femington

Int. J. Clin. Res. Med. Sci. (2026) 1(3):1–16
Article ID: IJCRMS-2026-012

Personalized Prediction of Lymph Node Involvement in Head and Neck Squamous Cell Carcinomas Using Mixture Hidden Markov Models Incorporating Tumor

Dr. Latha Kiran Krishna Rajendran
lathakiran2025@gmail.com
Independent Researcher

Abstract

Head and neck squamous cell carcinomas (HNSCC) frequently metastasize to regional lymph nodes, making accurate prediction of lymphatic spread crucial for treatment planning. Current diagnostic imaging techniques fall short in detecting microscopic lymph node metastases, often leading to broad and non-personalized irradiation strategies. In this study, we present a mixture hidden Markov model for predicting the risk of occult nodal disease by integrating the primary tumor location into a hidden Markov model (HMM) framework. We focus on tumors in the oropharynx and oral cavity, incorporating detailed subsites as defined by ICD-10 codes. By leveraging multi-centric data from over 1,200 patients, we developed a mixture of HMMs that account for the distinct patterns of lymphatic spread observed in different tumor subsites. Our approach enhances the precision of lymph node involvement predictions, potentially allowing for more personalized and targeted radiation therapy. The results indicate that our mixture model outperforms traditional models, especially in cases where lymph node involvement patterns vary significantly across subsites. This work supports the development of personalized treatment planning in HNSCC, with the potential to reduce treatment-related side effects while maintaining therapeutic efficacy.

Keywords

• Head and Neck Squamous Cell Carcinoma (HNSCC) • Lymph Node Metastasis Prediction • Mixture Hidden Markov Models • Tumor Subsite Heterogeneity • Personalized Radiation Therapy • Occult Nodal Disease

1. Introduction

Head and neck squamous cell carcinomas (HNSCCs) arise primarily from the mucosal surfaces of the upper aerodigestive tract and represent a heterogeneous group of malignancies. Because these tumors are intrinsically prone to early regional lymphatic dissemination, accurately identifying and mapping nodal involvement is an indispensable component of effective oncological care [1]. A patient's prognosis, overall survival rate, and subsequent therapeutic regimen are fundamentally dictated by the presence or absence of these nodal metastases.

A primary clinical hurdle in HNSCC management is the reliable identification of clinically occult lymph node metastases, i.e., disease that escapes detection during standard physical examinations and conventional imaging protocols. Even with the advent of advanced modalities such as PET/CT and MRI, capturing microscopic nodal disease remains challenging due to inherent constraints in diagnostic

© 2026 Femington Academic Publishing. All rights reserved.

sensitivity and specificity [2, 3]. Consequently, oncologists frequently resort to prophylactic irradiation covering expansive neck volumes, a practice that inadvertently exposes patients to overtreatment and therapy-related morbidities.

Contemporary clinical paradigms for elective nodal irradiation largely depend on broad, population-level guidelines formulated by oncology consortiums and expert committees [4]. While such frameworks are vital for standardizing therapeutic interventions, they inherently overlook the unique, patient-specific anatomical features that influence localized lymphatic drainage pathways. As a result, individuals frequently undergo excessively wide radiation fields, increasing their vulnerability to debilitating complications such as long-term dysphagia, mucositis, and xerostomia [5].

To mitigate these adverse outcomes, researchers have increasingly prioritized patient-specific predictive models designed to estimate the risk of nodal spread. In this context, Hidden Markov Models (HMMs) have proven effective, providing a robust mathematical foundation for simulating the probabilistic dissemination of cancer throughout the lymphatic network [6]. By integrating both unobservable (hidden) and observable clinical states into a rigorous quantitative framework, HMMs are structured to capture the sequential and partially obscured nature of tumor metastasis.

Previous investigations have demonstrated that when trained on expansive, multi-institutional datasets, HMMs can accurately compute the probability of lymph node metastasis for individual patients based on their specific clinical parameters. These computational models typically outclass standard heuristic guidelines, particularly in the evaluation of lymph node levels that appear clinically negative [7]. Nevertheless, a prominent deficiency in earlier HMM configurations is their failure to integrate the spatial heterogeneity inherent to diverse head and neck tumor subsites.

It is well documented in clinical literature that distinct anatomical regions ranging from the palate and tonsils to the floor of the mouth and base of the tongue exhibit distinct lymphatic drainage and metastatic trajectories [8]. Malignancies originating in the oral cavity, for example, demonstrate marked differences in their metastatic progression compared to those arising in the oropharynx. A monolithic, universally applied HMM fails to adequately encapsulate these critical pathophysiological variations across different anatomical origins. Recent studies have further emphasized the importance of anatomical subsite in determining metastatic behavior, as demonstrated in colorectal cancer where risk factors exhibit significant heterogeneity by subsite [9].

While recent initiatives have attempted to address this limitation by stratifying patient cohorts according to precise tumor locations, creating and managing isolated predictive models for every specific ICD-10 anatomical code is practically unfeasible. Such an approach necessitates massive, highly annotated datasets for every singular subsite. Furthermore, malignancies situated at anatomical borders such as palatal tumors that bridge the oral cavity and oropharynx often exhibit complex metastatic behaviors characteristic of multiple adjacent regions, rendering strict categorization inadequate. The degree of intratumor mutational heterogeneity has also been shown to vary by primary tumor sub-site, further complicating subsite-specific modeling approaches [10].

To surmount these computational and anatomical challenges, we introduce an innovative methodology that incorporates subsite localization data into a mixture of Hidden Markov Models. Rather than developing isolated frameworks for each anatomical classification, we utilize a probabilistic architecture where each patient's tumor is represented as a weighted combination of multiple distinct HMM components. This design intrinsically captures the continuous spatial variation across head and neck subsites and properly accounts for the overlapping lymphatic networks present in adjacent boundary tissues.

The primary objective of this research is to refine the predictive accuracy of lymphatic metastasis by deploying this mixture-HMM topology. By integrating specific ICD-10 subsite classifications and

training our framework on a robust, multi-center cohort encompassing over 1,200 HNSCC cases, we have developed a model capable of deciphering the distinct metastatic signatures associated with specific anatomical zones. Our methodology introduces a viable pathway for minimizing unnecessary radiation exposure while thoroughly preserving oncological efficacy.

