Titre: Title:	Risk Stratification for Adaptive Radiotherapy in Head and Neck Cancers Using Deep Learning Approaches
Auteur: Author:	Gautier Henique
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Henique, G. (2025). Risk Stratification for Adaptive Radiotherapy in Head and Neck Cancers Using Deep Learning Approaches [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/66824/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/66824/
Directeurs de recherche: Advisors:	Samuel Kadoury, & Houda Bahig
Programme: Program:	Génie biomédical

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Risk Stratification for Adaptive Radiotherapy in Head and Neck Cancers Using Deep Learning Approaches

GAUTIER HENIQUE
Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie biomédical

Juillet 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Risk Stratification for Adaptive Radiotherapy in Head and Neck Cancers Using Deep Learning Approaches

présenté par **Gautier HENIQUE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Julien COHEN-ADAD, président
Samuel KADOURY, membre et directeur de recherche
Houda BAHIG, membre et codirectrice de recherche
Arthur LALONDE, membre

DIDICTION

To my close ones

ACKNOWLEDGEMENTS

I would first like to thank my research supervisors, Professor Samuel Kadoury and Dr. Houda Bahig. Their constant support, wise advice, and sustained assistance made this work possible.

I would also like to thank all the colleagues and collaborators who contributed to this work, whether directly or indirectly. I wish you all success and happiness in your future endeavors.

Thank you to my colleagues at the MediCAL laboratory, who fostered a positive work environment leading to collaboration and mutual uplifting. Thank you Emilie Ouraou, Ralph Saber, David Grajales Lopera, Aloys Portafaix, William Le, Charmin Bang, Gabriel Dayan, Shima Sagordi, Amine El Foraici, and Augustin Rochereau.

Finally, thank you to the various interns I had the chance to mentor or supervise. Thank you to Marsien Boade, Eddy Manih, Coralie Brodeur, and Karim Bekthi.

RÉSUMÉ

Les cancers de la tête et du cou sont un enjeu majeur de santé publique, dont le traitement requiert une précision particulière en raison de la présence d'organes vitaux à proximité de la zone tumorale. La radiothérapie est l'un des traitements de référence pour cette pathologie, et a su bénéficier des nombreuses avancées en imagerie biomédicale pour permettre des traitements et un suivi de plus en plus adaptés aux spécificités de chaque patient. Toutefois, la complexité de ces données d'imagerie rend leur exploitation aussi difficile que prometteuse, alimentant une demande persistante de la part des cliniciens pour des outils robustes d'adaptation des traitements. Afin d'optimiser les chances de succès thérapeutique tout en limitant les effets indésirables, l'utilisation de biomarqueurs permettant de stratifier les risques spécifiques de complications liées au traitement constitue un axe de recherche prioritaire.

L'apprentissage profond s'impose aujourd'hui comme une voie prometteuse pour répondre à cet enjeu, en facilitant l'analyse de données d'imagerie hautement dimensionnelles et l'identification de biomarqueurs subtils, encore imperceptibles à l'œil humain. Néanmoins, pour que ces modèles puissent s'intégrer durablement dans les flux cliniques, ils doivent non seulement être performants, mais également robustes, interprétables et alignés sur les décisions thérapeutiques. Cette maîtrise s'inscrit dans ce contexte, en étudiant l'usage de l'intelligence artificielle pour prédire les complications et les issues thérapeutiques en radiothérapie des cancers de la tête et du cou, en intégrant les particularités anatomiques propres à chaque patient.

Deux axes méthodologiques principaux ont été développés, chacun correspondant à une phase distincte du parcours de soins.

Le premier axe porte sur la prédiction des toxicités radio-induites survenant au cours du traitement, avec pour objectif de modéliser l'interaction entre les doses effectivement délivrées et les déformations anatomiques visibles au fil du traitement sur les images CBCT (Cone-Beam Computed Tomography). L'étude a été menée sur une cohorte de 1012 patients traités par radiothérapie conformationnelle avec intention curative. Les scans CBCT, acquis quotidiennement, ont permis de calculer des champs de déformation vectorielle (DVF) à l'aide de VoxelMorph, un réseau de neurones convolutifs spécialisé dans le recalage d'images. Ces DVF ont ensuite servi à générer des cartes dose-déformation, intégrant de manière spatio-temporelle les doses cumulées et leur interaction avec les changements anatomiques individuels.

Un pipeline d'apprentissage profond en vision par ordinateur a été déployé pour extraire des biomarqueurs dosimétriques dynamiques, prédictifs de trois complications majeures : la dépendance à une sonde nasogastrique, la radionécrose et l'hospitalisation. Nous avons utilisé un modèle multi-branche combinant un réseau convolutif traitant les cartes dose-déformation et une branche clinique reposant sur un perceptron multi-couche. Les modèles obtenus ont atteint des précisions respectives de 74,1% (±2,9), 75,4% (±5,6) et 61,1% (±7,3), soulignant la pertinence de l'analyse conjointe dose-anatomie dans une optique d'adaptation thérapeutique en temps réel.

Le second axe méthodologique explore la valeur pronostique de l'extension ganglionnaire visible en imagerie (iENE) chez les patients atteints de cancer oropharyngé HPV-positif, à partir d'un sous-groupe de 397 patients issus de la cohorte précédente. Un pipeline complet et automatisé a été développé pour segmenter les iENE sur les scans CT de planification, en s'appuyant à la fois sur des réseaux convolutifs comme nnU-Net et sur des approches récentes basées sur des modèles fondamentaux avec segmentation guidée par prompt.

Des caractéristiques radiomiques classiques ainsi que des descripteurs profonds extraits des régions ganglionnaires et tumorales ont été analysés à l'aide de modèles d'apprentissage automatique, notamment XGBoost, enrichis par des techniques d'augmentation et d'équilibrage de données.

Par ailleurs, une architecture originale a été proposée, nommée AMO-ENE (Attention-based Multimodal Extranodal Extension), intégrant des mécanismes d'attention multi-têtes pour fusionner les représentations extraites de différentes modalités et sous-régions anatomiques, tout en évaluant leur importance relative. Cette architecture permet de pondérer de manière adaptative les informations hétérogènes et d'améliorer la stratification des risques oncologiques. AMO-ENE a ainsi été utilisé pour la prédiction du pronostic oncologique selon deux approches complémentaires : (1) une classification binaire de la réponse au traitement à deux ans, et (2) une modélisation fine de la survie via une régression multi-tâches temporelle. Le module de segmentation a atteint un coefficient de Dice de 74,4% (±24,9), tandis que le modèle de classification des iENE a obtenu une AUC de 79,92% (±2,51). Intégrés dans des modèles pronostiques, ces biomarqueurs ont permis de prédire avec précision plusieurs issues cliniques clés : métastases à distance à deux ans (AUC : 88,2% ±4,8), survie sans récidive (AUC : 78,1% ±8,6) et survie globale (AUC : 72,6% ±9,6). L'indice de concordance (C-index) pour la prédiction des métastases atteint 83,3% (±6,5), confirmant la valeur clinique de l'iENE et le potentiel des outils de décision fondés sur la radiomique.

En conclusion, cette maîtrise démontre la faisabilité et l'intérêt de l'intelligence artificielle pour améliorer la stratification des risques en radiothérapie, à travers l'exploitation conjointe d’imageries longitudinales et de planification. Les pipelines développés constituent une base concrète pour l’élaboration de protocoles thérapeutiques personnalisés. Les perspectives ouvertes incluent des stratégies de radiothérapie adaptative, une gestion proactive des toxicités, ainsi qu’un ajustement personnalisé des doses en fonction du risque individuel.

ABSTRACT

Head and neck cancers represent a complex and high-stakes clinical setting, where intricate anatomy, the proximity of critical organs, and intra-treatment anatomical changes during treatment pose significant challenges to both planning and delivery. Biomedical imaging plays a crucial role in guiding radiotherapy, benefiting from advances in high-quality scanner technologies to provide a wealth of multimodal information. However, the extraction of clinically actionable insights from this data remains limited, and the development of a robust dose adaptation framework continues to be a key demand from clinicians. Deep learning has emerged as a promising approach to address this gap by enabling the complex analysis of high-dimensional imaging data and the extraction of deep biomarkers not currently perceivable through conventional means. Yet, the successful integration of such models into clinical workflows requires more than predictive accuracy, it also demands robustness, interpretability, and a direct connection to clinical decision-making, especially in radiotherapy, where preserving patient health and quality of life is paramount.

This Masters thesis investigates the use of deep learning and medical imaging to predict adverse treatment outcomes and oncological prognosis in head and neck cancer, with a focus on incorporating patient-specific anatomical data to support personalized radiotherapy planning. Two primary methodological inquiries were pursued, each targeting distinct but complementary phases of treatment.

The first line of inquiry focused on in-treatment toxicity prediction by modeling the interaction between delivered dose distributions and anatomical deformations captured in longitudinal cone-beam computed tomography (CBCT) scans. A cohort of 1012 patients treated with curative-intent intensity-modulated radiotherapy (IMRT) was analyzed. CBCT scans acquired throughout the treatment course were used to compute deformation vector fields, capturing spatial anatomical changes over time. VoxelMorph, a convolutional neural network-based image registration framework, was employed to perform deformable registration between daily CBCTs.

Dose-deformation maps were constructed by combining daily dose distributions with the deformation fields representing anatomical changes from pre-treatment baselines, enabling the modeling of patient-specific spatio-temporal dose accumulation. A deep learning pipeline was then developed to extract dose-aware imaging biomarkers predictive of three clinically significant complications: nasogastric (NG) tube dependency, radionecrosis, and hospitalization. We employed a multi-branch toxicity prediction model, with a vision branch integrating the proposed dose-deformations maps through a convolutional neural residual network, and clinical information with a multi-layer perceptron branch. Multi-modal fusion was performed through late fusion concatenation of latent spaces. The resulting models achieved accuracies of 74.1% (±2.9) for NG tube dependency, 75.4% (±5.6) for radionecrosis, and 61.1% (±7.3) for hospitalization. These findings underscore the prognostic value of incorporating dynamic anatomical and dosimetric information, supporting the use of such models for real-time treatment adaptation and proactive toxicity management.

The second methodological component investigated the prognostic utility of imaging-visible extranodal extension (iENE) in HPV-positive oropharyngeal cancer patients. A sub-cohort of 397 patients from the original dataset was used. A fully automated pipeline was developed to segment extranodal extensions on pre-treatment CT scans, leveraging both convolutional neural networks (e.g., nnU-Net) and recent prompt-driven foundation models for segmentation.

Radiomic and foundation model-derived features were extracted from both nodal and tumoral regions and evaluated using statistical machine learning models for iENE grade classification. Among these, the XGBoost algorithm, augmented with data balancing techniques, demonstrated strong class stratification. These features were also integrated using a novel deep learning architecture, AMO-ENE (Attention-based Multimodal Extranodal Extension), which employs multi-head attention to extract and fuse deep representations across spatial subregions. This architecture enabled selective weighting of heterogeneous imaging features, providing enhanced classification of iENE status.

We applied AMO-ENE for oncological outcome prediction using two schemes: (1) binary classification of treatment response at the 2-year follow-up, and (2) full survival curve modeling using multi-task bin regression. The segmentation module achieved a Dice Similarity Coefficient of 74.4% (±24.9), while the iENE classification model attained an AUC of 79.92% (±2.51). Prognostic models incorporating these features demonstrated high predictive performance for key clinical endpoints, including 2-year distant metastasis (AUC: 88.2% ±4.8), disease-free survival (AUC: 78.1% ±8.6), and overall survival (AUC: 72.6% ±9.6). The model also achieved a concordance index (C-index) of 83.3% (±6.5) for metastasis prediction, validating iENE as a clinically relevant biomarker and emphasizing the potential of radiomics-driven, automated decision support tools in precision oncology.

In conclusion, this masters thesis demonstrates the feasibility and potential of deep learning in enhancing the stratification of radiotherapy risks and outcomes through the use of longitudinal and planning CT imaging. The developed pipelines contribute toward clinically actionable tools that could support more individualized treatment protocols. Future

applications include adaptive radiotherapy strategies, improved toxicity management, and risk-guided dose escalation or de-escalation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

HNC	Head and Neck cancer
OPC	Oropharyngeal cancer
HPV	human papillomavirus
OS	Overall survival
DM	Distant metastasis
DFS	Disease free survival
NGT	Feeding tube toxicity
OAR	Organ at risk
ENE	Extra nodal extension
iENE	Imaging visible extra nodal extension
RT	Radiotherapy
ART	Adaptive Radiotherapy
IMRT	Intensity modulated Radiotherapy
VMAT	Volumetric Modulated Arc Therapy
DVH	Dose Volume Histogram
CT	Computed tomography
CBCT	Cone-beam computed tomography
DVF	Deformation vector field
JDET	Jacobian Determinant
DA-JDET	Dose aware jacobian determinant
DL	Deep Learning
MLP	Multi Layer perceptron
CNN	Convolutional neural network
VIT	Vision transformer
AMO-ENE	Attention-based Multi-Omics fusion model
MHSA	Multi-head self attention
MRMR	Minimum Redundancy and Maximum Relevance
MTLR	Multi-task logistic regression
KM	Kaplan Meier
SHAP	SHapley Additive exPlanations
SR	Simulated reader
DSC	Dice Score
IoU	Intersection over union

C-index	Concordance index
AUC	Area under the curve

CHAPTER 1 INTRODUCTION

Each year, over 8,000 Canadians are diagnosed with Head and Neck Cancer (HNC) [21], contributing to 890,000 cases globally [22] and making it the seventh most prevalent type of cancer. Well-established risk factors include tobacco and alcohol consumption, with the Human Papillomavirus (HPV) emerging as a leading cause, particularly in oropharyngeal cancer. It's estimated that HPV will become the primary contributor to HNC by 2030, as its increasing prevalence in younger populations raises concerns about associated cancer risk. Incidences and mortality rates vary significantly by socio-demographic characteristics; for instance, in Canada, 72% of cases present in male patients, while 32% of fatalities occur in women. The burden created by HNC can be seen at multiple levels, with the obvious public health crisis present but also cost associated to treatment, burden created on hospital resources and overall poorer quality of life.

Standard of care to treat HNC typically involves a combination of therapies, aiming to control tumor proliferation while ensuring the preservation of healthy crucial organs. Treatments are highly individualized, adapted to the patient's specific condition, staging, and preferences, as the characteristics of tumor progression may necessitate different approaches. These often include radiation therapy, paired with chemotherapy and immunotherapy as systemic control agents, and surgery for the resection of affected tissues.

The head and neck region is a unique anatomical locality for cancer due to its high vascularity and the presence of numerous vital organs, posing significant treatment challenges. The intricate interactions between tumoral tissues and surrounding organs which are at risk of metastatic involvement or collateral damage from treatment create a complex anatomical landscape. This complexity holds potentially significant prognostic value for the early estimation of oncological outcomes through survival analysis.

While radiotherapy has proven effective and contributed to improved patient outcomes, it is also associated with side effects, commonly referred to as toxicities. These toxicities significantly diminish patients' quality of life and lengthen recovery processes. Controlling these toxicities is crucial as they can jeopardize the successful completion of therapies, negatively impacting oncological outcomes, and severely affect patients even after hospital discharge.

In the wake of significant computational and imaging progress, biomedical imaging has become an integral part of treatment planning and evaluation. Computed Tomography (CT) CT is routinely employed to visualize tumor volume and nearby organs at risk (OARs). These scans serve as the standard for segmenting various treatment targets, including the primary lesion and any metastatic involvement. The resulting annotations then serve as references for processing radiotherapy plans, guiding the prescription of appropriate radiation doses to specific structures according to clinical guidelines.

Beyond planning imagery, significant interest is now directed towards intra-treatment scans, such as Cone Beam Computed Tomography (CBCT) CBCT. These fast, lower-resolution scans enable longitudinal patient follow-up, offering crucial insights into disease progression and response to treatment. Such insights can warrant plan adaptation, including dose escalation or de-escalation, depending on favorable or unfavorable tumor control, respectively. However, CBCT scans suffer from low spatial resolution and poor soft tissue contrast, making human interpretation of these dynamic anatomical changes challenging.

Recent advances in computing, scanning technologies, and Deep Learning (DL) now open opportunities to explore subtle biomarkers currently imperceptible to the human eye. By leveraging large biobanks as statistical databases, DL models, with their millions of parameters, can be trained to identify complex patterns, demonstrating excellent performance and heralding the introduction of powerful tools for refining treatments and alleviating medical burden. However, effectively integrating and interpreting interactions across multiple heterogeneous input sources such as different scan sequences or composite organ statistics remains a significant challenge for even advanced statistical and deep learning models, limiting comprehensive multimodal analysis.

In this Master's thesis we aim to further develop the knowledge of HNC specific biomarkers as to predict and characterize the entire treatment procedure and patient response. By leveraging the various treatment accessible bio-imageries, treatment planning characteristics as well as patient clinical characteristics, we aim to propose frameworks leading to a fully adaptive radiotherapy.

We first propose a methodology in order to assess toxicity involvement during radiotherapy by utilising the longitudinal CBCT scans and pairing the anatomical deformations with dosimetry. We also evaluate the role of ganglion extra nodal extensions in oncological outcomes on CT imagery, as a significative organ at risk.

1.1 Plan of this Masters Thesis

In the following chapter 2, we will set the necessary background information by presenting clinical points relevant to this work. This section will be stratified by concept, firstly introducing the description of the studied cancer type followed by the relevant biomedical imagery used in the following experiments.

Chapter 3 introduces the necessary review of literature in order to grasp the methodologies employed in this work. We stratified this section the topics covered from our works.

