps will be essential to translate the present findings into decision-support systems that can guide personalized follow-up and improve outcomes for patients with head and neck cancer.

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Naive Bayes	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Linear SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
RBF SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Decision Tree	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Random Forest	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx

Table 3: Performance comparison of machine learning models using combined radiomic and clinical features. ROC-AUC with 95% CI values is reported for Train/Validation/Test

Classifier	LASSO	SelectKBest	PSO	WOA	GWO	GA	SA
Logistic Regression	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Naive Bayes	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Linear SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
RBF SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Decision Tree	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Random Forest	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx

Table 4: Performance comparison of machine learning models using hybrid feature selection pipelines (Bootstrap-LASSO + metaheuristics) with radiomic features. ROC-AUC with 95% CI values is reported for Train / Validation / Test sets

References

Classifier	**LASSO**	**SelectKBest**	**PSO**	**WOA**	**GWO**	**GA**	**SA**
Logistic Regression	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Naive Bayes	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Linear SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
RBF SVM	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Decision Tree	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx
Random Forest	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx	0.xx/0.XX/0.xx

Table 5: Performance comparison of machine learning models using hybrid feature selection pipelines (Bootstrap-LASSO + metaheuristics) with combined radiomics and clinical features. ROC-AUC with 95% CI values is reported for (Train/Validation/Test)