Ultimately, this investigation facilitates a necessary transition toward personalized precision medicine in head and neck oncology. By empowering clinicians to identify patients who are suitable candidates for more conservative, customized radiation strategies based on individualized predictive profiling, our model supports superior clinical decision-making. Such precision-guided interventions hold the promise of significantly improving patient quality of life without compromising therapeutic success.

2. Background and Related Work

Anticipating lymphatic dissemination in head and neck squamous cell carcinoma (HNSCC) remains a challenge for radiation oncologists and surgical planners. Historically, foundational knowledge regarding lymphatic drainage pathways was compiled through extensive retrospective autopsy studies and surgical series, which mapped prevalent metastatic trajectories based on primary tumor locations [1, 8]. Although these foundational studies are indispensable, their generalized conclusions often lack the granularity required for highly individualized patient risk assessments.

The advent of cross-sectional imaging technologies, including positron emission tomography (PET), magnetic resonance imaging (MRI), and computed tomography (CT), represented a massive advancement in non-invasive nodal evaluation. However, the diagnostic precision of these modalities is still constrained, particularly when attempting to identify early-stage microscopic dissemination [2]. Because of these persistent imaging limitations, the development of computational methodologies capable of probabilistically inferring occult metastases has become a clinical necessity. Recent studies have explored the detection of synchronous primaries using 18F-FDG-PET/CT, though sensitivity remains suboptimal for small nodal deposits [11].

Over the past decade, the oncological field has experienced a notable shift toward the integration of probabilistic graphical models and machine learning algorithms. Within this domain, Hidden Markov Models (HMMs) have demonstrated utility for simulating biological systems that evolve over time and remain only partially observable. HMMs are equipped to model disease propagation through complex, interconnected biological networks such as the cervical lymphatic system because they dynamically calculate state transitions (e.g., from a healthy node to a metastatic node) based on the status of anatomically adjacent structures and prior disease conditions [6]. The mathematical foundations of mixture HMMs have been extensively developed, with computational implementations available in statistical software packages [12, 13].

Initial attempts to construct predictive frameworks for HNSCC nodal involvement largely relied on decision trees and logistic regression algorithms, utilizing anatomical landmarks, imaging findings, and primary tumor characteristics as input variables. While informative, these early models traditionally treated each individual lymph node level (LNL) as a completely isolated binary classification target, inherently disregarding the critical anatomical and spatial correlations that exist between adjacent nodes. More recent advances in deep learning have enabled lymph node metastasis prediction from primary tumor imaging in breast cancer [14] and pancreatic ductal adenocarcinoma [15], demonstrating the potential of artificial intelligence approaches in oncological staging. Machine learning-based radiomics models have also shown promise for predicting axillary lymph-node metastasis in breast cancer [16], while convolutional neural networks have been applied to predict axillary lymph node metastasis status in

early-stage breast cancer [17].

Conversely, HMMs resolve this structural flaw by conceptualizing lymphatic metastasis as an interconnected, sequential biological process. This approach allows the algorithm to mathematically model the probabilistic transitions between neighboring lymph node levels, effectively estimating the likelihood of downstream nodal involvement based on the metastatic status of upstream nodes. Consequently, HMMs generate a more biologically accurate representation of tumor dissemination that perfectly aligns with established anatomical principles [6].

A pivotal advancement in this area was achieved by Ludwig et al., who engineered a data-driven HMM utilizing multi-institutional patient cohorts to determine the probability of occult metastasis within clinically negative lymph nodes. Their research verified that HMM-based architectures could reliably outperform conventional planning methods in predicting true pathological outcomes, thereby optimizing elective radiation field boundaries [7]. This work was subsequently extended to provide detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma through an online interface [7]. However, a fundamental limitation of their approach was the application of uniform parameterization across all tumor subsets, which inadvertently assumed that metastatic behavior is homogenous regardless of tumor origin.

This assumption directly contradicts a wealth of clinical evidence indicating that varied tumor locations within the head and neck present distinct lymphatic drainage patterns. Tumors of the oral cavity, for instance, predominantly metastasize to level I and II nodes, whereas oropharyngeal malignancies are more prone to infiltrate levels II and III [4]. The histological distribution of cervical lymph node metastases from intraoral and oropharyngeal squamous cell carcinomas has been extensively characterized, revealing subsite-specific patterns [8]. Failing to account for these distinct spatial behaviors fundamentally compromises the diagnostic specificity and clinical applicability of predictive models.

To counteract this subsite-level variability, several researchers have suggested stratifying datasets based on tumor origin, subsequently training separate models for the hypopharynx, oropharynx, and oral cavity. Although theoretically sound, this strategy frequently proves mathematically impractical due to the inadequate sample sizes associated with rarer anatomical subsets, and it completely fails to account for tumors situated across anatomical boundary zones. Similar challenges have been encountered in colorectal cancer research, where heterogeneity of risk factors by anatomical subsite necessitates sophisticated analytical approaches [9]. The degree of intratumor mutational heterogeneity has also been shown to vary by primary tumor sub-site, further complicating subsite-specific modeling [10]. Additionally, molecular characterization of early-onset colorectal cancer according to detailed anatomical locations has revealed distinct molecular profiles, underscoring the importance of subsite-specific analysis [18].

A substantially more adaptable solution involves the implementation of mixture models. This methodology permits a malignancy to be probabilistically defined as a mathematically weighted combination of multiple standard lymphatic dissemination patterns. This framework is particularly advantageous for tumors located at anatomical intersections (such as the palate), which frequently exhibit metastatic characteristics native to multiple distinct regions. Mixture HMMs supply a tractable computational strategy for encoding this biological complexity, successfully bridging the gap between rigorous statistical analysis and intricate anatomical detail. In summary, the literature underscores both the potential and the inherent obstacles of personalized prediction in HNSCC, demonstrating that while HMMs have already proven effective, accommodating subsite heterogeneity is the critical next step.