Chapter 4 consists of a presentation of the methodology put in place in the two papers presented in chapters 5 and 6. Each of these chapters will encompassed works as initially presented, and adapted to the format of this thesis.

Chapter 7 covers a general discussion covering the presented papers, by summarizing them and investigation their limitations.y proposing further advances and avenues for improvement, we also propose new development paths that may inspire the reader to pursue the works proposed.

The last chapter of this thesis will conclude on the presented work by providing insights on both the positive achievements of this program and the difficulties encountered along the way.

CHAPTER 2 BACKGROUND

2.1 Head and neck cancer anatomy

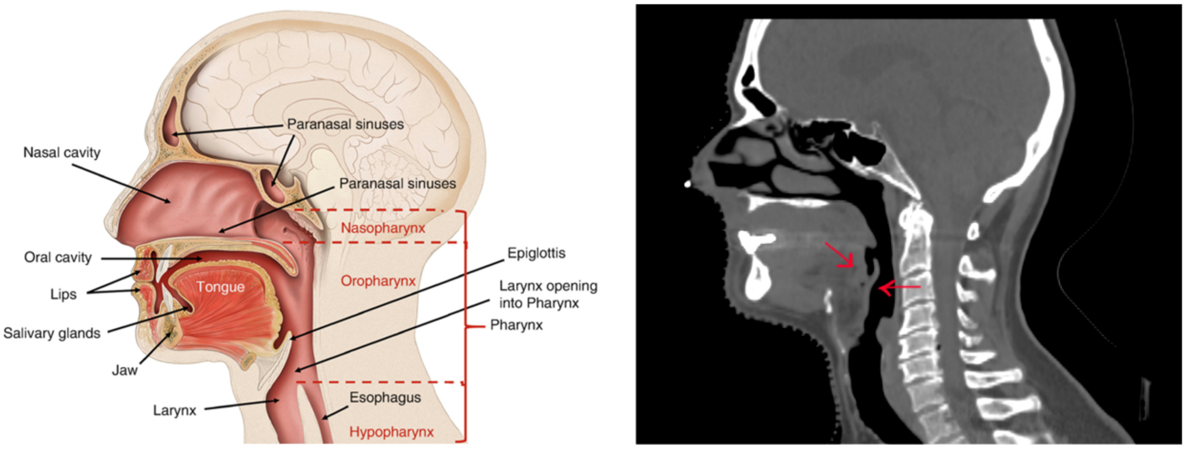


Figure 2.1 (Left) Diagram of the head and neck anatomy in sagittal view (adapted from [3]). (Right) Sagittal view of a computed tomography scan of a patient with head and neck cancer. Red arrows point to the affected tumoral region in the oropharynx

Head and neck cancer encompasses a group of malignancies that affect the oral region and the upper portions of the respiratory and digestive tracts. These cancers typically originate from squamous cells lining the mucosal surfaces of the head and neck, including the oral cavity, oropharynx, larynx, nasal cavity, sinuses, and, in some cases, the upper esophagus. Major risk factors include tobacco use, alcohol consumption, and infection with high-risk strains of the Human Papillomavirus, particularly in the development of oropharyngeal cancer.

Diagnosis typically involves a combination of physical examination, imaging studies (e.g., CT, MRI, PET), and biopsy procedures to confirm cancer type and extent. From diagnosis, treatment strategies are selected depending on cancer stage and location, and may involve surgery, radiation therapy, chemotherapy, or immunotherapy.

2.1.1 Anatomical Considerations and Extranodal Extension in Head and Neck Cancer

Due to the dense anatomical landscape, high vascularity, and proximity of vital structures, treating HNC presents unique challenges in preserving function while achieving oncological control. During treatment, a wide array of Organs at Risk (OARs) must be considered, including the salivary glands, optic systems, auditory organs, voice and feeding structures, and the brain, among others. As damage to these structures during radiation or surgical interventions can lead to significant impacts on core physiological functions such as breathing, swallowing, and speech, HNC treatment can profoundly impact the patient's quality of life. Early intervention and functional preservation are therefore key treatment goals.

A particularly important prognostic feature in HNC is extranodal extension (ENE), a condition in which metastatic tumor cells within lymph nodes breach the nodal capsule and invade surrounding soft tissue. ENE is commonly associated with more aggressive tumor behavior and an elevated risk of both local recurrence and distant metastasis. ENE is recognized as a high-risk feature in HPV-negative cases and is included in staging guidelines from the American Joint Committee on Cancer (AJCC). For HPV-positive cases, its prognostic value remains under investigation.

2.2 Current standard of care for radiotherapy treatment of HNC

Radiation therapy is one of the primary treatment options for HNC, with the goal of administering targeted levels of radiation, most often through X-rays, to damage the genetic material of malignant cells and interrupt their proliferation. Delivering safe levels of radiation doses to specific organs while accurately targeting the tumor volume necessitates a combination of meticulous pre-treatment planning and pre-delivery alignment using fast scanners.

In the treatment room, external radiotherapy involves utilizing a dose delivery system, such as a linear accelerator (LINAC), to produce and target energy beams at selected localizations. Figure 2.2 presents a diagram of a LINAC system and a picture representing a treatment session. While the patient lies flat on the treatment bed and their head is secured with a thermoplastic mask, the LINAC administers dosage to targeted tissues.

2.2.1 Intensity modulated external radiotherapy

Intensity Modulated Radiotherapy (IMRT) is a specialized radiotherapy procedure, made possible by the advanced imaging available in the clinical setting. It aims at specifically targeting determined regions with selected levels of radiation, in order to deliver the desired clinical prescription locally, allowing for more precise tumor targeting and the sparing of healthy tissue. This technology, however, comes at the cost of significantly enhanced planning and delivery complexity, especially when compared to 3D conformal radiotherapy.

In addition to IMRT, Volumetric Modulated Arc Therapy (VMAT) consists of a similar

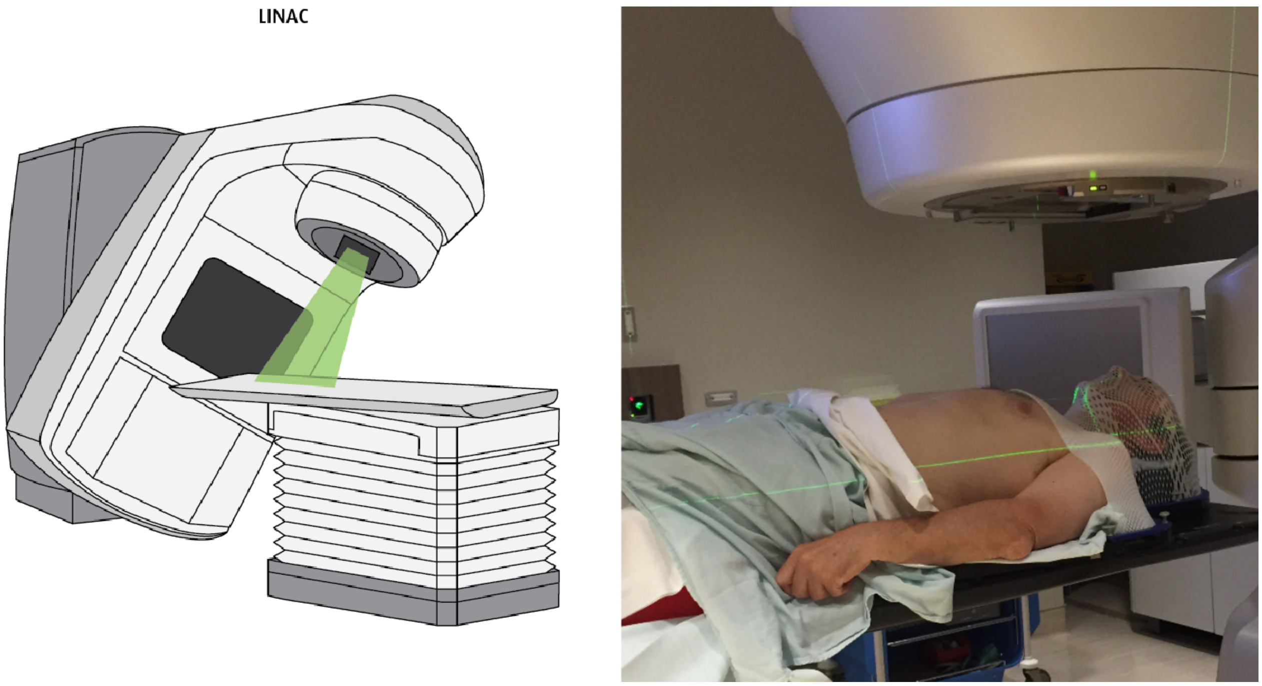


Figure 2.2 (Left) Diagram of a LINAC system. (Right) Photography of a patient with mask on under a VMAT system. Taken from [4]. Photography licensed by Michael Goodyear, CC BY-SA 4.0

procedure but involves rotating the delivery system around the patient, allowing for an accelerated treatment while preserving delivery accuracy.

For curative-intent HNC treatment, a typical course consists of 30 to 35 treatment days, where a fractionated portion of the total dose is administered to the patient daily. Before each treatment session, a CBCT scan is acquired, effectively allowing for longitudinal monitoring of anatomical changes in the patient. This information can confirm tumor shrinkage or necessitate re-adaptation of the dosimetry.

2.2.2 IMRT Planning

Intensity-Modulated Radiation Therapy Planning

Planning of external beam radiotherapy is a critical process led by dosimetrists and radiation oncologists. The goal is to create a customized dose distribution that conforms to the patient's unique anatomy and tumor characteristics, delivered through a linear accelerator (LINAC). This planning phase is essential for maximizing tumor control while minimizing radiation exposure to healthy tissues.

For curative-intent treatments, this process begins with the delineation of three hierarchical target volumes:

- **Gross Tumor Volume (GTV):** The visible or palpable extent of the tumor, as identified through imaging or clinical examination.
- **Clinical Target Volume (CTV):** An expansion beyond the GTV that encompasses areas suspected of harboring microscopic disease, accounting for potential subclinical tumor spread. The CTV aims to ensure oncologic control beyond visible disease boundaries. **Planning Target Volume (PTV):** A geometric margin added to the CTV to account for uncertainties in patient positioning, internal motion (e.g., swallowing or breathing), and daily setup variation. The PTV ensures that the prescribed dose is effectively delivered to the entire CTV throughout treatment.

Prescribed radiation doses vary based on tumor stage and treatment protocol. In definitive radiotherapy for head and neck cancer, the CTV typically receives between 66–70 Gy in 2 Gy fractions, while elective nodal regions may receive 50–60 Gy. Organs at Risk (OARs) are simultaneously contoured and assigned dose constraints to limit toxicity [23].

From the described target doses and annotations of structures, a dosimetry plan will be derived, a process referred to as inverse planning. Dose distribution plans are generated using advanced computational algorithms such as the Monte Carlo method or Acuros XB [23], which simulate radiation transport with high accuracy. Figure 2.3 proposes a visual depiction of IMRT planning on a real case.

In order to process these simulations, objectives of coverage (how completely the target volume is irradiated to the target dose), conformity to the prescribed dose, and homogeneity in the treatment volume are established. These metrics are iteratively optimized through inverse planning, often over several hours, to balance target coverage against OAR sparing [24].

Final plans are always reviewed by the radiation oncologist [24], who may modify contours or priorities based on individual patient characteristics, such as comorbidities, prior treatments, or specific anatomic variations. Additionally, during treatment, significant anatomical changes such as tumor shrinkage or patient weight loss may necessitate adaptive re-planning [24] to maintain treatment accuracy and minimize toxicity.

2.2.3 Radiation therapy toxicities

Radiotherapy treatment may present side effects, denoted as toxicities. While they are a testament to the effectiveness of the procedure, potentially grave complications for the pa-

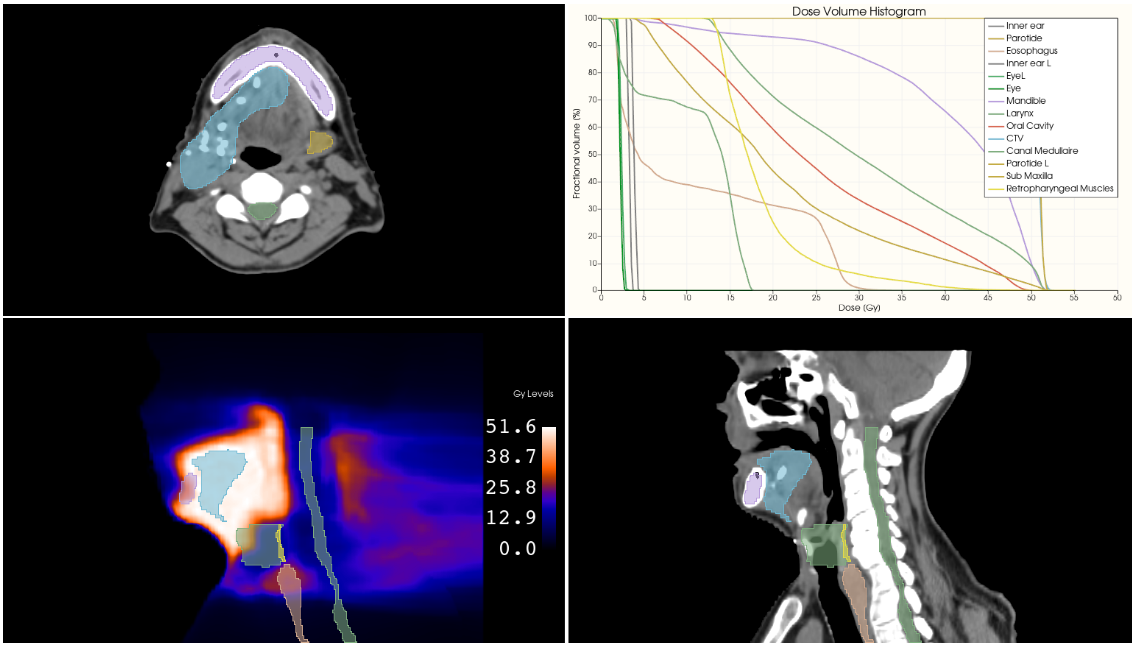


Figure 2.3 Figures taken from IMRT radiotherapy planning. A CT scan with segmentation maps of relevant clinical OARs and CTV target can be observed. The figure on the bottom left displays a dosimetry map of radiation intensities. The top right figure represents the dose-volume histogram of clinical targets and OARs, mapping the proportion of organs receiving specific dose levels.

tient's well-being may arise. These toxicities can be acute (occurring during or shortly after treatment) or late (developing months or years later). They often manifest as mucositis (inflammation of mucous membranes), xerostomia (dry mouth), dysphagia (difficulty swallowing), dermatitis (skin inflammation), or long-term damage to salivary glands, nerves, or bone structures.

Several factors can make dose plans more likely to cause toxicities in head and neck IMRT. One key factor is the proximity of the tumor to organs at risk (OARs), which makes it difficult to spare healthy tissues while delivering an effective dose to the tumor. Anatomical complexity or difficult access to certain tumor sites also increases planning difficulty. Additionally, larger tumors or those that infiltrate surrounding structures often require broader radiation fields, exposing more normal tissue. Finally, patient-specific anatomy and changes over the course of treatment, like weight loss or tumor shrinkage, can also affect dose delivery and toxicity risk.

Controlling toxicities during the course of radiotherapy is a main priority and demand from clinical oncologists, as their early detection can benefit both the quality of life for patients during recovery and overall treatment success. Re-planning may help mitigate severe side effects, although the complexity and time requirements, as stated previously, make this approach situational. Multidisciplinary interventions, including nutritional support, speech and swallowing therapy, and pain management, are often necessary to ensure treatment tolerance and recovery.

2.3 Role of Biomedical Imaging in HNC Treatment

2.3.1 Computed Tomography

Physical Principle

Computed Tomography (CT) is a widely used imaging technology in clinical care, offering good anatomical detail at a relatively low cost compared to other modalities such as Magnetic Resonance Imaging (MRI).

CT scans are a type of X-ray imaging modality in which cross-sectional slices of the body are generated by analyzing the response of anatomical structures to transmitted X-ray photons. The resulting images are organized into a grid, with each two-dimensional slice consisting of pixels and the full volume represented as voxels. The spacing between slices is adjustable, depending on the spatial resolution requirements.

Specific clinical needs influence scan parameters, such as the selected field of view (the anatomical region to be imaged), the use of contrast agents, and scanner-specific settings related to beam filtration and reconstruction algorithms.

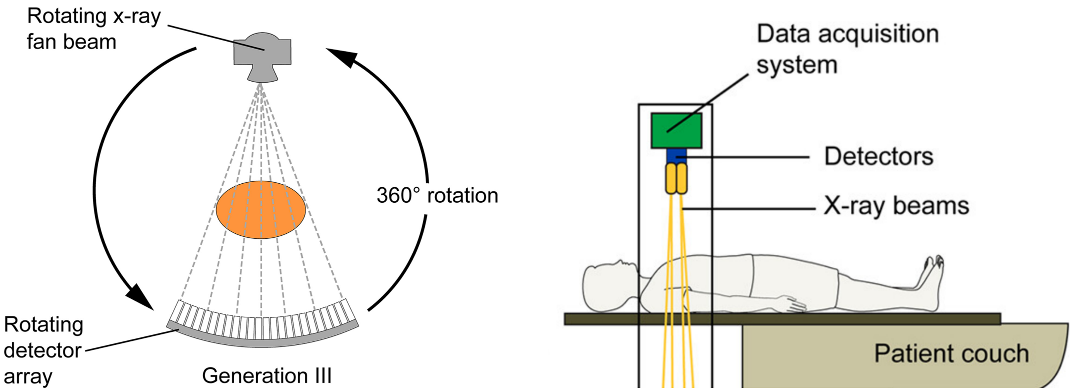


Figure 2.4 Diagram of a 3rd generation CT scanner. A rotating source of X-rays emits in a fan distribution, which is captured by a paired rotating detector array. Licensed and adapted from [5].