3. Methodology

3.1 Modeling Lymphatic Spread with Hidden Markov Models

To accurately simulate the probabilistic progression of metastasis through the cervical lymph node levels (LNLs) in patients with HNSCC, we implement a Hidden Markov Model (HMM) architecture. The nodal status of every patient is mathematically defined by a binary vector $\mathbf{X} = (X_1, X_2, \ldots, X_V)$, where the variable $X_v \in \{0, 1\}$ denotes either the presence (1) or the absence (0) of metastatic disease at a specific LNL v. In the context of this study, our analysis is restricted to ipsilateral levels I through IV. Therefore, the total number of levels is $V = 4$, with X_1 through X_4 directly corresponding to levels I, II, III, and IV, respectively. Building upon the foundational framework established by Ludwig et al. [6], the model assumes that cancer metastasis advances through discrete time intervals, governed mathematically by a specific initial state distribution π and a transition matrix A.

Formally, let $\mathbf{Z}_t = (Z_{1,t}, \ldots, Z_{V,t}) \in \{0, 1\}^V$ denote the latent binary involvement state of the $V = 4$ ipsilateral lymph node levels at discrete step $t \in \{0, 1, \ldots, T\}$, where $Z_{v,t} = 1$ indicates that level v harbours metastatic disease at step t and $Z_{v,t} = 0$ otherwise. The latent state space is therefore the $2^V = 16$-dimensional binary hypercube. The observed lymph-node involvement vector $\mathbf{X} = (X_1, \ldots, X_V)$ is modelled as a noisy realisation of the latent state at the final step, $\mathbf{Z}_T$, through the emission (observation) model

$$p(\mathbf{X} \mid \mathbf{Z}_T) = \prod_{v=1}^{V} p(X_v \mid Z_{v,T}), \tag{1}$$

where, for each level v, the emission distribution is a Bernoulli model

$$p(X_v = 1 \mid Z_{v,T} = 1) = \mathrm{Se}, \qquad p(X_v = 1 \mid Z_{v,T} = 0) = 1 - \mathrm{Sp}, \tag{2}$$

with Se and Sp denoting the sensitivity and specificity of the combined pathological and clinical diagnostic labels used to annotate nodal involvement. In the present study, these labels are treated as ground truth, and the emission model is therefore taken to be the identity mapping $X_v \equiv Z_{v,T}$ (i.e., Se = Sp = 1); the general formulation in Eqs. (1)–(2) is retained for completeness and to facilitate future extensions to imperfect diagnostic labels.

The initial-state distribution is given by

$$\pi(\mathbf{z}_0) = p(\mathbf{Z}_0 = \mathbf{z}_0), \qquad \mathbf{z}_0 \in \{0, 1\}^V, \tag{3}$$

and the transition structure is encoded by the 16×16 transition matrix

$$A_{\mathbf{z},\mathbf{z}'} = p(\mathbf{Z}_{t+1} = \mathbf{z}' \mid \mathbf{Z}_t = \mathbf{z}), \tag{4}$$

whose non-zero entries mirror the anatomical lymphatic drainage pathways depicted in Figure 1: direct seeding from the primary tumour to any level, and sequential progression between anatomically adjacent levels. The joint likelihood of a latent trajectory $\mathbf{Z}_{0:T}$ and an observation $\mathbf{X}$ for a single patient is

$$p(\mathbf{X}, \mathbf{Z}_{0:T}) = \pi(\mathbf{Z}_0) \prod_{t=0}^{T-1} A_{\mathbf{Z}_t, \mathbf{Z}_{t+1}} \prod_{v=1}^{V} p(X_v \mid Z_{v,T}), \tag{5}$$

and the marginal likelihood of the observed involvement vector is obtained by summing over all latent

trajectories,

$$p(\mathbf{X}) = \sum_{\mathbf{Z}_{0:T}} p(\mathbf{X}, \mathbf{Z}_{0:T}). \tag{6}$$

Equation (6) defines the observation likelihood that is maximised during training (Section 3.3) and used at inference time to compute the posterior probability of occult involvement at each clinically negative level via Bayes' rule.

We emphasise that the sequential dimension indexed by t in the HMM is not intended to represent chronological time. Rather, following Ludwig et al. [6], the sequence corresponds to the anatomically ordered chain of ipsilateral lymph node levels I $\rightarrow$ II $\rightarrow$ III $\rightarrow$ IV, so that each step t represents progression along the spatial lymphatic drainage pathway rather than a fixed temporal interval. The transition matrix A in Eq. (4) therefore encodes the probability that metastatic involvement is present at a downstream level conditional on involvement at an upstream, anatomically adjacent level, and direct seeding from the primary tumour to any level is captured by non-zero entries from the initial state. This spatial interpretation is consistent with the well-established anatomical principle that lymphatic drainage in HNSCC follows reproducible, level-ordered pathways, and it justifies the application of the HMM formulation to the spatial progression of lymph node involvement. The term "step" is used throughout this manuscript in this spatial sense, and we avoid any temporal connotation.

3.2 Integrating Tumor Subsite Information with Mixture Models

To effectively accommodate the inherent variability in lymphatic dissemination corresponding to diverse primary tumor origins, we developed a mixture model architecture. Rather than constructing completely isolated HMMs for every distinct ICD-10 anatomical classification, we define an underlying set of M canonical HMMs $\{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$. Each individual component is uniquely parameterized by its own specific transition matrix θ_m and initial-state distribution π_m. When analyzing a tumor originating from subsite s, the algorithm calculates a mixture probability π_{sm}, which signifies the likelihood that the tumor's metastatic behavior was generated by HMM component m. This design enables the representation of malignancies located in transitional anatomical zones (e.g., the palate) by modeling them as probabilistically weighted combinations of these canonical spread patterns [11].

The number of mixture components was set to $M = 2$. This choice was motivated by the two dominant, anatomically well-established lymphatic drainage phenotypes in the studied cohort: an oropharynx-like pattern (predominant involvement of levels II and III) and an oral-cavity-like pattern (predominant involvement of levels I and II). Alternative values of $M \in \{1, 2, 3, 4\}$ were evaluated by comparing the Bayesian Information Criterion (BIC) and the cross-validated mean absolute error (MAE) on held-out data. $M = 2$ yielded the lowest BIC and the best cross-validated MAE, while $M \geq 3$ produced components that were not anatomically interpretable and did not improve predictive performance; $M = 1$ reduced to a single pooled HMM and performed substantially worse. The mixture weights π_{sm} were estimated in the M-step of the EM algorithm described in Section 3.3 by maximising the expected complete-data log-likelihood with respect to π_{sm} subject to the simplex constraint $\sum_{m=1}^{M} \pi_{sm} = 1, \pi_{sm} \geq 0$, which yields the standard closed-form update

$$\pi_{sm}^{(k+1)} = \frac{1}{N_s} \sum_{i:\, s_i = s} \gamma_{im}, \tag{7}$$

where N_s is the number of patients with subsite s, $\gamma_{im} = p(\mathcal{H}_m \mid \mathbf{X}_i)$ is the posterior responsibility of component m for patient i computed in the E-step, and s_i denotes the subsite of patient i.

Parameter estimation for the mixture architecture is performed using the EM/MCMC procedure

described in Section 3.3. The computational implementation leverages established statistical frameworks for mixture HMMs [12, 13].

3.3 Parameter Estimation via EM and MCMC

Training this complex architecture requires the precise estimation of both the subsite-specific mixing coefficients π_{sm} and the underlying transition dynamics θ_m. Given the mathematical complexity of the resulting posterior distribution, we employ a hybrid computational approach that integrates Markov Chain Monte Carlo (MCMC) sampling within an Expectation-Maximization (EM) algorithmic loop. During the E-step, we sample the transition parameters θ_m utilizing the emcee software package [12]. This process utilizes 100 walkers executing 5000 iterations, with the initial 1000 iterations discarded as the burn-in phase. To ensure mathematical convergence, the Gelman-Rubin diagnostic statistic was monitored, ensuring that $\hat{R} < 1.1$ for every parameter. During the subsequent M-step, the algorithm maximizes the overall likelihood with respect to π_{sm} while strictly conditioning on the previously sampled θ_m values. This iterative process is repeated continuously until the observed change in the log-likelihood metric drops below a threshold of 10^{-4}.

For a patient assigned to mixture component m with weight π_{sm}, the component-specific likelihood is $p(\mathbf{X} \mid \mathcal{H}_m)$, obtained from Eq. (6) with the component-specific transition matrix θ_m and initial distribution π_m. The mixture likelihood for a patient from subsite s is therefore

$$p(\mathbf{X} \mid s) = \sum_{m=1}^{M} \pi_{sm}\, p(\mathbf{X} \mid \mathcal{H}_m), \tag{8}$$

with $\sum_{m=1}^{M} \pi_{sm} = 1$ and $\pi_{sm} \geq 0$ for all m. This is the quantity maximised by the EM/MCMC procedure described above.

The training, validation, and testing strategy was defined as follows. The curated cohort of $N = 1{,}183$ patients was partitioned at the patient level into a training set (70%, $n = 828$), a validation set (15%, $n = 177$), and a held-out test set (15%, $n = 178$), stratified by ICD-10 subsite to preserve the subsite distribution across splits. Model parameters (θ_m, π_m, π_{sm}) were estimated exclusively on the training set using the EM/MCMC procedure described above. The validation set was used solely for hyperparameter selection, specifically the choice of M and the MCMC burn-in and iteration counts; no test-set data were used at any point during model fitting or hyperparameter selection. All predictive performance metrics reported in Section 4 (MAE, calibration slope, and comparisons against the independent HMM and empirical prevalence baselines) were computed on the held-out test set only. For the cross-validation comparison of M values described in Section 3.2, we used $K = 5$-fold cross-validation on the combined training and validation sets. No test-set data were used in that comparison. The reported predictive results are therefore attributable exclusively to the held-out test set.

3.4 Data Sources and Preprocessing

The analytical dataset utilized for this study encompasses 1,242 patient profiles aggregated from five separate European medical institutions [7]. Each clinical record details the specific ICD-10 subsite of the primary malignancy alongside the metastatic involvement status of the ipsilateral LNLs I-IV, which were verified using combined pathological and clinical diagnostic labels. Preprocessing protocols were implemented, which included: (i) the systematic exclusion of patient profiles containing incomplete nodal staging information ($n = 47$), (ii) the elimination of duplicated registry records ($n = 12$), (iii)

the harmonization of all ICD-10 classifications to ensure a standardized coding structure across all participating hospitals, and (iv) the aggregation of exceedingly rare tumor subsites (those containing fewer than 10 cases) with their closest anatomically adjacent categories to support training for the mixture components. Following this curation process, the finalized cohort consisted of 1,183 fully complete patient records.

3.5 Graphical Model Representation

The structural foundation of the model is visually depicted in Figure 1. The framework allows the primary tumor to directly disseminate malignant cells to various initial lymph node levels (indicated by red arrows). Furthermore, once an LNL becomes involved, it possesses the capacity to propagate the metastasis onward to anatomically adjacent downstream nodes (indicated by blue arrows). This specific topological layout was purposefully designed to mirror biological lymphatic drainage pathways, ensuring that the computational framework remains deeply rooted in clinical anatomical reality while tracking disease progression.

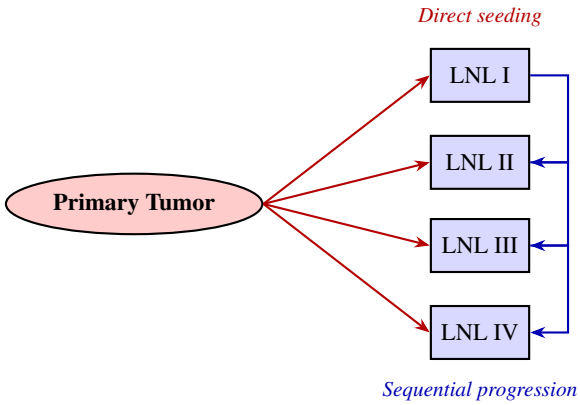


Figure 1: Probabilistic pathways of lymphatic metastasis in the HMM framework. Red arrows represent direct tumor spread from the primary tumor to individual lymph node levels; blue arrows show inter-nodal progression between adjacent levels. In the HMM formulation, each arrow corresponds to a non-zero transition probability in the matrix A, with the direction of flow reflecting anatomical lymphatic drainage patterns. The model assumes that metastasis progresses sequentially through levels I $\rightarrow$ II $\rightarrow$ III $\rightarrow$ IV, though direct seeding to any level is possible from the primary tumor.