Each voxel in a CT scan represents the attenuation of X-ray photons by the tissue, which is influenced by the tissue's density and atomic number. This attenuation is quantified using standardized Hounsfield Units (HU), which compare the attenuation coefficient of a given material to that of water, as shown in Equation 2.3.1, where μ denotes the attenuation coefficient:

$$HU = 100 \times \frac{\mu_{material} - \mu_{water}}{\mu_{water}}$$

(2.1)

Typical HU values are -1000 for air voxels, 0 for water, above 1000 for all dense bone structures. A window of -50 - 300 HU can be selected to visualize soft tissue, however this range depends on the parameters of the scanner and if contrast agents are used.

Because tumors often present as abnormal masses with distinct densities, CT is a crucial tool in cancer detection. In head and neck cancer treatment, CT is the primary modality for planning, as it enables accurate localization of the tumor and delineation of critical structures.

Role in Organ Segmentation

CT imaging serves as the standard imagery in HNC due to its relatively low cost compared to magnetic resonance imaging, and its ability to provide good resolution and contrast for differentiating anatomical structures and malignancies. Contrast-enhanced CT scans can further improve the visibility of vascular structures and tumor margins.

CT imaging plays a vital role in organ segmentation during radiotherapy planning. It allows clinicians to identify and delineate tumors and Organs at Risk (OARs), which is essential for the aforementioned optimization of radiation dose delivery.

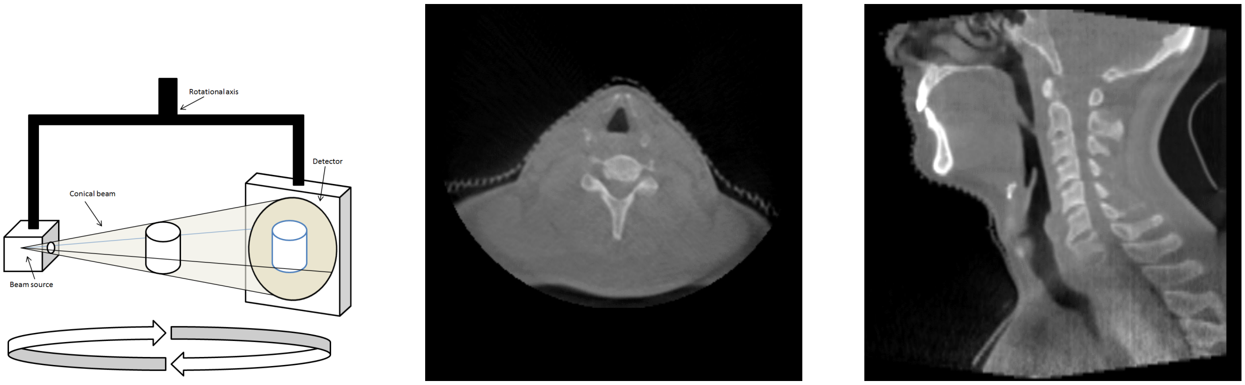
2.3.2 Cone Beam Computed Tomography

Cone Beam Computed Tomography (CBCT) is a pivotal imaging modality in the context of modern IMRT. It is designed for rapid, low-dose volumetric imaging, offering the capability to acquire three-dimensional anatomical information immediately before or during treatment. Despite its lower resolution and poorer soft tissue contrast compared to conventional fan-beam CT, CBCT's accessibility, speed, and real-time imaging capacity make it indispensable in Image-Guided Radiation Therapy (IGRT).

Imaging Principles and Limitations

CBCT operates by acquiring a series of two-dimensional X-ray projections using a cone-shaped beam and a flat-panel detector as the gantry rotates 360 degrees around the patient. The collected projections are reconstructed into a volumetric dataset, typically within one minute. The advantage of this rapid acquisition is reduced patient discomfort and minimal motion artifacts. However, this fast scanning sequence presents increased scatter radiation and image noise, limited soft tissue contrast, inaccurate Hounsfield Unit (HU) values due to inconsistent X-ray spectra, and reconstruction artifacts and potential distortions. Figure 2.5 presents a visual depiction of the physical principle of CBCT capture and resulting visual examples.

Because of these limitations, CBCTs are not typically used for dose calculation or primary diagnostic assessment but remains highly effective for anatomical localization. The inaccurate HU values, in particular, prevent direct use of CBCT for dosimetric calculations, making the development of methods to derive accurate synthetic CTs from CBCT a crucial area of ongoing research.



(a) CBCT Principle (b) Axial View (c) Sagittal View

Figure 2.5 Diagram of CBCT applied to head and neck imaging. Diagram (a) courtesy of Aron Saar. The cone beam nature of the capture can be seen through the conic artifact across the contour of all views.

Role in Longitudinal Treatment Evaluation

In IMRT, particularly for head and neck cancers, CBCT serves two main purposes: daily patient setup verification and adaptive treatment planning.

Daily Setup Verification As explained in the IMRT planning section, the anatomy of the head and neck region requires millimetric precision in dose delivery. CBCT imaging enables clinicians to verify and adjust patient positioning just before each fraction of radiation is delivered. This allows for the detection and correction of translational and rotational setup deviations, alignment of the patient based on internal anatomy, and monitoring of tumor position when using immobilization devices.

To do so, rigid registration between planning CT and daily CBCT ensures that high-dose regions are aligned with the tumor while sparing adjacent OARs. This daily verification process is essential for reducing setup errors, compensating for patient motion, and ensuring inter-fraction consistency.

Monitoring Anatomical Changes and Adaptive Radiotherapy (ART) Over the 6–7 weeks of typical radiotherapy, patients may experience considerable anatomical changes—such as tumor regression, weight loss, and muscle atrophy—which can alter dose distribution and compromise treatment quality.

CBCT provides a means to monitor these changes longitudinally by visualizing gross tumor shrinkage, indicative of treatment response, or by tracking deviations from the original anatomy. While daily rigid registration aligns the overall patient position, larger anatomical changes and internal organ motion over the course of treatment often necessitate more advanced techniques like deformable image registration (DIR) to accurately map the evolving anatomy and its impact on dose distribution. This allows clinicians to modify the treatment in order to restore optimal tumor coverage and maintain organ sparing.

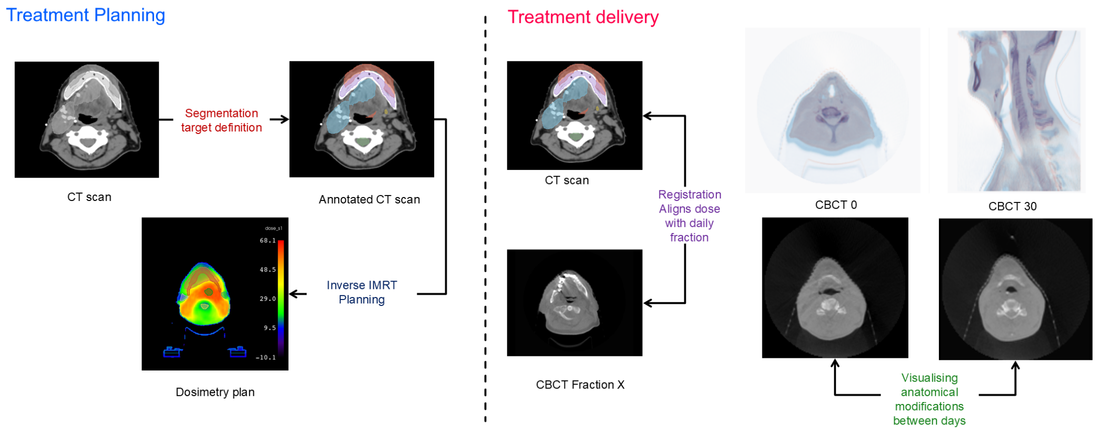


Figure 2.6 Overview of the use of medical imaging and CBCT in IMRT. The left portion represents the contouring and dose planning steps necessary for the creation of a dosimetry plan. The right portion visualizes the use of CBCT images during the course of treatment. Registration between the CT image and any CBCT fraction allows for rigid movement of the dose plan to ensure alignment. Visualizing pairs of CBCTs over time allows for the monitoring of daily anatomical and tumoral changes.

CHAPTER 3 LITERATURE REVIEW

3.1 Notions of Deep Learning

Deep Learning (DL) is a specific subfield of Machine Learning that aims to train statistical algorithms to learn complex, non-linear representations from input data in order to solve specific tasks. Training is achieved using artificial neurons as the fundamental computational units, which are stacked into layers—creating the depth characteristic of deep models.

Given a layer of neurons, the vectorized form of the layer's output can be expressed as:

$$\mathbf{h}^{(l)} = \varphi\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{3.1}$$

where: $\mathbf{h}^{(l-1)}$ is the input vector to layer l (i.e., the activations from the previous layer) ; $\mathbf{W}^{(l)}$ is the weight matrix for layer l, ; $\mathbf{b}^{(l)}$ is the bias vector ; $\varphi(\cdot)$ is the activation function (e.g., ReLU, sigmoid) ; $\mathbf{h}^{(l)}$ is the output (activation) of layer l.

By combining such artificial neurons into layers and stacking them, the simplest form of a deep learning model is the Multi-Layer Perceptron (MLP) . MLPs are capable of learning deep features from underlying data characteristics. Through this process of combining non-linear representation transformations, low-level underlying features may be extracted to approximate complex functions.

Training a model involves minimizing an objective function by updating weights to reduce prediction error. When a training sample is passed through a deep learning model, all intermediate activation states are cached and later used in backpropagation, the update principle where gradients are propagated backward through the network using the chain rule of calculus.

The core update rule using gradient descent for a parameter θ is:

$$\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta} \tag{3.2}$$

where: $\mathcal{L}$ is the loss function ; η is the learning rate ; $\frac{\partial \mathcal{L}}{\partial \theta}$ is the gradient of the loss with respect to the parameter θ.

The introduction of mini-batch training enabled more robust estimation of gradients by averaging over small subsets of data, improving convergence stability. Further enhancements have come from advanced optimization algorithms (e.g., Adam, RMSprop) and techniques

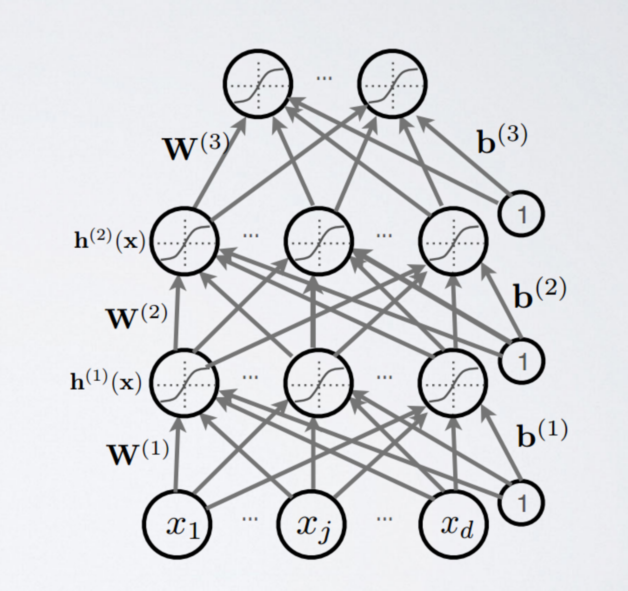


Figure 3.1 A MLP neural network. For the i-th input and l-th layer, x_i is the i-th input at layer 0, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector, and $\mathbf{h}^{(l)}$ the output (activation). Taken from [6].

like learning rate scheduling and regularization (e.g., dropout, L2 regularization) to improve generalization and convergence.

3.1.1 Deep learning powered computer vision

The application of neural networks to the analysis of images is a well-established and popular framework, with many recent advances and significant applications in fields such as medical imaging.

Convolutional Neural Networks: Convolutional Neural Networks (CNNs) , initially introduced by Yann LeCun in the early 1990s with LeNet-5, gained widespread popularity with AlexNet [2]. AlexNet's dominant win in the 2012 ImageNet competition sparked a major resurgence of this principle in deep learning for computer vision.

CNNs apply convolution operations over input images using learnable kernels that slide across spatial dimensions. These filters capture local patterns such as edges, textures, and shapes. To manage computational complexity and extract hierarchical features, CNNs often incorporate downsampling operations—such as max pooling or average pooling—which reduce spatial resolution while increasing the receptive field of subsequent layers, enabling robust feature extraction across scales.

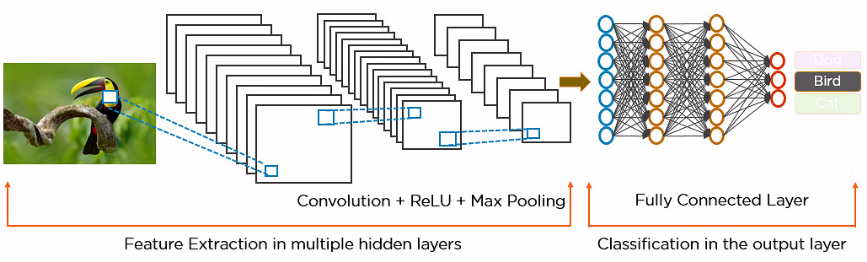

Dog
Bird
Cat
Convolution + ReLU + Max Pooling
Fully Connected Layer
Feature Extraction in multiple hidden layers
Classification in the output layer

Figure 3.2 A CNN neural network applied to imaged to image classification. Convolution layers and pooling operations are stacked together to reduce dimensionality, and a last connected layer is used to obtain the class scores. Taken from [7]

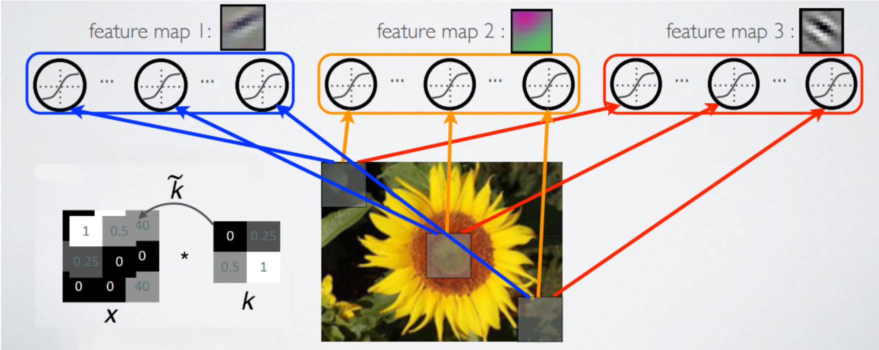


Figure 3.3 Illustration of the convolution principle, each convolution kernel allows the extraction of feature maps. Taken from [6]

CNN architectures build upon the basic convolution operation by increasing depth and introducing innovations to train deeper and more expressive models. A notable milestone is the introduction of the residual block in ResNet [17], which allows for the propagation of gradients through skip connections, facilitating the training of much deeper networks without degradation in performance.

Since ResNet [17], research on CNNs has focused on enhancing pre-training capabilities over large-scale datasets, allowing models to learn general representations that can later be fine-tuned for specific downstream tasks. Architectures such as EfficientNet [25] and ResNeXt [26] exemplify this trend by improving computational efficiency and scaling strategies while maintaining or increasing accuracy. CNNs remain a competitive and resource-efficient choice in computer vision, especially for applications involving limited data, such as those in medical image analysis.

Vision Transformers: Introduced in 2020, Vision Transformers (ViTs) [1] propose an alternative framework for image analysis. Inspired by the success of the attention mechanism in natural language processing, images are treated as sequences of tokenized patches, where their relative importance is assessed using self-attention. Refer to Figure 3.4 for a visual depiction.

Given an input image of size $H \times W \times C$, it is first divided into non-overlapping patches of size $P \times P$, resulting in $N = \frac{HW}{P^2}$ patches. Each patch is flattened and linearly projected into a D-dimensional embedding, forming a sequence $X \in \mathbb{R}^{N \times D}$.

Self-attention is then applied to this sequence to compute interactions between all pairs of tokens. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3.3}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are the query, key, and value matrices computed from the input sequence X, and $W_Q, W_K, W_V \in \mathbb{R}^{D \times d_k}$ are learned projection matrices.

ViT networks have been shown, under the right pre-training conditions [1], to enable more powerful representation learning than traditional CNN architectures. This success confirms the potential of the attention mechanism in capturing long-range dependencies and contextual information across spatial regions in images.

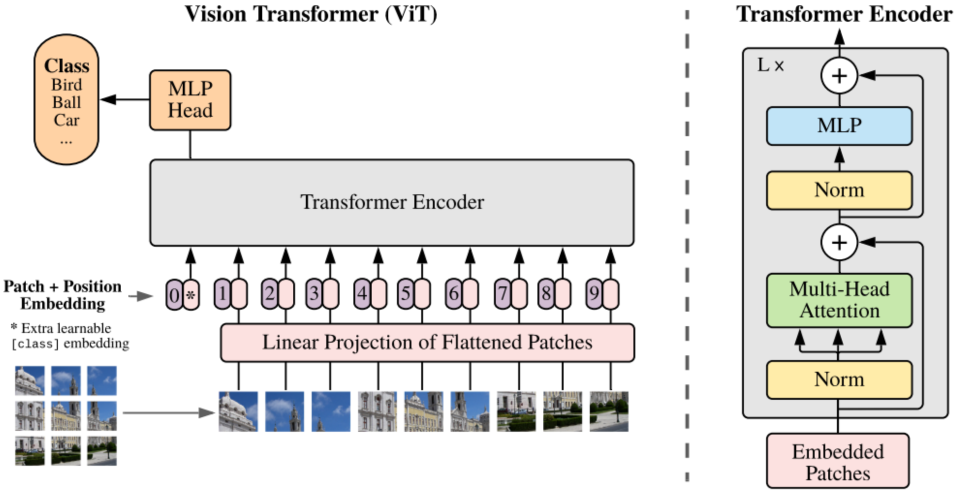


Figure 3.4 The original ViT model, processing images as sequences of 16x16 pixels patches. Taken from [1].

ViT networks, however, do not consistently outperform other architectures across all tasks. In particular, the attention mechanism tends to be less effective at capturing fine-grained local pixel-to-pixel relationships, which are crucial in tasks such as image segmentation. Moreover, ViTs typically require large-scale training datasets with substantial instance variability to achieve strong generalization performance, making them less suitable for smaller datasets.

3.2 Application to Medical Image Classification

mage classification involves assigning a category from a predefined set of possible labels to an input image. This task holds substantial importance in the medical field, where scans, or specific regions within them, can be analysed to extract imaging biomarkers relevant to diagnosis, prognosis, or the characterization of disease-specific features.

Various approaches to feature extraction in medical images have been developed and can often be combined. The following sections provide an overview of prominent methodologies and their application to HNC.

3.2.1 Radiomic Feature Analysis

Radiomics, introduced over the past decade, has emerged as a pivotal technique in medical image analysis for extracting hand-crafted features from regions of interest (ROIs). These features aim to quantitatively describe tissue characteristics captured in imaging modalities.

Radiomic features are categorized into 11 distinct families, encompassing a spectrum of descriptors from basic shape metrics to advanced filtered intensity-based statistics. From a user-defined ROI—such as a segmented lesion or an OAR—these features are computed using standardized extraction parameters. The resulting high-dimensional feature set typically undergoes selection procedures tailored to the specific prediction task.

Due to their hand-crafted nature, simpler radiomic features often retain a high degree of interpretability, making them attractive for use in statistical modeling. However, as feature complexity increases, interpretability tends to decline, and these features may exhibit reduced robustness and generalizability across datasets.

In the context of HNC, radiomics has demonstrated strong potential in capturing relevant phenotypic traits of primary tumors and lymph nodes [27, 28]. These features have been shown to effectively differentiate between aggressive and indolent disease presentations. Additionally, radiomic descriptors have been successfully integrated with dose distribution metrics to improve the prediction of treatment-induced toxicities, such as xerostomia [29].

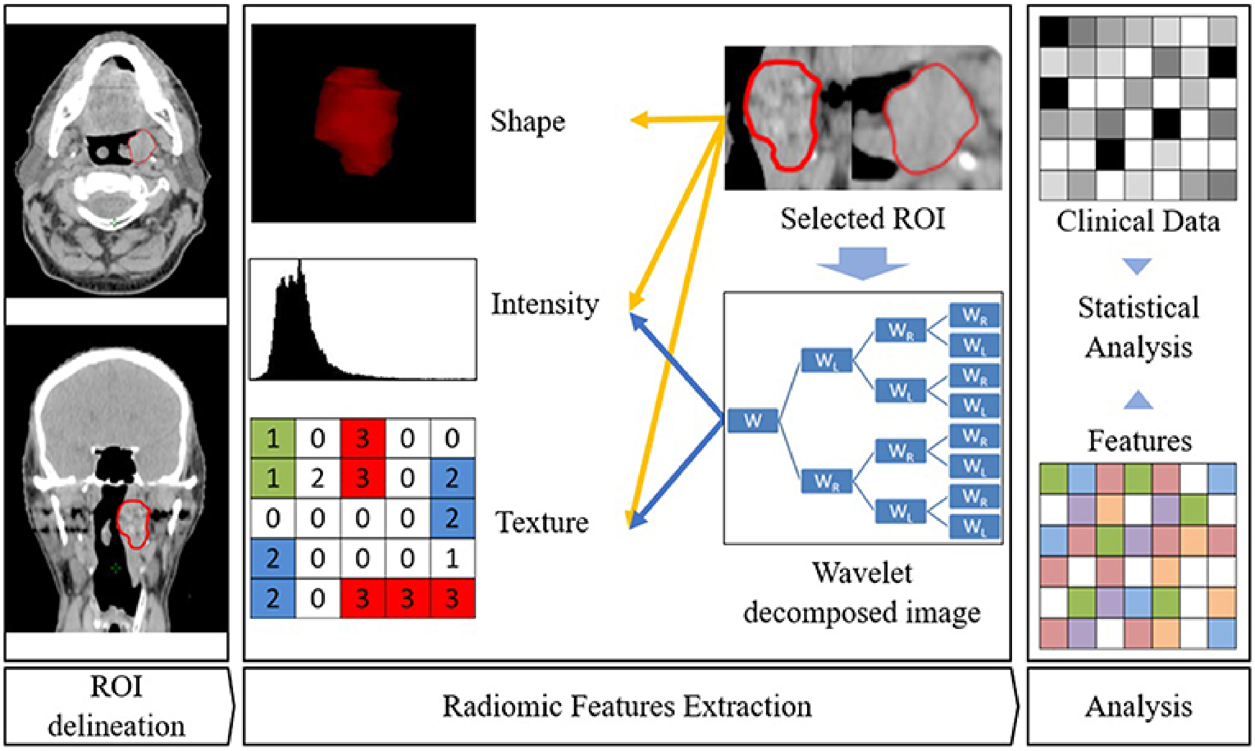


Figure 3.5 A radiomic feature pipeline applied to HNC. Features are first extracted from a segmentation on the scan, then they are analyzed through statistical models. Licensed from [8]

Despite its continued success, radiomics faces notable limitations, including sensitivity to image noise, dependency on precise segmentation masks, feature instability, and limited representational power. These challenges motivate the adoption of alternative approaches such as deep learning, which can learn complex representations directly from raw imaging data.

3.2.2 Vision Classifier Models

Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) offer a powerful alternative to traditional hand-crafted feature extraction by learning representations directly from raw image inputs. Unlike radiomics, which relies on predefined ROIs and engineered descriptors, deep learning models apply hierarchical, non-linear transformations to the entire image (or relevant subregions), enabling task-specific feature learning.

This paradigm, known as representation learning, allows models to discover complex and potentially subtle imaging biomarkers embedded within the data. Although such features are typically less interpretable than those derived from radiomics, they can provide more discriminative power for classification tasks.

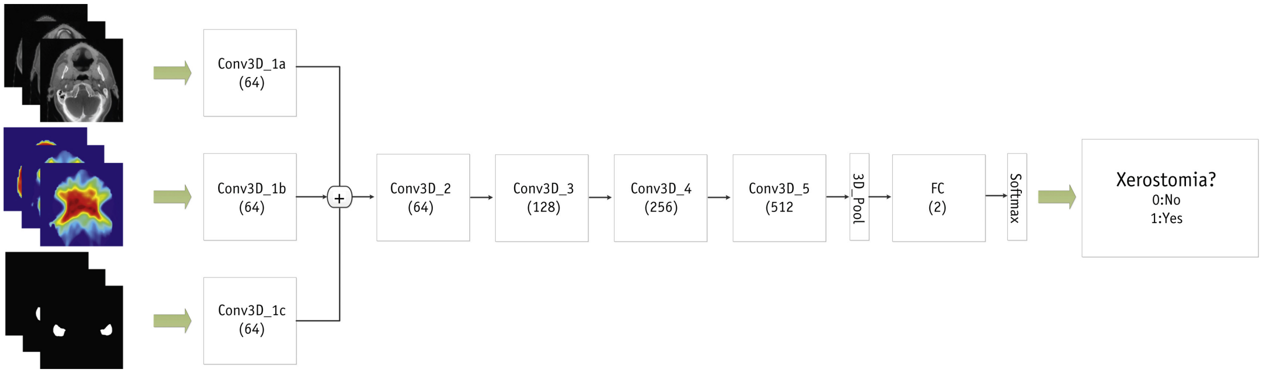


Figure 3.6 A multi-channel CNN trained for toxicity prediction in HNC. Deep features are extracted from the combination of a CT scan, a dosimetry plan and segmentation maps of OAR (parotid glands). Licensed granted from [9]

In medical imaging, deep vision models have demonstrated strong performance in a variety of tasks, including distinguishing between healthy and pathological tissues [30] and predicting treatment responses based on planning image characteristics [31]. Specifically in HNC, CNN-based models have been employed to classify radiation-induced toxicities using deformation vector fields [32]. Hybrid models combining CNN and ViT architectures have also been proposed for multitask learning frameworks aimed at predicting oncologic outcomes [33].

Self-supervised learning and foundation models

As described previously, with the development of large multi-center medical datasets, foundation models have emerged as a viable alternative to training architectures from scratch for downstream tasks. Self-supervised pre-training is employed to leverage this large volume of unlabeled data, typically by solving proxy tasks designed to learn meaningful representations without requiring any annotations. These representations can then be fine-tuned for specific clinical objectives, such as segmentation, classification, or outcome prediction.

SwinUNETR [19] was previously pre-trained [34] on five thousand CT scans, and the authors provide pretrained weights for the Swin ViT encoder that can be adapted to a variety of CT-based tasks. A 3D ResNet-50 CNN model was also trained [10] on eleven thousand radiologic lesions, offering an out-of-the-box method for lesion characterization. These pretrained features have previously been shown to be salient for prognosis prediction [35], leveraging their capacity to capture lesion aggressiveness and clinical relevance.

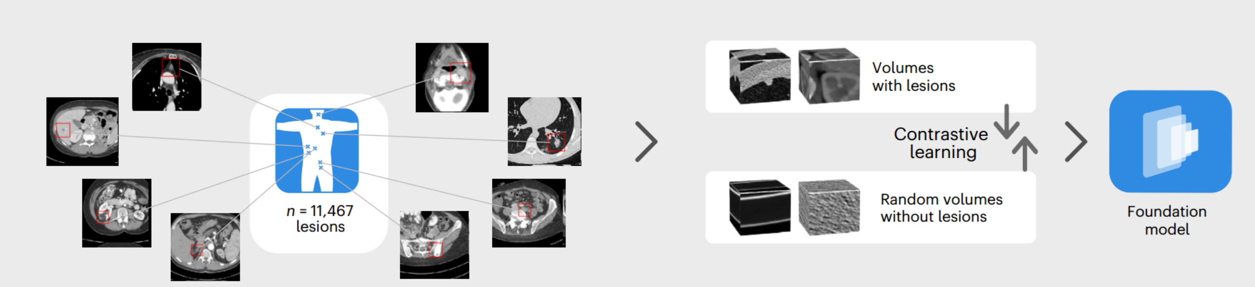


Figure 3.7 The Foundation model for cancer imaging biomarkers from [10]. Contrastive semi-supervised learning is performed on a large number of pairs coming from healthy and tumoral sites, exploring a deep representation of CT tissues. Licensed and modified by cropping from [10].

3.2.3 Application to medical image Segmentation

Medical image segmentation is the task consisting in automatically generating delineations of structures on medical scans, such as tumor volumes or OARs. This task holds significant clinical interest, as it could provide automated tools to help alleviate the annotation burden from radiologists, and more generally allow for the improved detection of relevant structures.

Vision networks are adapted to the medical segmentation task by integrating the U-shaped network initially proposed in U-Net [11] (see Figure 3.8). An initial encoder network constructs a latent representation of salient features in the training distribution through either convolutions or attention, and from this latent distribution, a decoder network resamples and upscales the representation back to a 2D or 3D output, generating segmentation.

CNNs remain today a strong contender in segmentation challenges, especially as the size of medical datasets often impairs harnessing attention to its fullest ability. nnU-Net [18] is a framework integrating multiple convolutional encoders along with fully automated preprocessing and hyper-parameter suggestions, rendering it a powerful out-of-the-box solution. In 2024 [36], this model has shown a strong ability to perform OAR segmentation on a relatively small dataset of 50 patients. Other U-Net CNN architectures have shown strong ability for head and neck structures: [37] integrated attention between skip connections to increase performance on small organs, and [38] improved on nnU-Net by including multimodal registration as a pre-processing step.

Vision transformers present a strong alternative to CNN models, especially relevant when trying to harness large quantities of data. SwinUNETR [19] increments on the attention mechanism by shifting computation windows through varying resolutions, effectively performing the same mechanism proposed by down-sampling operations. [12] and [39] demon-

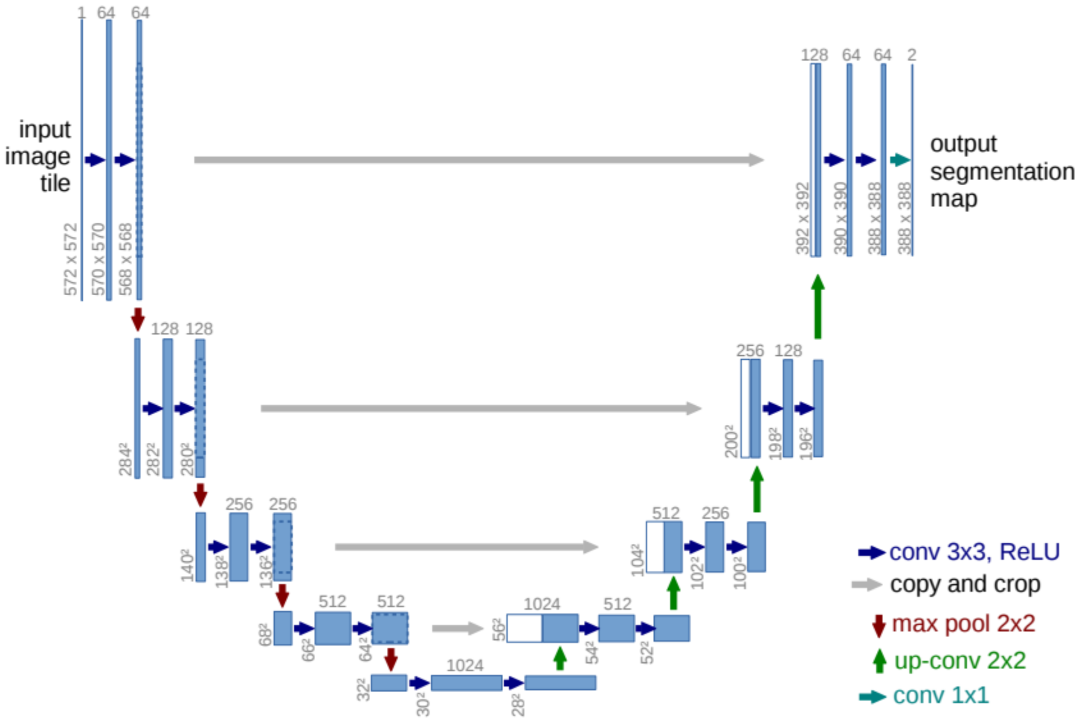


Figure 3.8 The original U-Net network introduced in 2015. Taken from [11].

strated that self-supervised pre-training, which trains models to solve simple transformation correction and in-painting on scans without the need for labels, allowed for strong leverage of large available datasets, resulting in improved fine-tuned downstream organ segmentation results compared to CNN models.

Current advances in ViT models aim today at creating foundation models, which are models pre-trained on huge quantities of data and fine-tunable to various applications. The Segment Anything Model (SAM) [40] first introduced the use of foundation models for the task of segmentation, and it was later adapted to 3D medical imaging by SamMed3D [20]. These models all utilize a ViT image encoder with an additional prompt encoder, in the form of bounding boxes, points, or even text. While these models demonstrate high performance on lesion and organ segmentation [41], their performance needs to be analyzed through the lens of the quantity of data that was used for pre-training, as well as the task simplification provided by prompts to the model, where points and bounding boxes provide the detection of structure to delineate.

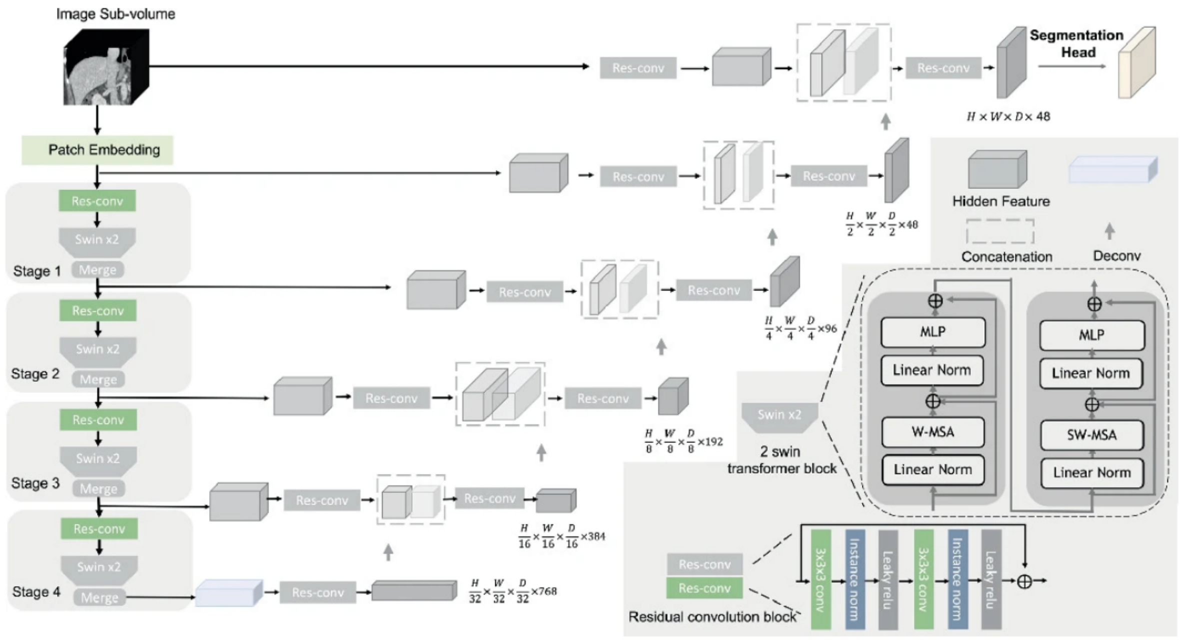


Image Sub-volume
Patch Embedding
Res-conv
Swin x2
Merge
Stage 1
Stage 2
Stage 3
Stage 4
Segmentation Head
$H \times W \times D \times 48$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$
Hidden Feature
Concatenation
Deconv
MLP
Linear Norm
W-MSA
SW-MSA
2 swin transformer block
Residual convolution block
3x3x3 conv
Instance norm
Leaky relu

Figure 3.9 SwinUNETRV2, a hybrid CNN-ViT model applied to 3D medical image segmentation. Licensed by the authors of [12]

3.2.4 Deformable image registration

Deformable image registration is the task of predicting the deformation field necessary to align two input images. In medical applications, this process may be necessary to propagate segmentations from a modality to another, in order to analyse the deformation between different patient anatomy at different time periods, or even to align treatment plans such as dosimetry with the position of a patient. Figure 3.10 presents a general overview on medical image registration.