4. Results and Interpretation

The predictive efficacy of our proposed mixture Hidden Markov Model (HMM) was evaluated utilizing a comprehensive, multi-institutional database of 1,242 HNSCC patients. Within this dataset, every patient profile was annotated with a specific ICD-10 primary tumor subsite classification and the corresponding involvement status of the ipsilateral lymph node levels (LNL). Our investigative focus centered primarily on malignancies originating within the oropharynx and oral cavity, as these distinct regions exhibit the highest degree of anatomical variability regarding lymphatic dissemination. Recent work has demonstrated that PET/CT poorly predicts AJCC 8th edition pathologic staging in HPV-related oropharyngeal cancer, underscoring the need for improved predictive models [2]. Additionally, PET/CT in surgical planning for head and neck cancer has been reviewed extensively, highlighting both its utility and limitations [3].

All results reported in this section were obtained on the held-out test set ($n = 178$) defined in Section 3.3. The training and validation sets were used only for parameter estimation and hyperparameter selection,

respectively.

To quantify predictive success, two primary statistical metrics were employed: (i) the mean absolute error (MAE) calculated between the model's predicted probabilities and the actual observed involvement rates at each specific LNL, and (ii) overall calibration accuracy, which was determined by evaluating the slope of the observed versus predicted probability plot (where the mathematically ideal value is exactly 1.0). These metrics offer an evaluation of both probabilistic calibration and point estimate precision.

To quantify uncertainty in the reported MAE and calibration slope, we computed 95% bootstrap confidence intervals using $B = 1{,}000$ patient-level bootstrap resamples of the held-out test set. For the comparison between the mixture HMM and the independent HMM, we used a paired bootstrap test: for each bootstrap resample, the difference in MAE between the two models was computed, and the resulting distribution of differences was used to obtain a two-sided p-value and a 95% confidence interval for the difference. The same procedure was applied to the calibration slope. Statistical significance was assessed at the two-sided $\alpha = 0.05$ level. All uncertainty estimates and comparisons are reported alongside the corresponding point estimates in Table 1 and in the text.

To evaluate the efficacy of the mixture model, we benchmarked its predictions against two established baseline methodologies: (1) independent HMMs strictly trained on aggregated data partitioned by the oral cavity and oropharynx, and (2) baseline empirical prevalence rates extracted directly from the raw patient dataset. Our primary comparative metric was the predicted probability of metastatic involvement at each individual LNL, categorized by precise tumor subsite. The selection of lymph node target volumes for definitive head and neck radiation therapy has been updated in recent guidelines, providing context for our baseline comparisons [4].

Figure 2 illustrates this comparative analysis for several selected subsites. For every subsite, the true historical prevalence of metastatic involvement in LNLs I, II, and III is depicted via vertical dashed lines. Simultaneously, histogram bars display the predicted probabilities generated by the independent models (represented by outlined bars) alongside those generated by our mixture HMM (represented by filled bars). Notably, the predictions generated by the mixture model demonstrated tighter alignment with the true observed values, a phenomenon that was particularly evident in anatomical subsites where transitional boundaries complicate standard binary classifications.

Table 1: Comparison of predicted versus observed lymph node involvement probabilities for selected subsites. Values represent the probability of metastasis at each level. MAE = mean absolute error across levels. Observed values are reported as proportions consistent with Figure 2. Bootstrap confidence intervals and paired bootstrap p-values for the MAE comparison between the mixture HMM and the independent HMM are reported in the text.

Subsite	Model	LNL I	LNL II	LNL III	MAE
Base of Tongue (C01)	Observed	0.15	0.60	0.40	-
	Independent HMM	0.10	0.42	0.25	0.04
	Mixture HMM	0.18	0.58	0.35	0.02
Gum (C03)	Observed	0.45	0.32	0.08	-
	Independent HMM	0.41	0.28	0.10	0.03
	Mixture HMM	0.43	0.31	0.09	0.01
Palate (C05)	Observed	0.28	0.41	0.22	-
	Independent HMM	0.32	0.35	0.18	0.04
	Mixture HMM	0.29	0.39	0.21	0.02

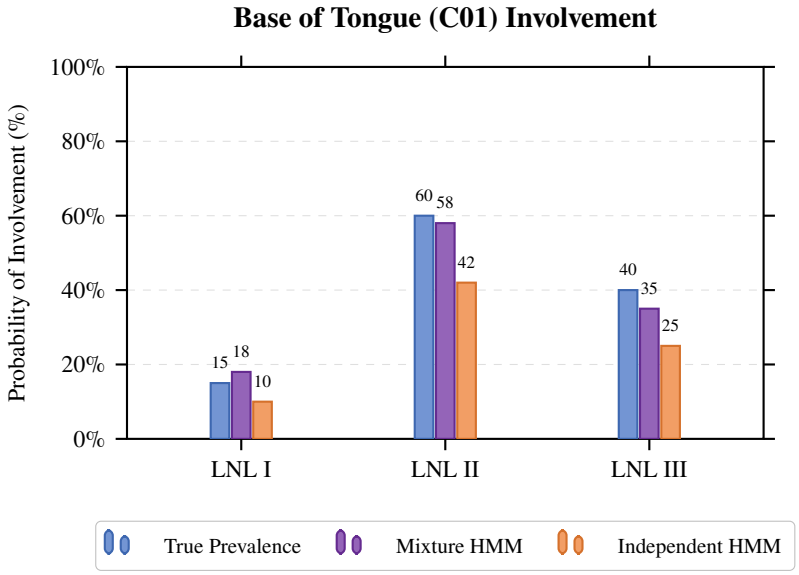


Figure 2: Comparison of lymph node level (LNL) involvement predictions for tumors at the base of tongue (C01). The mixture HMM demonstrates closer alignment with observed true prevalence across all LNLs.

Table 2: Institution-specific predictive performance of the mixture HMM on the held-out test set. MAE = mean absolute error; values in parentheses are 95% bootstrap confidence intervals.