Similar to image segmentation, neural networks employ the U-shaped family of algorithms in order to predict the deformation vector field between two input images, in either an unsupervised scheme requiring no targets, or in a supervised scheme where evaluation is based on known structures in the target image. CNN models such as VoxelMorph or SynthMorph [13] have been shown to align anatomical structures with better consistency than traditional affine and b-spline based deformable methods. Le et al have demonstrated [32] that Voxelmorph can be applied to CBCT pairs in order to extract the deformation map, between radiotherapy fractions.

ViT networks are also utilised in medical registration, models such as TransMorph [42] have shown strong inter and cross modality registration, surpassing CNN while also proposing

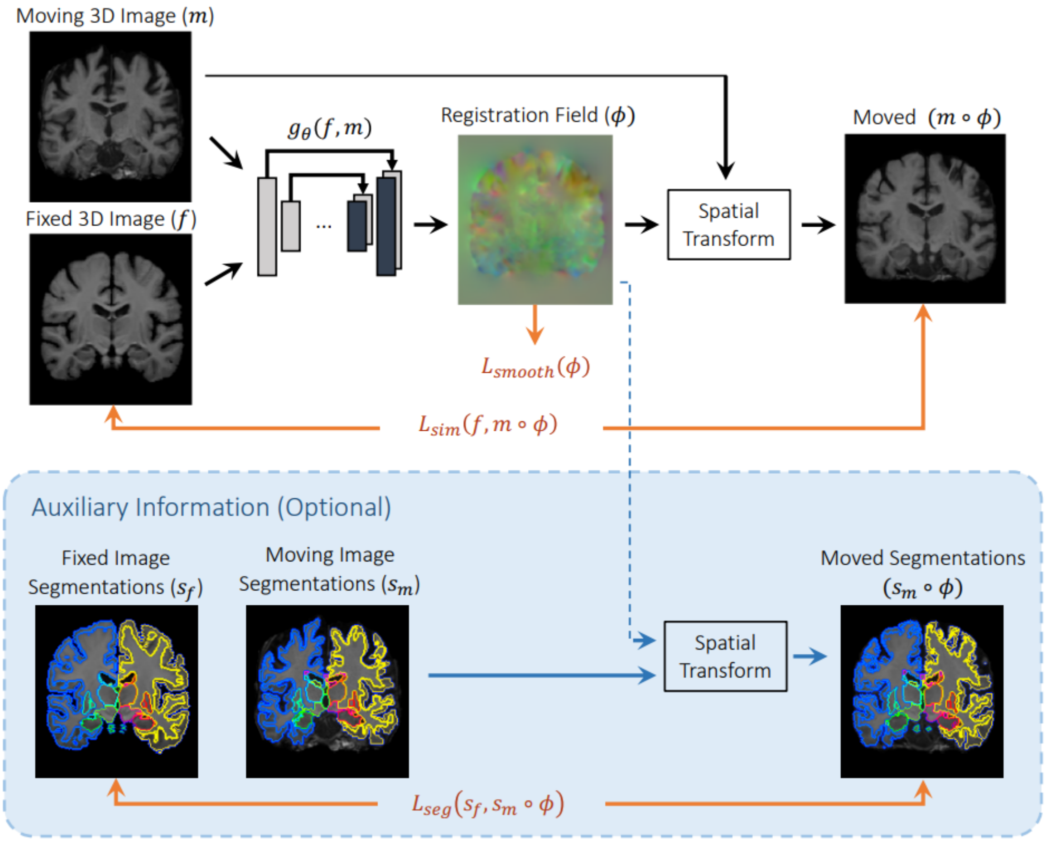

Moving 3D Image (m)
Fixed 3D Image (f)
$g_\theta(f, m)$
Registration Field (ϕ)
Spatial Transform
Moved ($m \circ \phi$)
$L_{smooth}(\phi)$
$L_{sim}(f, m \circ \phi)$
Auxiliary Information (Optional)
Fixed Image Segmentations (s_f)
Moving Image Segmentations (s_m)
Spatial Transform
Moved Segmentations ($s_m \circ \phi$)
$L_{seg}(s_f, s_m \circ \phi)$

Figure 3.10 General pipeline of deep medical registration. A deep learning generative model gθ (where θ are the model parameters) aims at estimating the deformation vector field ϕ that aligns a moving (m) image to a fixed (f) reference. The generated field can be applied with a spatial transformer network to transport images and auxiliary modalities to the target fixed space. Taken from [13]

competitive diffeomorphic variants.

3.3 Multi modal integration

Medical cohorts can comprise various modalities, including data from different scanners and their sequences, as well as tabular clinical, genomic, and pathoanatomic information, among others. From this multitude of potentially useful data arise two opposing trends: while combining modalities may enable interactions between sources that improve overall predictive performance, the increased input dimensionality can degrade performance due to the curse of dimensionality affecting statistical models.

To mitigate this issue and enable effective multimodal integration, several strategies can be employed. These strategies are typically categorized based on the stage at which fusion occurs in the processing pipeline. Figure 3.11 depicts various modality fusion strategies.

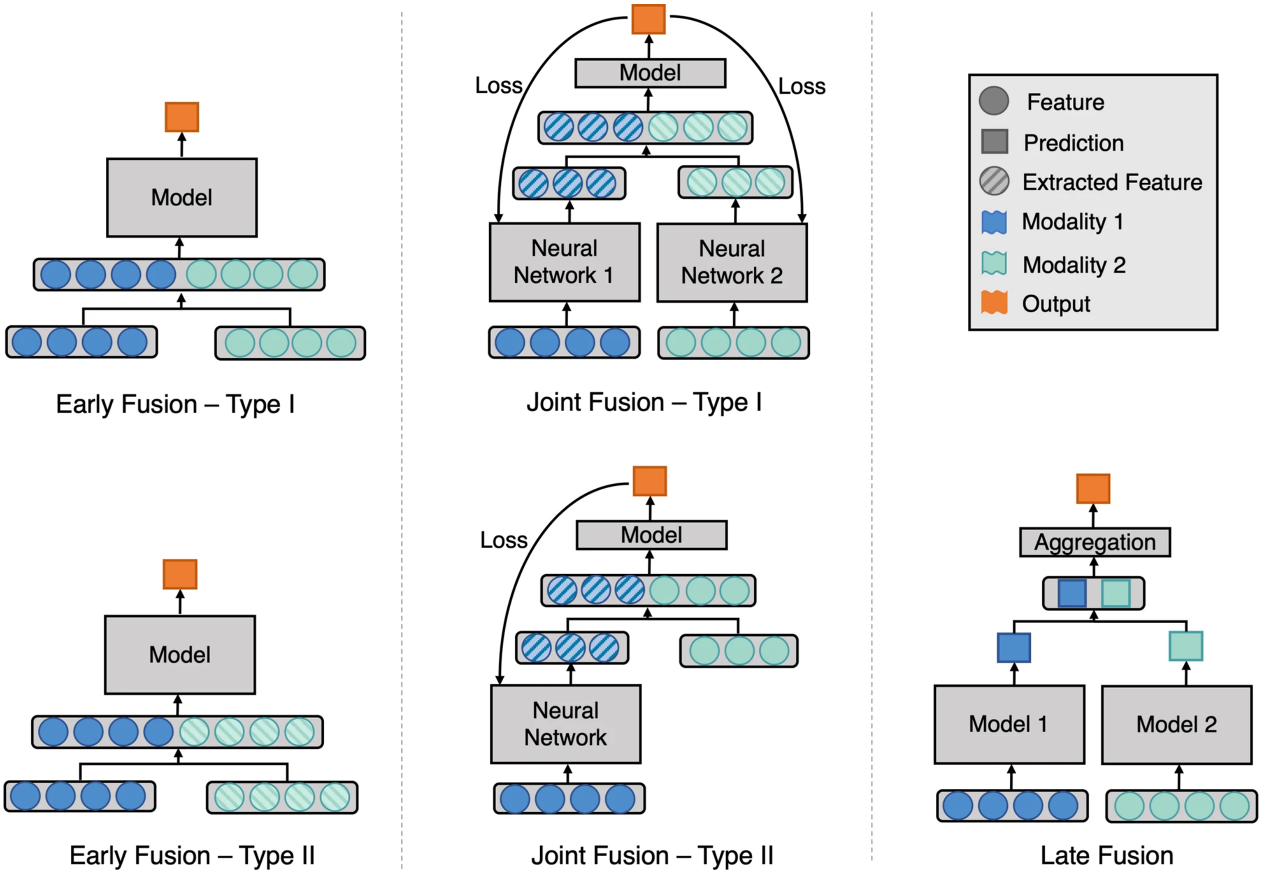


Figure 3.11 An overview of fusion strategies employed for medical modality aggregation. Based on the position of fusion, several differences in cross-interaction exploration power and computational complexity arise. Licensed from [14]

The simplest of these is early fusion, where multimodal inputs are concatenated or stacked together before being passed into a shared model. This approach treats the input modalities as a single, unified input vector:

Given two modalities x1 and x2 , early fusion involves concatenating these inputs:

$$x_{\text{early}} = [x_1; x_2] \in \mathbb{R}^{d_1+d_2} \tag{3.4}$$

This combined representation is then input into a model

$$\hat{y} = f(x_{\text{early}}) \tag{3.5}$$

where f represents the statistical learning model and $\hat{y}$ the output logit.

Early fusion is most effective when the overall dimensionality of the input remains relatively low or when all input modalities come from similar data sources, e.g., multiple scan sequences. Concatenating high-dimensional inputs can lead to over-fitting, especially in the absence of strong regularization, as the model may struggle to extract relevant features..

In [43], multimodal integration of positron emission tomography and CT images was performed by stacking the two scans channel-wise, yielding improved survival prediction results over the stand-alone modalities. In [31], a CNN model was trained to predict toxicities using stacked OAR, CT, and dose plan channels as input.

An alternative to early concatenation is late fusion, where each modality is first processed independently through a dedicated feature encoder. The resulting latent embeddings are then combined at a later stage using a fusion mechanism. This allows the model to learn modality-specific features in isolation before investigating cross-modal interactions.

Let x_1 and x_2 be two input modalities with respective encoders f_1 and f_2. We first compute the latent representations:

$$z_1 = f_1(x_1), \quad z_2 = f_2(x_2) \tag{3.6}$$

In simple late fusion, these representations can be concatenated and passed to a prediction head:

$$z_{\text{late}} = [z_1; z_2], \quad \hat{y} = g(z_{\text{late}}) \tag{3.7}$$

Where g is a downstream prediction model, typically composed of fully connected layers.

Attention-Based Late Fusion To enable more flexible interactions and allow the model to focus on the most relevant modality information, attention mechanisms can be used to fuse modalities. Specifically, a late fusion framework using multi-head attention can be formulated as follows:

Let $Z = [z_1; z_2] \in \mathbb{R}^{2\times D}$ be the stacked latent representations of the two modalities, where each $z_i \in \mathbb{R}^D$. The attention mechanism is then parametrized with with $W_Q, W_K, W_V \in \mathbb{R}^{D\times d_k}$ as learnable projection matrices.

To capture multiple interaction subspaces, multi-head attention is employed:

$$h_i = \text{Attention}(ZW_Q^{(i)}, ZW_K^{(i)}, ZW_V^{(i)}) \tag{3.8}$$

Here, H is the number of attention heads, and $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)} \in \mathbb{R}^{D \times d_k}$ are the projection matrices for the i^{th} head, while $W_O \in \mathbb{R}^{Hd_k \times D}$ projects the concatenated outputs back to the model dimension.

The final fused representation is:

$$z_{\text{attn}} = \text{MultiHead}(Z) \tag{3.9}$$

And the model prediction is:

$$\hat{y} = g(z_{\text{attn}}) \tag{3.10}$$

This formulation enables the model to dynamically weigh modality-specific contributions and to learn complex interactions between features extracted from each modality.

Late attention fusion is particularly advantageous in multimodal medical applications involving heterogeneous datasuch as the integration of clinical variables, imaging features, and molecular data where simple concatenation fails to capture subtle interdependencies. In [44] Dyam, a similar strategy was used to fuse genomic, pathomic, and radiomic embeddings to predict oncologic outcomes, leveraging the flexibility and representational power of attention mechanisms.

3.4 Survival Analysis

Survival analysis in statistical oncology refers to modelling and predicting the time until a specific clinical event occurs for a patient. These events, often referred to as outcomes, may include disease recurrence, metastasis, or overall survival. Conducting such analysis requires longitudinal follow-up data, as outcomes can occur at varying time points depending on cancer type and treatment modality. This task is of major importance to both clinicians and patients, as it allows for the identification of individuals at higher risk of recurrence and supports treatment personalization, including possible de-escalation to minimize toxicity in cases of high confidence in tumoral control.

In HNC radiotherapy, survival prediction is particularly valuable because it provides a time-dependent assessment of treatment efficacy. Unlike binary classification tasks that focus solely on outcome occurrence, survival models are designed to distinguish between early and late events, capturing meaningful differences in disease biology and progression dynamics.

To evaluate the performance of survival models in the presence of censored data, a generalization of the Area Under the Curve (AUC) has been proposed: the Concordance Index (C-index). The C-index measures the degree of concordance between the predicted risk scores and the actual survival times, accounting for censoring. It is defined as:

$$C = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\left(\hat{h}_i > \hat{h}_j\right) \tag{3.11}$$

where $\mathcal{P}$ is the set of all comparable (i.e., non-censored) patient pairs, $\hat{h}_i$ and $\hat{h}_j$ are the predicted risk scores for patients i and j, and $\mathbb{I}$ is the indicator function. The C-index reflects the ability of the model to correctly rank patients by risk.

Kaplan–Meier and Cox Modelling

When evaluating the impact of a single covariate, such as a biomarker on survival outcomes, the Kaplan–Meier (KM) estimator can be used to visualize time-to-event distributions for stratified subgroups. These curves display the probability of event-free survival over time and allow for the identification of significant group differences. A strong separation between curves suggests that the covariate meaningfully affects the outcome. To statistically validate this separation, a log-rank test can be performed, yielding a p-value that indicates whether survival differences between groups are significant.

While KM curves are valuable for univariate group comparisons, they do not provide patient-specific risk scores. For multivariate modeling and individualized risk estimation, the Cox proportional hazards model is widely employed. This semi-parametric model assumes that covariates have a multiplicative effect on the baseline hazard, and that these effects remain constant over time. The hazard function in the Cox model is defined as:

$$h(t|X) = h_0(t)\exp(\beta^\top X) \tag{3.12}$$

where $h(t|X)$ is the hazard function at time t for a patient with covariates X, $h_0(t)$ is the baseline hazard function, and β is the vector of learned coefficients corresponding to each covariate.

Cox models have been effective in predicting oncological outcomes and remain a standard in clinical survival analysis. However, their reliance on linear relationships between covariates and the log-risk function, as well as the proportional hazards assumption, can limit their expressiveness in capturing complex, non-linear interactions within high-dimensional clinical data.

Deep survival

To build upon classical statistical survival analysis methods, deep learning models were first introduced as powerful covariate extractors for downstream Cox models. Leveraging their capacity for nonlinear feature extraction, these models have demonstrated superior survival risk estimation compared to traditional machine learning approaches, especially in high-dimensional biomedical datasets.

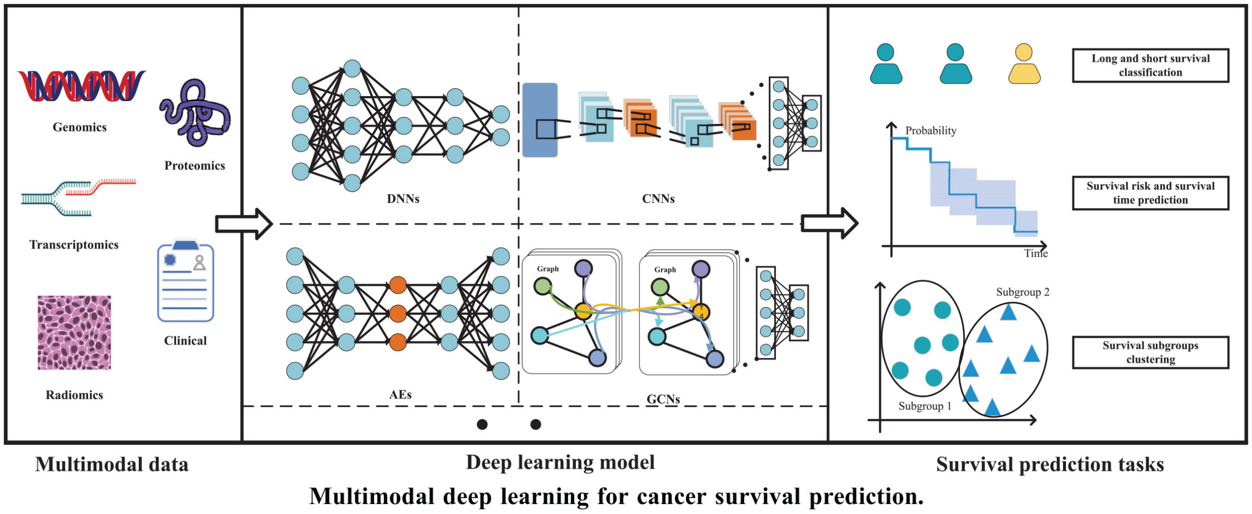

Genomics
Proteomics
Transcriptomics
Clinical
Radiomics
DNNs
CNNs
AEs
Graph
Graph
GCNs
Long and short survival classification
Probability
Time
Survival risk and survival time prediction
Subgroup 2
Subgroup 1
Survival subgroups clustering
Multimodal data
Deep learning model
Survival prediction tasks
Multimodal deep learning for cancer survival prediction.

Figure 3.12 An overview of deep survival methods. From multi-modal medical input, several feature extractor deep learning models can be used to group populations in relative risk classes. Licensed from [15]

In order to enable end-to-end training, two primary frameworks have emerged for integrating Cox modeling within neural networks:

Cox-loss Penalized Deep Learning Models. DeepSurv [45] is one of the most prominent frameworks implementing the Cox proportional hazards model directly within a deep neural network. The loss function is formulated as the negative log partial likelihood of the Cox model. Given an input feature vector x_i, a neural network $\hat{h}_\theta(x_i)$ outputs a risk score. The objective function to be minimized is:

$$\mathcal{L}(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j \in \Re(T_i)} e^{\hat{h}_\theta(x_j)} \right) + \lambda \|\theta\|_2^2 \tag{3.13}$$

where $\mathfrak{R}(T_i)$ denotes the risk set of patients still under observation at time T_i, and $\lambda\|\theta\|_2^2$ is an L_2-regularization term. This approach has proven effective for outcome prediction in oncology. It was shown to outperform decoupled feature-extraction and Cox modeling pipelines in colorectal cancer prognosis [30], and has also been applied in head and neck cancer (HNC) survival estimation.

Discrete Time Models and MTLR. An alternative family of deep survival models seeks to avoids the assumptions of the Cox, model particularly the proportional hazards assumption, by discretizing the time axis and directly modeling event probabilities in each interval. These models predict the likelihood that a patient will experience the event in specific time bins, offering more flexible time-dependent risk estimates.

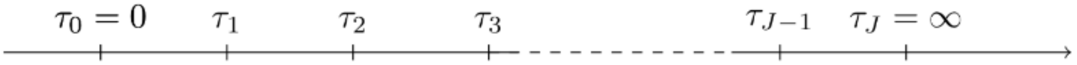


Figure 3.13 Time bin discretisation diagram used in MTLR modelling. τ_i represents the follow up time at i-th bin. Figure taken from [16]

Multi-Task Logistic Regression (MTLR) [46] and its neural adaptation (N-MTLR) model survival as a sequence of binary classification tasks across time intervals. Each task estimates the probability of surviving beyond a specific time point. The survival function for a patient with feature vector $\vec{x}$ over the interval $[\tau_{s-1}, \infty)$ (refer to figure 3.13 for a visualisation) is computed as:

$$S(\tau_{s-1}, \vec{x}) = \sum_{k=s}^{J} \frac{\exp\left(\sum_{l=k+1}^{J} \psi_l(\vec{x})\right)}{Z(\psi(\vec{x}))} \tag{3.14}$$

where $\psi : \mathbb{R}^p \to \mathbb{R}^J$ is a neural network that maps the input vector $\vec{x} \in \mathbb{R}^p$ to a score vector across J discrete time intervals, and $Z(\psi(\vec{x}))$ is the normalization factor:

$$Z(\psi(\vec{x})) = \sum_{j=1}^{J} \exp\left(\sum_{l=j+1}^{J} \psi_l(\vec{x})\right) \tag{3.15}$$

This framework produces a smooth, continuous approximation of the survival function and does not rely on the restrictive assumptions of the Cox model. However, to ensure consistency across time intervals and interpretability of risk scores, additional regularization is often required.

This methodology has previously been applied to HNC survival prediction in [47] oropharyngeal cancer, where primary tumor radiomics were used as the input to a N-MTLR feed forward network.

3.5 Summary of Literature review

In this overview of methods from the literature, we have established a foundation for employing computer vision techniques in the context of HNC cases. Deep learning-based statistical models have shown considerable promise in extracting meaningful features from medical imaging, enabling their application across a wide range of clinical tasks.

We have highlighted how these models can be used to predict both treatment-related toxicities and oncological outcomes, thereby contributing to more personalized treatment planning. Nevertheless, their application in HNC remains relatively under explored, primarily due to the scarcity of multi-modal datasets in this domain.

CHAPTER 4 METHODOLOGY OF THE RESEARCH PROPOSAL

The preceding sections have outlined the key clinical challenges that persist in current HNC radiation therapy, along with a review of state-of-the-art methodologies in computer vision-guided clinical analysis. This work is divided into two complementary parts: the first introduces a novel approach for radiation toxicity prediction based on the relationship between dosimetry and anatomical deformations; the second proposes a new framework for oncological outcome prediction that incorporates extranodal extension (ENE) features.

4.1 Problem Statement

Adaptive radiotherapy for HNC offers the potential for more effective oncological treatment with a reduced risk of adverse events. Specifically, reducing the radiation dose to regions associated with favorable prognoses may help limit radiation-induced toxicities while maintaining effective tumor control. However, the successful implementation of such an approach requires robust stratification of both cancer types and patient profiles, enabling the classification of patients into high and low risk groups of toxicities and tumoral recurrence.

Through the methodologies developed in this work, we aim to better characterize and model both static and dynamic contributors to treatment response. Ultimately, this could support improved clinical decision-making, including personalized treatment planning, dose de-escalation strategies, and early identification of non-responders.

4.2 Chronology and Interactions Between Studies

The first component of our study investigates anatomical changes that occur during treatment, as patients' anatomies evolve in response to both radiation and concurrent therapies. While changes in tumor volume can indicate treatment efficacy, alterations in the shape or position of organs-at-risk (OARs) may invalidate the original dose plan, resulting in sub-optimal dosing and increased toxicity risk. Investigating the interaction between dosimetry and anatomical deviations from the planned anatomy may enhance our ability to predict treatment-related toxicities.

Evaluating the interaction between dosimetry and CBCT-based deformation maps proved challenging due to the noisy nature of the imaging and the difficulty in accurately mapping deformations across longitudinal instances. These limitations made it difficult to assess the adequacy of the initial treatment plan. Consequently, we decided to refocus initially on a more tractable target, as a step toward advancing the proposed strategy.

In the second phase, we focused on treatment-related toxicities as they are represented in the radiation therapy plan. Certain OARs may possess anatomical or functional features that complicate dose delivery, thereby increasing the likelihood of adverse effects. In some cases, the dosimetric configuration itself may predispose patients to toxicities, particularly when tumor geometry limits the ability to adequately spare surrounding healthy tissues.

In this context, the identification of extranodal extension (ENE) in HPV-positive patients emerged as a clinically significant target and a potentially valuable prognostic marker. We therefore integrated this pathology into our study to enhance data representation in a future pipeline for toxicity prediction that includes OARs.

4.3 Hypotheses

Based on the challenges outlined in the previous section and our research rationale, we have formulated the following hypotheses:

- **Hypothesis 1**: Investigating the interactions between daily anatomical deformations and dosimetry will improve the prediction of radiation-induced toxicities.
- **Hypothesis 2**: Incorporating extranodal extension and tumor-related features will enhance predictions of oncological outcomes in HPV-positive oropharyngeal cancer patients.

4.4 Objectives

1. **Development of a deep learning framework that integrates daily anatomical deformations and dosimetric data to enable early and accurate prediction of radiation-induced toxicities in head and neck cancer patients, facilitating risk stratification during treatment delivery.**
2. **Design and exploration of a multimodal prognostic model incorporating extranodal extension (ENE) features and tumor characteristics to predict key oncological outcomes in HPV-positive oropharyngeal cancer patients, enabling patient risk stratification at radiotherapy planning.**

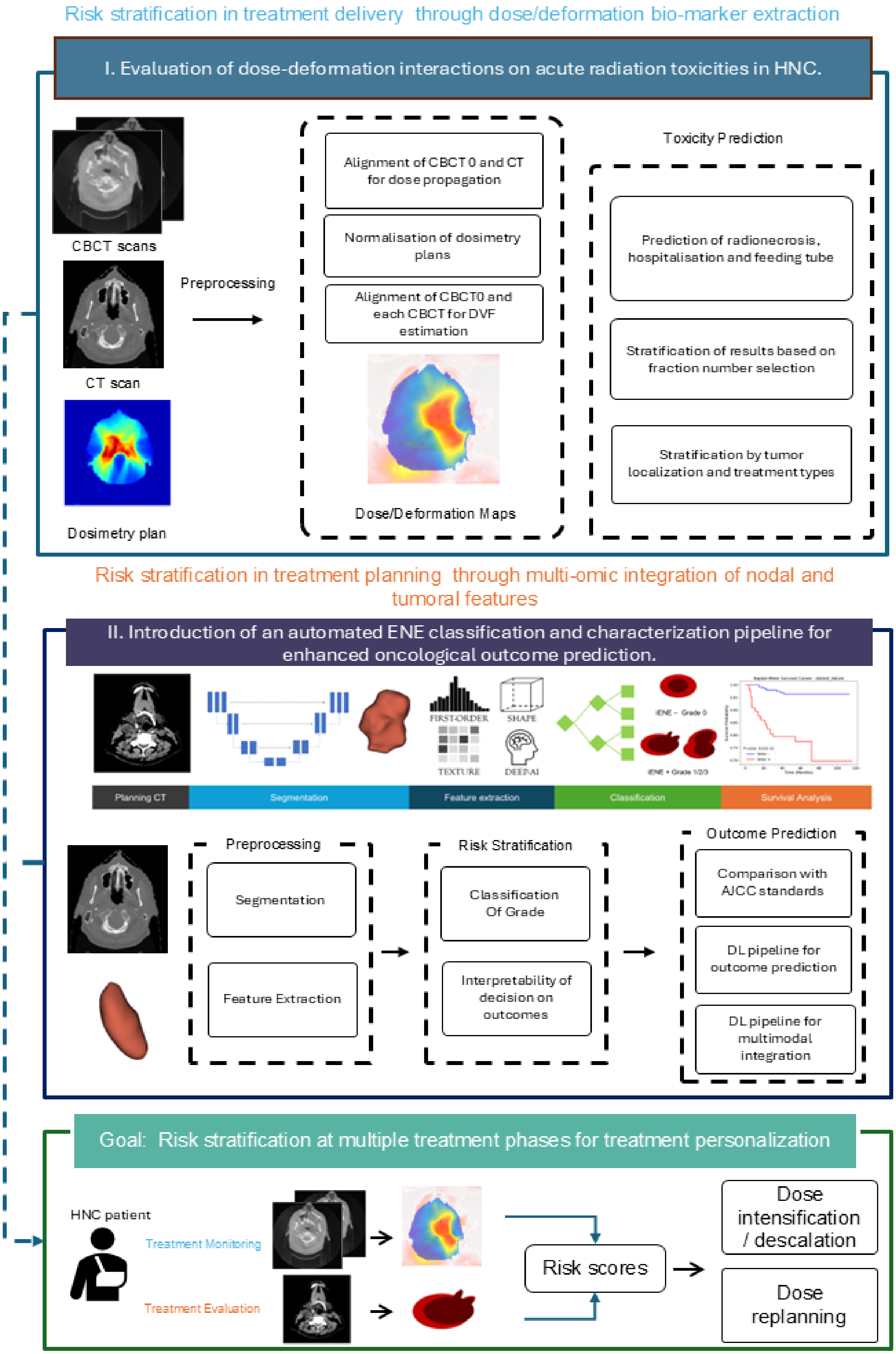

Risk stratification in treatment delivery through dose/deformation bio-marker extraction
I. Evaluation of dose-deformation interactions on acute radiation toxicities in HNC.
CBCT scans
CT scan
Dosimetry plan
Preprocessing
Alignment of CBCT0 and CT for dose propagation
Normalisation of dosimetry plans
Alignment of CBCT0 and each CBCT for DVF estimation
Dose/Deformation Maps
Toxicity Prediction
Prediction of radionecrosis, hospitalisation and feeding tube
Stratification of results based on fraction number selection
Stratification by tumor localization and treatment types
Risk stratification in treatment planning through multi-omic integration of nodal and tumoral features
II. Introduction of an automated ENE classification and characterization pipeline for enhanced oncological outcome prediction.
FIRST-ORDER
SHAPE
TEXTURE
DEEP AI
iENE - Grade 0
iENE + Grade 1/2/3
Planning CT
Segmentation
Feature extraction
Classification
Survival Analysis
Preprocessing
Segmentation
Feature Extraction
Risk Stratification
Classification Of Grade
Interpretability of decision on outcomes
Outcome Prediction
Comparison with AJCC standards
DL pipeline for outcome prediction
DL pipeline for multimodal integration
Goal: Risk stratification at multiple treatment phases for treatment personalization
HNC patient
Treatment Monitoring
Treatment Evaluation
Risk scores
Dose intensification / descalation
Dose replanning

Figure 4.1 Overview of the methodologies and chronological structure of the work presented in this proposal.

4.5 Overview of the methods employed

4.5.1 Dose-Aware toxicity prediction

Design

To build upon previous work in toxicity prediction for head and neck cancer (HNC), we investigated the interaction between dosimetry and patient anatomy through longitudinal morphological comparisons. Our working hypothesis is that radiation toxicities are exacerbated by anatomical shifts that result in inadequate dose coverage in regions originally planned to spare OAR. By combining the amplitude of anatomical deformations with local dose levels, we propose the use of dose-deformation maps to highlight regions where significant anatomical shifts, coupled with high radiation exposure, may contribute to increased toxicity.

To identify zones at risk for toxicity, we normalized the dosimetry plans and applied them as attention maps to the Jacobian determinants of the deformation vector fields (DVFs) between a reference $CBCT_0$ scan and a subsequent $CBCT_t$, where t corresponds to a treatment fraction between 1 and 30. This results in a dose-aware Jacobian determinant (DA-JDET) map, defined as:

$$DA_JDET(0 \rightarrow t) = Dose \times JacDet(DVF_{0 \rightarrow t}) \tag{4.1}$$

where $DA_JDET0 \rightarrow$ t represents the proposed dose-aware deformation map from fraction 0 to fraction t, Dose CBCTis the normalized dose distribution mapped onto the initial CBCT0, and JacDet DVF0 $\rightarrow$ t is the Jacobian determinant of the deformation vector field between $CBCT_0$ and $CBCT_t$. This formulation highlights regions where high-dose areas intersect with significant anatomical deformation, indicating a potential for increased radiation-induced toxicity. Normalization of the dose plan was performed by gathering the distribution of intensities over the entire training sets, and z-score normalization was applied. This operation is essential for the stability of the training procedure, and allows for the model to understand normal dosimetry in HNC radiotherapy treatment.

A significant challenge in this process involved transferring the dose plan to $CBCT_0$, as the reference CT scan on which the dosimetry is originally based, is typically acquired weeks or months before the first treatment fraction. During this time, substantial anatomical changes can occur, making registration between the pre-treatment scans difficult.

These challenges are compounded by the considerable differences in domain and field of view between modalities. Moreover, CBCT scans are inherently noisy and subject to artifacts, necessitating filtering procedures to mitigate disruptions to network training.

Evaluation of the Method

A retrospective cohort of 1,012 HNC patients was collected and utilized to validate the toxicity prediction models. Each patient had an associated planning CT scan, a set of 30 to 33 CBCT scans acquired during treatment, a corresponding dosimetry plan, and relevant clinical data.

The CT scans were first resampled to an isotropic voxel spacing of 2mm, then cropped to a size of 128×128×128 voxels centered around the clinical target volume to align with the field of view of the initial CBCT scan.

An affine registration was performed between the planning CT and the first CBCT to warp the original dosimetry plan into the CBCT space. Dosimetry plans were transported to CBCT space accordingly, with no deformable deformation of the intensity distribution. Subsequently, a deformable registration model based on VoxelMorph [13] was employed to compute the deformation vector fields (DVFs) between pairs of CBCT scans. These DVFs were further processed to derive their Jacobian determinant maps, capturing local anatomical volume changes over time.

A multi-branch architecture combining a ResNet-50 backbone and a multilayer perceptron was then trained to predict radiation induced toxicities specifically, feeding tube requirement, radionecrosis, and unplanned hospitalization, based on the combination of dose maps and CBCT scans.