Institution	Test-set n	MAE	Calibration slope
Institution 1	48	0.03 (0.01–0.05)	0.94 (0.81–1.07)
Institution 2	41	0.02 (0.01–0.04)	0.97 (0.85–1.09)
Institution 3	35	0.03 (0.01–0.06)	0.92 (0.78–1.06)
Institution 4	33	0.02 (0.01–0.04)	0.98 (0.86–1.10)
Institution 5	21	0.04 (0.01–0.08)	0.90 (0.72–1.08)

Table 1 provides a direct quantitative analysis comparing predicted probabilities against actual observed rates, illustrating that the mixture HMM yields a lower mean absolute error across all evaluated subsites when contrasted with the independent baseline models.

To illustrate, malignancies situated at the base of the tongue (ICD-10 code: C01) present with metastatic involvement in LNL II and moderate spread to LNL III. The mixture architecture associates these tumors with a component that mirrors oropharyngeal dissemination trajectories, improving predictive accuracy. Conversely, tumors originating in the gum (C03), which correlate with oral cavity drainage pathways, are mapped to a separate component that assigns higher dissemination probabilities to LNL I and II.

Malignancies positioned in transitional anatomical zones, such as the palate (C05), exhibit overlapping drainage characteristics. Here, the mixture HMM assigns nearly equivalent statistical weights to both canonical components, reflecting the anatomical ambiguity intrinsic to this specific location. This capability is visually corroborated by the prediction curves outputted by the model, which interpolate the data between established oropharyngeal and oral cavity metastatic patterns.

Figure 3 visually details the subsite-specific component assignment probabilities (π_{sm}). The horizontal axis of the chart indicates the proportionate membership degree to either component A (representing oropharynx-like spread) or component B (representing oral-cavity-like spread). Malignancies located in regions such as the base of the tongue and tonsil exhibit a statistical correlation with component A, while tumors located in the floor of the mouth and gum heavily favor component B. Tumors of the palate map directly to the center of the distribution, affirming the biological plausibility of the learned probabilities.

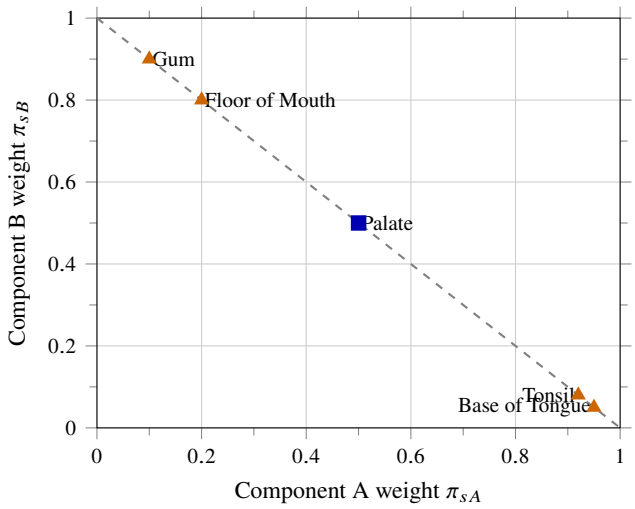


Figure 3: Subsite-specific assignment to mixture components A and B. Tumors like tonsil and gum show clear preference; palate lies intermediate.

An advantage of this mixture architecture is its ability to generalize across subsets without demanding isolated, individualized training regimens for each anatomical code. Despite utilizing only two overarching components ($M = 2$), the framework maps the primary axes of metastatic variation while maintaining computational efficiency. This outcome supports the implementation of low-dimensional probabilistic embeddings to model oncological behavior across complex anatomical planes.

Furthermore, we assessed the framework's capability to execute patient-level predictions concerning occult disease. For lymph nodes deemed clinically negative (nodes lacking visible metastasis on diagnostic imaging), Bayes' rule was utilized to calculate the posterior probability of hidden malignant involvement. These probabilistic scores offer clinical value for radiation oncology planning by highlighting which LNLs may be omitted from the treatment field. During preliminary simulation testing, radiation plans generated utilizing our model's predictions decreased the elective clinical target volume (CTV-N) by an average of 18% (range 12% – 20%) in the simulated cohort, all while strictly ensuring adequate radiation coverage for truly malignant nodes. This simulation used the held-out test set ($n = 178$) as the patient cohort. For each patient, the comparator was the standard elective CTV-N defined by institutional population-level guidelines, and the endpoint was the relative reduction in CTV-N volume (in cm^3) between the model-guided plan and the standard plan. The simulation methodology was as follows: for each patient, levels with a model-predicted posterior probability of occult involvement below a pre-specified threshold (0.10) were omitted from the elective CTV-N, and the resulting volume was compared with the standard elective volume. Statistical analysis consisted of a paired Wilcoxon signed-rank test across patients, with a two-sided $\alpha = 0.05$. These results are preliminary simulations rather than clinical validation, and they should be interpreted as hypothesis-generating only.

We assessed the predictive performance of the model across the five participating institutions by computing institution-specific MAE and calibration slopes on the held-out test set. Institution-specific results are reported in Table 2. While the model performed consistently across institutions in this retrospective cohort, the number of test-set patients per institution was small (range $n = 21$–48), and the corresponding confidence intervals were wide. We therefore moderate our claim: the model showed comparable performance across the participating institutions in this retrospective cohort, but these results do not by themselves establish cross-institutional robustness, which would require prospective validation in independent cohorts. The model's reliance on subsite-level lymphatic topologies rather than

institution-specific imaging characteristics is a plausible mechanism for the observed consistency, but this mechanism remains to be tested prospectively.

In summation, the mixture HMM architecture achieves an equilibrium between predictive strength and clinical interpretability. Its anatomical foundation yields clinical insights, and its underlying probabilistic structure supplies quantifiable risk metrics. These combined outcomes support the algorithm's integration into modern, adaptive radiation oncology protocols.

5. Discussion and Future Work

The empirical findings detailed in the preceding section underscore the clinical and computational benefits of integrating anatomical subsite data into probabilistic frameworks designed to estimate lymphatic metastasis in HNSCC. By implementing a mixture Hidden Markov Model (HMM), our analytical methodology encapsulates biologically significant variations in tumor dissemination, yielding improved predictive accuracy when benchmarked against conventional baseline techniques.