Experimental details

Model performance was assessed using 5-fold stratified cross-validation, with results reported as the mean and standard deviation of each evaluation metric across the folds.

Each model was trained for 100 epochs using the Adam optimizer and the OneCycle learning rate scheduling strategy. Data augmentation techniques were applied to improve generalizability and robustness. Augmentation was performed using the TorchIO framework and included the following transformations: affine rotations in the axial plane of up to 20°, random flipping, and translations in the plan of up to 60 mm.

Model hyperparameters were optimized via grid search over the following configurations: learning rates 0.01, 0.007, 0.001, batch sizes 2, 4, 8, and dropout ratios 0.3, 0.4.

4.5.2 AMO-ENE : attention based multi omic fusion for survival analysis in HPV ENE associated OPC.

Design

In the second part of this work, we investigate ENE as a biomarker for poor oncological outcomes in OPC. The proposed AMO-ENE framework consists of two main components: an automated pipeline for grading imaging-based ENE (iENE) and a multi-modal outcome prediction model.

The first component addresses a critical clinical need for a more robust and automated iENE grading system, as current evidence regarding the role of pathological lymph nodes in HPV-positive cancers remains limited. The second component, AMO-ENE, aims to integrate heterogeneous data sources through a multi-modal approach, leveraging deep feature extraction and robust fusion mechanisms to enable cross-modality interactions and improve outcome prediction performance.

This study contributes to the broader objective of enhancing our understanding of head and neck cancers by introducing a method for characterizing a key structure at risk and laying the groundwork for a more comprehensive, multi-organ modelling framework.

Grading iENE poses significant challenges. Pathological lymph nodes often have poorly defined borders in ct imaging, particularly in regions where infiltration into surrounding tissues occurs, leading to low-resolution and ambiguous anatomical boundaries. Furthermore, patients may present with multiple pathological nodes at varying stages of iENE, necessitating effective node selection or aggregation strategies to isolate the clinically most relevant instance.

Traditional deep survival analysis methods in the literature either make strong assumptions about the risk distribution, such as the proportional hazards assumption in Cox models, or are vulnerable to over-fitting when confronted with high-dimensional inputs. To address these limitations, we employ a late fusion strategy based on latent-modality multi-head self-attention fusion, which allows for interaction across low-level features and the selection of salient information from each modality.

To model time-dependent risk scores, we adopt the MTLR framework, avoiding the need for assumptions about the underlying hazard function and thereby offering greater flexibility in modeling complex survival outcomes.

Evaluation of Method

We evaluated our method using a retrospective cohort of 397 oropharyngeal cancer (OPC) patients, 346 of whom were also included in the toxicity prediction study. For each patient, a planning CT scan was provided along with the corresponding clinically segmented ground-truth mask and an associated IENE (Ipsilateral Extranodal Extension) score. For 55 patients within this cohort, clinical grading of IENE was independently performed by three board-certified neurologists. This allowed us to assess inter-rater agreement among clinicians and compare it against the model's predictions.

Segmentation Task Evaluation Details All CT images were resampled to a voxel spacing of 1.5 × 0.9766 × 0.9766 mm using third-order B-spline interpolation for the CT volumes and nearest-neighbour interpolation for segmentation masks. Intensities were clipped to the 0.5th and 99.5th percentiles of the Hounsfield Unit (HU) distribution, then normalized using standard z-score scaling.

Training and inference were conducted on cropped patches of size 80 × 192 × 160 voxels. A sliding-window inference approach was used during testing to ensure full-volume coverage and improved robustness.

Segmentation performance was evaluated using 5-fold stratified cross-validation. The Dice Similarity Coefficient (DSC) was used as the primary metric, reported as the mean and standard deviation across all folds. To mitigate potential bias from small nodal structures, we calculated DSC only on the largest pathological node in each case.

The nnU-Net framework was trained for a maximum of 750 epochs; however, convergence was typically achieved earlier. Training used the default data augmentation pipeline and learning rate scheduling provided by the framework.

IENE Grade Classification Task Evaluation Details Radiomic features were extracted from both the primary tumor volume and the predicted ENE masks. All 11 families of radiomic features were extracted using PyRadiomics [48], with a fixed HU bin width of 10 units. CT scans were resampled to an isotropic voxel spacing of 1 mm using B-spline interpolation. No additional intensity normalization was applied during feature extraction.

Grades were grouped in binary classification schemes, in order to account for the high class imbalance : 0 - 1/2/3 represents the most important case considering presence/absence of ENE regardless of severity ; 0/1 - 2/3 aggregates low severity classes, relevant as significant doubt between these two classes exists ; and 0/1/2 - 3 in order to separate the high severity cases from the rest. To further combat class imbalance, synthetic minority over-sampling [49] is applied to training folds.

A comprehensive grid search was conducted for model hyperparameter optimization. The search space included:

- Classifier models: $\{RF, MLP, XGBOOST\}$
- Number of estimators: $\{50, 100, 200, 500, 1000\}$
- Tree depths: $\{3, 5, 7, 10\}$
- Initial learning rates: $\{0.01, 0.1, 0.2\}$
- L1 ratios: $\{0.001, 0.01, 0.1, 1\}$
- Number of PCA components: $\{5, 10, 15, ..., 100\}$

Outcome Prediction Task Evaluation Details The AMO-ENE model was evaluated on two clinically relevant outcome prediction tasks designed to capture both short-term and long-term disease trajectories:

- **Binary 2-year outcome classification**: This task focuses on predicting whether patients experience an adverse event (e.g., recurrence or death) within two years following diagnosis. The choice of a 2-year threshold is clinically motivated, as it aligns with standard oncology follow-up protocols and ensures all patients in the cohort had sufficient follow-up time.
- **Continuous risk assessment**: To capture the full temporal dynamics of disease progression, we also modeled time-to-event outcomes using a continuous survival prediction framework, thus enabling individualized risk assessment across the entire follow-up distribution.

Input features for the model included radiomic descriptors extracted from both the primary tumor and the predicted ENE region, using the same extraction parameters as described in the classification task. In addition to radiomics, we incorporated clinical variables (e.g., age, stage, performance status) to enable a multi-modal input structure. This combination reflects our multi-organ, multi-source learning approach, designed to capture both local and systemic indicators of prognosis.

Time bins for the MTLR modelling were obtained through the method described in [46], where the number of bins is set by obtaining the square root of number of observations, and the bin value obtained by attributing the correspondent quantile of follow-up time.

All models were evaluated using 5-fold stratified cross-validation, ensuring consistent splits across experiments to facilitate fair comparison. Performance was assessed using the Area Under the Curve (AUC)for the binary classification task, and the Concordance index (C-index) for the survival analysis task.

Hyperparameter tuning was performed via grid search over the following ranges:

- Learning rate: $\{0.01, 0.001, 0.007\}$
- Batch size: $\{8, 16, 32\}$
- Hidden layer sizes (latent space): $\{256, 512, 1024\}$
- Dropout ratios: $\{0, 0.3, 0.4\}$
- Number of attention heads: $\{1, 2, 3, 4\}$

To prevent overfitting, early stopping was employed during training. If the validation loss did not improve for 20 consecutive epochs, training was halted. This approach helped ensure model generalizability without compromising performance

CHAPTER 5 ARTICLE 1 : DOSE AWARE TOXICITY PREDICTION IN HEAD AND NECK CANCER PATIENTS USING A DEFORMABLE 3D CNN ON DAILY CBCT ACQUISITIONS

Gautier Henique[†], Chulmin Bang[⋆], William Le[†], Edith Filion[⋆], Phuc Felix Ngyuen-Tan[⋆], Houda Bahig[⋆], Samuel Kadoury[⋆†]

[†] Polytechnique Montréal, Montréal, Québec, Canada
[⋆] Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, Québec, Canada

5.1 Publication

This paper has been published in the IEEE transactions on the 21st International Symposium on Biomedical Imaging (ISBI 2024) and presented in Athens as an oral presentation in the Surgery guidance and planning special session. The official publication was rendered on the 27th of May 2024.

5.2 Contribution

My contributions to this work covers all the coding, planning, experiments, and writing.

5.3 Abstract

With recent advancements in intensity-modulated radiotherapy and image guidance for cancer treatment, there is growing interest in Deep Learning-based Adaptive Radiotherapy due to its potential to mitigate radiation toxicity. During the course of treatment, daily acquisitions of cone beam computed tomography (CBCT) allows to monitor significant anatomical changes around the tumour target area and the response to treatment. Previous works have demonstrated the benefits of using anatomical deformation features as predictors of early toxicities. The objective of this work is to investigate the use of radiomic distributions to predict early reactive nasogastric tubing (NG tube), radionecrosis and broad hospitalisation based on initial dosimetry plans. We propose a method combining inter-fractional anatomical deformation, dosimetry and clinical information to improve the prediction performance. For this work we implement a deformable registration pipeline and train a 3D convolutional neural network model on the Jacobian determinants of the deformation vector fields between

different fractions of treatment. We exploit the dose plans as feature maps to focus the attention of the network on areas susceptible to radiation toxicity. We obtain balanced accuracy scores of 75.4% for radionecrosis at fraction 20, 61.1% for hospitalisation after the first week of treatment and 74.1% for NG tube insertion after the 5th week.

5.4 Introduction

In recent years, cone-beam computed tomography (CBCT) has become an integral part of the standard radiation therapy treatment workflow to improve patient positioning during radiotherapy treatments. For head and neck cancer (HNC), most tumour sites require 60–70 Gy fractioned over 6–7 weeks of 5-day weekly treatments, totalling 30–35 longitudinal volumetric CBCT images [50]. Furthermore, as the recent increase of deep learning based predictive models is linked in part to the large availability of data, it becomes logical to leverage both of these aspects — such as the convolutional neural network (CNN) [51] and its ResNet [17] variant that have become widely popular across every domain — with the large amount of CBCTs obtained during HNC treatment.

Despite these daily image acquisitions, no consistent clinically actionable information have been extracted from this wealth of patient specific data, in spite of several studies [52–55] showing the existence of dynamic inter fraction biomarkers potentially helping to re-adapt treatment plans.

Using multivariate regression, Qin et al. [56] showed that progression-free survival and lung toxicity could be predicted from biomarkers extracted from serial CBCTs during non-small cell lung cancer treatment on 34 stage I patients. In HNC, Rosen et al. [57] performed a deformable image registration between the planning CT and the daily CBCT to measure the change in delivered dose, where a univariate regression model was trained to predict xerostomia, yielding an 0.77 AUC in outcome prediction. Recently, Le et al. [58] combined features extracted from the deformation vector field (DVF) — similarly obtained from a registration of CT to CBCT — with patient clinical data to predict hospitalisation, reactive nasogastric (NG) tube and radionecrosis risk via a multi-branch ResNet model on a cohort of 292 HNC patients. Furthermore, while much recent efforts were focused on exploiting this imaging modality for dose calculations and radiomic feature analysis, physiology is only one part of the patient specific information also available during radiotherapy.

Optimisation of dose plans is a complex and multivariate process that must take into account factors such as the dose delivery system, the organs at risk (OAR), and the specific anatomy of the patient, with outcomes affected by dose conformity, physiological response, and clinical comorbidity. It therefore becomes evident that the dose plan, itself a 3D map of tissue irradiation levels, must be included in any predictive method or model that aims for patient-specific adaptive treatment. Recently, Wentzel et al. [59] combined standard patient information with hand-crafted spatial features on the dose-organ interaction with a logistic regression model to predict radiation associated dysphagia on a cohort of 200 HNC patients. In Men et al. [60], xerostomia risk was classified using a residual 3D CNN with a combination of CT, dose plans, clinical parameters, and parotid and submandibular glands contouring, achieving 0.84 AUC when trained on 784 patients.

The purpose of this study is to investigate the incorporation of dosimetric information to a previous method combining clinical parameters with anatomical deformations between serial CBCT and pre-treatment CT in predicting post-treatment radionecrosis, reactive early NG tube and global hospitalisation. Specifically, we use a dataset of 1012 HNC patients using a deep learning based 3D deformable registration and multi-branch classification models to create deformation maps between CBCT fractions, while applying normalised attention maps of the dose plans on the Jacobian matrices of the DVFs. The work is based on the hypothesis that alterations in the anatomy can result in a misalignment between the initially planned radiotherapy distribution and the targeted zones before treatment. Dose distributions are typically focused around the tumoural sites, however OARs will still receive residual gradient dose. Because changes in the patient anatomy are likely to occur during treatment, radiation therapy and hospitalisation can cause tissue shrinkage and overall weight loss, while chemotherapy may result in tissue swelling due to fluid injection. These factors collectively contribute to a shift in the patient's original anatomical configuration. This discrepancy between the initial plan and the modified anatomical structure further heightens the risk of toxicities in the affected regions, as healthy tissue may be irradiated with non-standard dosage.

The contributions of this study are as follows:

1. The introduction of a joint deformation-dose toxicity prediction process based on a 3D ResNet.
2. The evaluation of dose distributions in HNC radiotherapy toxicity prediction on a dataset of 1012 patients.
3. The assessment of the toxicity predictive performance while controlling for different tumour sites.

5.5 Materials and Methods

5.5.1 Dataset

A cohort of 1012 HNC patients was retrospectively collected, with patients undergoing histologically confirmed radiotherapy at our research hospital between 2016 and 2022. This dataset includes patients undergoing primary or adjuvant radiotherapy, or chemoradiotherapy; all selection criteria, toxicity definitions and treatments follow the same protocol. The following primary tumour sites were included: oropharynx, hypopharynx, larynx, oral cavity, sino-nasopharynx and unknown primaries. In this dataset, toxicity incidences are 16.6% for reactive NG tubing, 4.2% hospitalisation rate during treatment, and 4.6% observed radionecrosis up to 6 months after treatment. Exclusion criteria include subjects in palliative care and intubation before treatment.

5.5.2 Preprocessing

We follow a similar preprocessing pipeline to [58], where all images are resampled to 2mm isotropic spacing via nearest neighbour interpolation, then the CTs were automatically cropped around the clinical target volume to a margin of 1.4 times the corresponding CBCT0 volume size in order to match the field of views. The volumes were then interpolated to a matching 128x128x128 volume size via nearest usampling. A first pretrained VoxelMorph [13] model trained to register the planning CT (pCT) to CBCT_0, allows to non-rigidly deform both OARs and dose distributions on the initial CBCT. Equation 5.1 described the proposed process:

$$\text{DVF}_{Dose/OAR} = \text{pCT} \rightarrow \text{CBCT}_0\ . \tag{5.1}$$

An additional pretrained VoxelMorph model trained on a separate in-house subset of pairs of 392 CBCT volumes, allows to register the baseline CBCT_0 with each serial $\text{CBCT}_{t>0}$. The composition of the rigid and deformable registration models allows an affine alignment of CBCT volumes and calculating the DVF between the CBCT_0 and the desired daily CBCT as shown in Eq. 5.2 where → indicates a left-associative registration between scanning pairs and N the total number of fractions:

$$\text{DVF}_t = \text{CBCT}_0 \rightarrow \text{CBCT}_t\ \forall t \in [1, N]. \tag{5.2}$$

Due to computational resource limitations, three channels DVFs were transformed to their corresponding 3D Jacobian determinants ($J\text{det}$) using the SimpleITK library. This Jacobian matrix represents the first order partial derivatives of the DVF in each dimension. Here, we define $f : R^3 \rightarrow R^3$ (x,y,z base) as the deformation from CBCT$_0$ to CBCT$_t$, and Eq. 5.3 represents the $J\text{det}$ of f:

$$J\text{det}(f) = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right]. \tag{5.3}$$

Following the registration process, broad contours of the head and neck are created with a binary threshold and morphological opening, in order to mask non anatomical structures in the scan. The generated mask is applied before each pass in the imaging branch of the model.

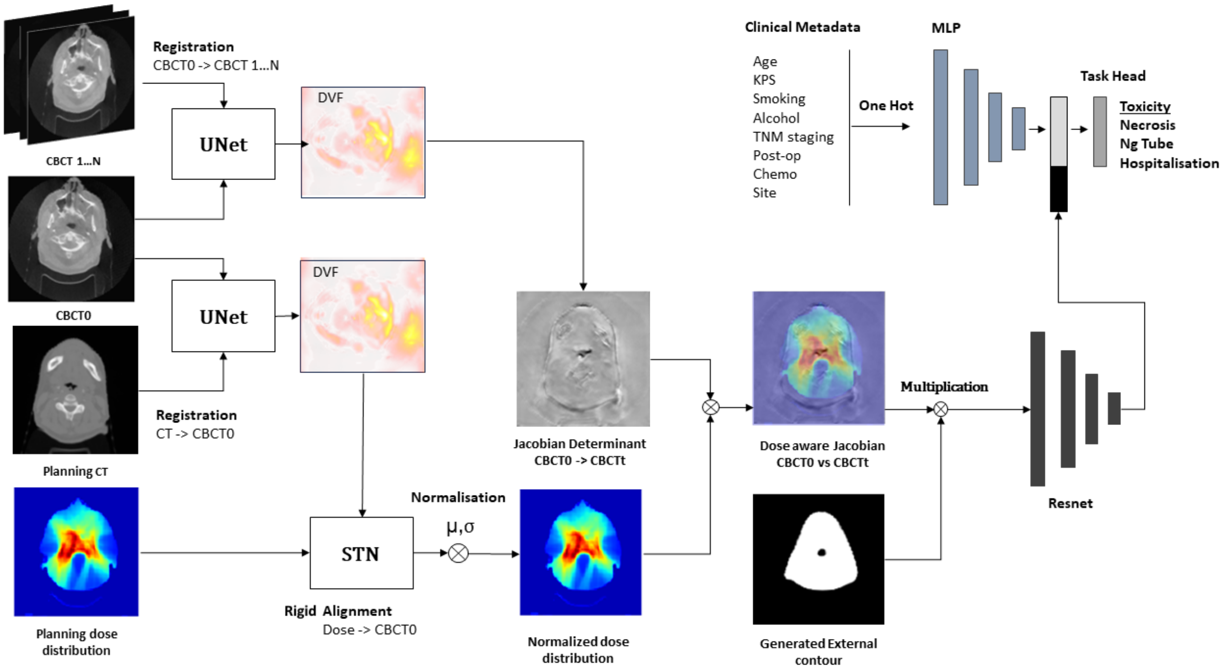


Figure 5.1 Proposed multi-modal deformable registration process and model architecture. A first UNet VoxelMorph [13] model processes deformation between CBCT$_0$ $\rightarrow$ CBCT$_t$, the resulting DVF converted to its Jacobian matrix. A second VoxelMorph processes the deformation between pCT $\rightarrow$ CBCT$_0$; the resulting DVF is applied to the dose distribution for alignment. Dose distributions are normalised with population statistics before multiplying the Jdet and generated contours. A ResNet50 [17] branch is trained on the resulting dose-Jacobian volume. A secondary branch consisting in a multi-layer perceptron (MLP) incorporates one hot encoded clinical data from subjects. Latent feature vectors are stacked before generating the toxicity outcomes.