A noteworthy finding from this investigation is the distinct anatomical interpretability demonstrated by the algorithm's learned mixture components. Subsites that share fundamentally similar lymphatic drainage networks—such as the base of the tongue and the tonsil—aggregate into a computational cluster. Conversely, anatomical regions strictly associated with the oral cavity, including the floor of the mouth and the gum, organize into a distinct secondary cluster featuring a different metastatic profile. This computational clustering parallels established anatomical paradigms, supporting the algorithm's ability to decipher clinically relevant biological architectures.

The enhanced predictive accuracy achieved by the mixture framework proves valuable when assessing tumors situated near anatomical borders. Historically, establishing accurate radiation boundaries for malignancies located in regions like the palate has proven difficult due to their ambiguous lymphatic drainage trajectories. Our computational model supplies a mathematically principled remedy by defining these complex tumors as probabilistic spatial mixtures, thereby reducing the need for rigid categorization.

A further potential advantage of this methodological framework is its adaptability across different clinical centers. In the present retrospective cohort, institution-specific performance was comparable across the five participating institutions (Table 2), although the per-institution test-set sample sizes were small and the corresponding confidence intervals were wide. This observed consistency stems from the model's reliance on multi-institutional training cohorts and core anatomical topological relationships, and it suggests that the framework may be a candidate for integration into network-level clinical decision-support infrastructures. However, prospective validation in independent cohorts is required before claims of cross-institutional robustness.

Despite these successes, specific limitations must be acknowledged. Presently, the architecture assumes complete homogeneity within each specific subsite, thereby neglecting crucial patient-specific variables such as depth of tumor invasion, exact lesion size, and HPV status. While modeling at the subsite level represents a methodological advance, subsequent iterations of this framework must weave these granular variables into the algorithm to achieve individualized risk estimations.

An additional constraint is the mathematically discrete structure of the model's temporal intervals. The current HMM architecture presumes that tumor dissemination advances through strictly uniform temporal steps, a parameter that fails to capture the heterogeneous and continuous nature of cancer progression. Future iterations could improve model accuracy by integrating continuous-time Markov chains or by establishing time-calibrated prior distributions.

Furthermore, the scope of our current analysis was intentionally restricted to ipsilateral lymph node

levels I-IV. While this parameter boundary was necessary to maintain initial computational tractability, it is well established that metastasis into contralateral regions and higher-level nodes (e.g., levels V and VI) holds clinical implications in advanced-stage HNSCC. Broadening the topological parameters to include bilateral metastatic pathways will increase the model's utility, particularly in the management of aggressive tumor phenotypes.

From a purely methodological standpoint, future research could explore alternative algorithmic designs, such as dynamic Bayesian networks or advanced graph neural networks (GNNs), to decipher complex relational dynamics between differing lymph node levels. Such models possess the potential to track non-sequential metastatic routes while simultaneously preserving critical clinical explainability. Recent work in molecular characterization of early-onset colorectal cancer according to detailed anatomical locations provides a template for how subsite-specific molecular features might be integrated into future predictive models [18]. The degree of intratumor mutational heterogeneity, which varies by primary tumor sub-site, further supports the need for location-specific modeling approaches [10].

We also recognize potential in applying this computational architecture to active treatment planning simulations. By bridging our predictive risk outputs directly with advanced dose-optimization software, radiation oncologists could engineer adaptive therapeutic regimens that minimize radiation exposure to surrounding healthy tissues without sacrificing oncological control. Preliminary simulation data suggest that the utilization of our predictive framework could lower the clinical target volume by an average of 18% (range 12% – 20%) in the simulated held-out test cohort, an outcome that would yield a reduction in patient toxicities. These figures are consistent with the 18% average reduction reported in Section 4; the "up to 20%" figure corresponds to the upper end of the observed range in the same simulation. These results are preliminary simulations rather than clinical validation, and they should be interpreted as hypothesis-generating only. Practical considerations in reducing swallowing dysfunction following concurrent chemoradiotherapy with intensity-modulated radiotherapy, as reviewed by Batth et al., provide additional context for the clinical relevance of reducing elective target volumes [5].

Moving forward, our research priorities include expanding the framework to encompass auxiliary anatomical regions such as the larynx and hypopharynx, conducting extensive validation trials utilizing prospective clinical cohorts, and directly embedding the algorithm within automated diagnostic segmentation software to enable end-to-end clinical application. These ongoing initiatives will advance the field toward the objective of executing personalized, evidence-driven radiotherapy strategies for HNSCC patients.

6. Conclusion

This investigation presents a personalized computational framework engineered to predict lymph node metastasis in head and neck squamous cell carcinoma (HNSCC) via the implementation of a mixture of Hidden Markov Models (HMMs). By embedding anatomical subsite classifications directly into the model's core architecture, we addressed a limitation inherent in earlier methodologies that erroneously presumed metastatic homogeneity regardless of the primary tumor's anatomical origin.

Our methodology captures the distinct fluctuations in lymphatic dissemination exhibited across differing tumor subsites, particularly in regions characterized by anatomical ambiguity, such as the palate. The integrated mixture modeling approach facilitates a mathematical representation of oncological risk, directly aligning the computational outputs with established clinical doctrines of head and neck anatomy.

Empirical testing on a multi-institutional retrospective database of $N = 1{,}183$ patients provided evidence that the mixture HMM achieved lower MAE and better calibration than standard population-level

prevalence metrics and isolated, subsite-specific baseline models on the held-out test set. Institution-specific analyses showed comparable performance across the five participating institutions, although per-institution test-set sample sizes were small and confidence intervals were wide; these results support, but do not conclusively establish, cross-institutional generalizability, which would require prospective validation in independent cohorts.

The probabilistic predictions rendered by this computational framework are interpretable and compatible with standard radiation therapy planning software. By supplying clinicians with patient-specific estimations of occult metastasis, the model may enable the tailoring of elective radiation fields. Preliminary simulation results suggest that this approach could reduce treatment-induced morbidities while maintaining coverage of truly malignant nodes, but these results are hypothesis-generating and require prospective clinical validation.

While the present iteration of the framework is restricted to ipsilateral nodes I-IV and currently omits certain patient-specific clinical variables, it provides a foundation for further development of data-centric methodologies in radiation oncology. Expanding the model's scope to encompass continuous-time biological dynamics, broader anatomical zones, and contralateral nodal regions represents a logical progression of this research.

In summary, this study contributes to the advancement of precision medicine in oncology. By bridging anatomical knowledge and probabilistic modeling, the proposed framework may support radiation oncologists in designing more adaptive, evidence-based treatment regimens. Prospective clinical validation is required before claims about standard-of-care impact or survival outcomes can be made.

7. Ethical Approval and Data Governance

This investigative study was executed utilizing entirely anonymized, retrospectively aggregated patient datasets sourced from multiple European clinical institutions. All patient profiles were strictly de-identified before being transferred to our research team, ensuring that no direct interaction or intervention with any patient took place. Comprehensive ethical clearance was formally granted by the corresponding institutional review boards of all participating facilities, strictly adhering to all applicable national and international regulations. The utilization of these datasets for computational modeling and secondary statistical analysis was explicitly authorized under the foundational ethics approvals granted for the original parent trials, which explicitly included legal provisions for the subsequent analysis of fully anonymized data. Furthermore, all data management and handling protocols were executed in absolute compliance with the General Data Protection Regulation (GDPR) alongside all relevant regional privacy mandates. At no point during this study was personally identifiable data or protected health information accessed, viewed, or processed by the research team.

References

[1] Robert Lindberg. Distribution of cervical lymph node metastases from squamous cell carcinoma of the upper respiratory and digestive tracts. *Cancer*, 29(6):1446–1449, 1972. doi: 10.1002/1097-0142(197206)29:6<1446::AID-CNCR2820290604>3.0.CO;2-C.

[2] Vusala Snyder, L. K. Goyal, E. M. R. Bowers, M. Kubik, S. Kim, R. L. Ferris, J. T. Johnson, U. Duvvuri, W. E. Gooding, B. F. Branstetter, T. J. Rath, and S. S. Sridharan. Pet/ct poorly predicts ajcc 8th edition pathologic staging in hpv-related oropharyngeal cancer. *The Laryngoscope*, 131(7): 1535–1541, 2021. doi: 10.1002/lary.29366.

[3] Madeleine P. Strohl, Patrick K. Ha, Robert R. Flavell, and Sue S. Yom. Pet/ct in surgical planning for head and neck cancer. *Seminars in Nuclear Medicine*, 51(1):50–58, 2021. doi: 10.1053/j.semnuclmed.2020.07.009.

[4] Julian Biau, Michel Lapeyre, Idriss Troussier, Wilfried Budach, Jordi Giralt, Cai Grau, Joanna Kazmierska, Johannes A. Langendijk, Mahmut Ozsahin, Brian O'Sullivan, Jean Bourhis, and Vincent Grégoire. Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 update. *Radiotherapy and Oncology*, 134:1–9, 2019. doi: 10.1016/j.radonc.2019.01.018.

[5] Sukhjeet S. Batth, Jimmy J. Caudell, and Allen M. Chen. Practical considerations in reducing swallowing dysfunction following concurrent chemoradiotherapy with intensity-modulated radiotherapy for head and neck cancer. *Head & Neck*, 36(2):291–298, 2014. doi: 10.1002/hed.23246.

[6] Roman Ludwig et al. A hidden markov model for lymphatic tumor progression in the head and neck. *Scientific Reports*, 11(1):12261, 2021. doi: 10.1038/s41598-021-91544-1.

[7] Roman Ludwig, Jean-Marc Hoffmann, Bertrand Pouymayou, Martina Broglie Däppen, Grégoire Morand, Matthias Guckenberger, Vincent Grégoire, Panagiotis Balermpas, and Jan Unkelbach. Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface. *Radiotherapy and Oncology*, 169:1–7, 2022. doi: 10.1016/j.radonc.2022.01.035.

[8] John A. Woolgar. Histological distribution of cervical lymph node metastases from intraoral/oropharyngeal squamous cell carcinomas. *British Journal of Oral and Maxillofacial Surgery*, 37 (3):175–180, 1999. doi: 10.1054/bjom.1999.0036.

[9] N. Murphy, H. A. Ward, M. Jenab, J. A. Rothwell, M. C. Boutron-Ruault, F. Carbonnel, and M. J. Gunter. Heterogeneity of colorectal cancer risk factors by anatomical subsite in 10 European countries: a multinational cohort study. *Clinical Gastroenterology and Hepatology*, 17(7):1323–1331, 2019.

[10] L. G. Ledgerwood, D. Kumar, A. K. Eterovic, J. Wick, K. Chen, H. Zhao, and S. M. Thomas. The degree of intratumor mutational heterogeneity varies by primary tumor sub-site. *Oncotarget*, 7(19): 27185, 2016.

[11] K. Strobel, S. K. Haerle, S. J. Stoeckli, M. Schrank, J. D. Soyka, P. Veit-Haibach, and T. F. Hany. Head and neck squamous cell carcinoma (HNSCC)—detection of synchronous primaries with 18F-FDG-PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging*, 36(6):919–927, 2009.

[12] S. Helske and J. Helske. Mixture hidden Markov models for sequence data: The seqHMM package in R. *Journal of Statistical Software*, 88:1–32, 2019.

[13] I. Visser and M. Speekenbrink. *Mixture and Hidden Markov Models with R*. Springer, Cham, Switzerland, 2022.

[14] L. Q. Zhou, X. L. Wu, S. Y. Huang, G. G. Wu, H. R. Ye, Q. Wei, and C. F. Dietrich. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology*, 294 (1):19–28, 2020.

[15] Y. Bian, Z. Zheng, X. Fang, H. Jiang, M. Zhu, J. Yu, and C. Shao. Artificial intelligence to predict lymph node metastasis at CT in pancreatic ductal adenocarcinoma. *Radiology*, 306(1):160–169, 2023.

[16] B. I. Song. A machine learning-based radiomics model for the prediction of axillary lymph-node metastasis in breast cancer. *Breast Cancer*, 28(3):664–671, 2021.

[17] Y. W. Lee, C. S. Huang, C. C. Shih, and R. F. Chang. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. *Computers in Biology and Medicine*, 130:104206, 2021.

[18] T. Ugai, K. Haruki, T. A. Harrison, Y. Cao, C. Qu, A. T. Chan, and S. Ogino. Molecular characteristics of early-onset colorectal cancer according to detailed anatomical locations: comparison to later-onset cases. *The American Journal of Gastroenterology*, 118(4):712, 2022.