5.5.3 Dose-aware models and training protocol

The proposed model shown in Fig. 1, is a 3D Residual network based on Torchvision's implementation. Specifically, all experiments use the ResNet50 (Bottleneck ResNet v1.5 base with 3D convolution and [3, 4, 6, 3] layering) architecture variant [17], pretrained on the MedicalNet database [61] to take advantage of pre-existing available medical imaging data. Dose plans and Jacobians matrices follow an early fusion scheme [62] to avoid redundant information, limit memory impact and focus attention on high dose locations. Dose plans are normalised between 0-1 using statistics computed on the entire population of doses (dosage $\sim \mathscr{Z}(0, 1)$), then multiplied by the pixel intensities of the Jacobians, producing deformation-dose joined pixel intensities. Higher pixel intensities correspond to a balance of high dosage and/or high deformation in regards to the study population. Clinical and imaging modalities follow a late fusion process with joined latent representations being horizontally stacked ($512 + 256$ features latent vector) before input into the last prediction linear layer(size: batchsize,768).

A baseline multi-layer perceptron (MLP) clinical model is implemented with three layers (64, 128, 256) consecutive dense blocks containing each a linear layer, batch normalisation and ReLU non-linear activation function. Dropout with ratio of 0.4 was applied between blocks.

The training process employs the binary cross-entropy with logits loss function, using the ADAM optimiser with a learning rate of 0.007 and a batch size of 8 for 100 epochs with early stopping of 10 epochs set on a patience of 10-4. To expedite the convergence rate of the model, the OneCycle learning rate scheduling policy as proposed originally by Smith et al. [63] is used. All models in this study are trained on a 2080Ti NVIDIA GPU, separately for each toxicity and experiment.

We apply a data augmentation scheme to prevent over-fitting, using the following parameters: random affine rotation by up to 20 degrees, image centred translation by up to 60mm in the axial plane and random flipping with probability of 0.5. To account for the high class imbalance during training and testing, we apply stratified 5-Fold cross validation to ensure each partition contains the same distribution as the original population, thus avoiding a potential undersampling of toxicities in the test set. We also apply positional weights to the loss function to re-scale the importance of toxicity-positive samples to the same apparition rate as negative samples.

5.6 Results

Qualitative results of the DVFs and dose plans are shown in Fig. 2 alongside their reference CBCT fraction. At fraction day 10, deformations appear mainly outside the dose covered area, with a tissue expansion concordant with chemotherapy induced swelling. At fraction 25, tissue swelling has reverted to its original state and high deformations appear in high dosage areas.

Table.1 displays the classification performance obtained for all source modalities available in this study, at the optimal fraction for each toxicity. Results show an accuracy improvement in predicting radionecrosis when incorporating dosimetry alongside clinical data over the clinical-only model ($70.2 > 69.3\%$). However, the joined dose-Jacobian model trained for fraction 20 deformation outperforms all other ablation models for this toxicity ($75.4 > 70.2\%$) suggesting that a combination of modality including dose, tissue deformation and clinical information is optimal for this toxicity. For NG tube at fraction 25 the dose-Jacobian model outperforms all modalities ($74.1 > 68.0\%$ clinical). Our results are coherent with previous observations and quantitative results presented as deformations in the anatomy take time to manifest, and competing effects from chemotherapy appear in first 2 weeks of treatment. Deformations in sensitive airway and feeding tracts appear later in treatment and are exploited by our model to improve the detection of toxicity. The model did not perform better than random using any imaging modality on the hospitalisation task with 53.1% and 55.4% for Jacobians and their dose normalised counterpart respectively. Hospitalisation remains extremely difficult to predict, with a low occurrence (4.2%) and a variety of multi-system global involvement to consider. Concurrent treatment, other diseases, undocumented patient preferences may result in particular course of actions. Furthermore, COVID-19 related modifications in treatment scheduling may have also influenced hospitalisation and in turn model accuracy.

Table. 2 presents the classification performance obtained on two different tumoural sites. Our results indicate improved results for all toxicities when including different sites, with ($74.3 > 66.7\%$) for radionecrosis when using Jacobians and doses, ($66.3 > 57.8\%$) when using dose distributions for NG tube and ($57.6 > 53.7\%$) for hospitalisation with dose distributions. Although toxicity rates may be higher and treatments more similar when focusing on oropharynx only cases, we obtain higher prediction rates for all toxicities when using the mixed cohort. These results suggest that the inclusion of multiple tumoural sites did not hinder the generalisation ability of our CNN and rather displays the adaptability of such models.

Fig. 5.3 represents the prediction performance of both $J\det$ only and $J\det$+dose models over the different fraction days selected for the $\text{CBCT}_0 \rightarrow \text{CBCT}_t$ deformation. Results indicate no significant changes for radionecrosis over the fraction days. Tissue necrosis is a long range affection, with symptoms occurring months after treatment; the temporal range

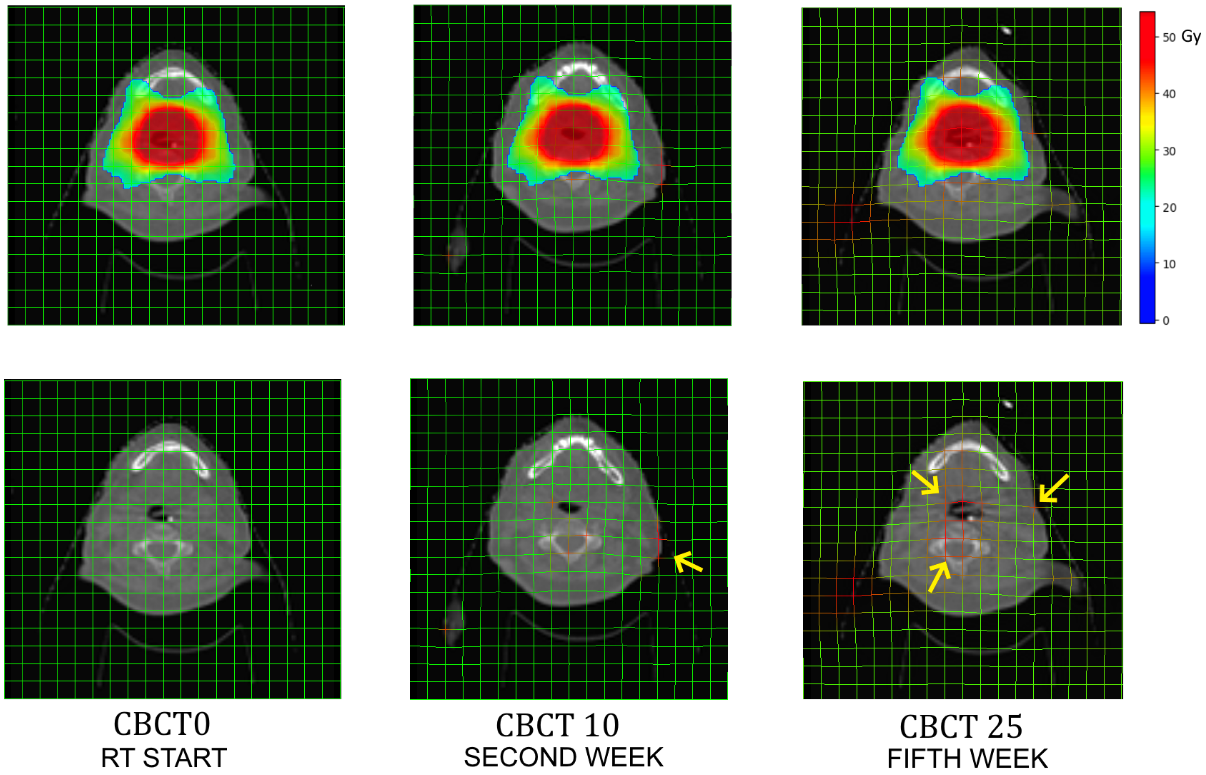


Figure 5.2 Sample CBCT fractions at different treatment weeks. *Top*: Superimposed CBCT and dose distribution. *Bottom*: CBCT and deformation grid corresponding to the DVF of $\mathrm{CBCT}_0 \rightarrow \mathrm{CBCT}_t$.

covered by CBCTs may not be able to capture significant changes in biomarkers during treatment. The NG tube toxicity displays a significant improvement in performance between fraction 10 and 25, with a maximum of 74.1% obtained after the 5^{th} week of treatment. This confirms our previous assumption with anatomical deformation as changes correlated to NG toxicity appear later in treatment. The dose aware Jacobian modality did not follow this steady improvement rate as expected, suggesting that initial dose plans show poor ability in correlating with long term deformations. Anatomical deformations appearing in areas with low planning dosage may have resulted in feeding impairment, which in turn were covered up by the multiplication with the dose matrix. It is also significant to state that re-planning was not included in this study.

Table 5.1 Classification performance on $n = 1012$ patients using CBCT deformations at the best fraction for each toxicity. Figures are presented as mean (SD) over a 5-fold stratified cross-validation. Baseline models include the clinical branch only, while multi-modal models combined both the clinical and imaging path in the ResNet architecture.

	Fraction	*Modality*		*Metric (% + SD)*		
Toxicity		**Dose**	J**det**	**Accuracy**	**Specificity**	**Sensitivity**
Hospitalisation	**10**			**61.1** ± **7.3**	**65.0** ± **18.4**	59.0 ± 20.6
		✓		57.6 ± 7.6	55.7 ± 9.4	59.7 ± 14.4
			✓	53.1 ± 2.7	61.6 ± 11.8	47.6 ± 17.6
		✓	✓	55.4 ± 7.7	54.4 ± 20.4	**61.8** ± **27.3**
NG tube	**25**			68.0 ± 1.7	56.6 ± 2.8	79.4 ± 4.9
		✓		66.3 ± 4.2	60.3 ± 7.9	72.2 ± 14.8
			✓	**74.1** ± **2.9**	**68.0** ± **4.4**	**80.5** ± **6.3**
		✓	✓	64.5 ± 3.7	54.3 ± 9.3	79.1 ± 11.0
Radionecrosis	**20**			69.3 ± 2.6	**71.9** ± **5.2**	67.7 ± 15.4
		✓		70.2 ± 7.6	62.4 ± 6.1	76.8 ± 15.5
			✓	69.0 ± 12.1	57.7 ± 6.8	80.3 ± 19.9
		✓	✓	**75.4** ± **5.6**	60.7 ± 3.4	**94.5** ± **9.6**

Table 5.2 Classification performance of clinical + imaging models on subsets of the initial cohorts by primary cancer site. The fraction used for all models is 10. Figures are presented as mean (SD) over a 5-fold cross-validation.

Toxicity	**Cancer site**	**Incidence (%)**	J**det**	**Accuracy (%)**
Hospitalisation	Mixed	4.2	✓	55.4 ± 7.7
				57.6 ± **7.6**
	Oropharynx	5.2	✓	53.7 ± 12.5
				45.9 ± 8.6
NG tube	Mixed	17	✓	62.0 ± 3.8
				66.3 ± **4.2**
	Oropharynx	23	✓	57.8 ± 5.2
				57.1 ± 6.5
Radionecrosis	Mixed	4.56	✓	**74.3** ± **6.0**
				57.6 ± 7.6
	Oropharynx	6.5	✓	66.7 ± 10.0
				65.9 ± 8.2

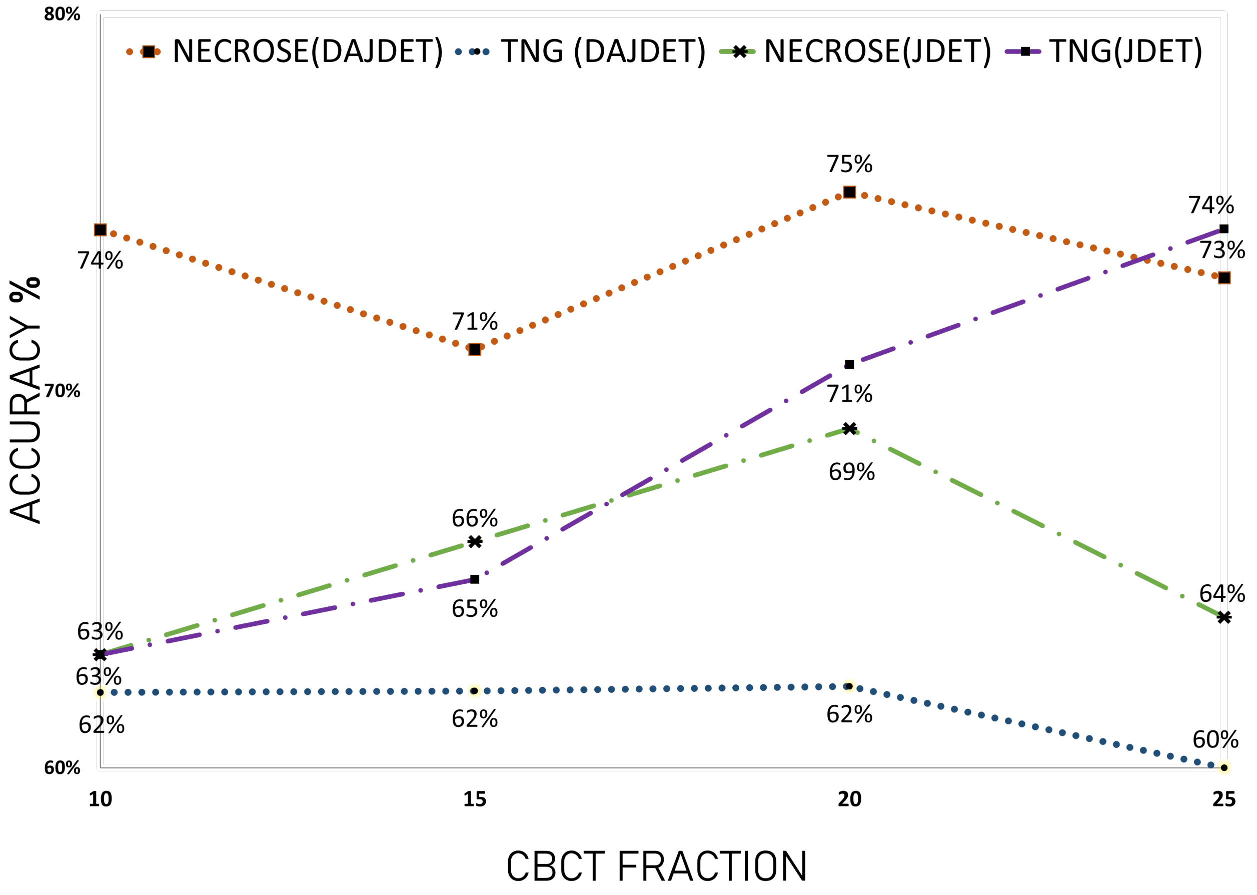


Figure 5.3 Accuracy of toxicity prediction models in function of the daily fraction used to compute the CBCT deformations.

5.7 Conclusion

In this study, we explored the significance of dosimetric data in the prediction of radiotherapy toxicity for individuals with head and neck cancer. We also examined its potential when combined with the daily monitoring of anatomical deformations using longitudinal CBCTs. While raw dose plans demonstrated predictive capabilities for toxicity, our findings encourage the adoption of a multi-modal approach, combining toxicity relevant information from multiple sources. Considering the inclusion of additional soft tissue modalities such as MRI could prove advantageous in future investigations. This warrants further exploration into modality fusion to enhance research outcomes and prevent redundancy. Limitations of this study are the omission of re-planning strategies and the use static deformable registration strategy, potentially preventing significant inter-fraction deformations from being detected. This suggests that the integration of a serial temporal prediction process could further improve performance.

5.8 Acknowledgements

This research has been funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors have no relevant financial or non-financial interest to disclose.

5.9 Compliance with ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the CHUM's Ethics Committee. (Date : April the 26th 2019)

CHAPTER 6 ARTICLE 2 : AMO-ENE ATTENTION BASED MULTI OMICS FUSION MODEL FOR OUTCOME PREDICTION IN EXTRANODAL EXTENSION AND HPV ASSOCIATED OROPHARYNGEAL CANCER

Gautier Hénique[1,2], William Le, BSc[1,2], Gabriel Dayan, MD[2,3], Coralie Brodeur[1], Edith Filion, MD[3], Phuc-Felix Nguyen-Tan, MD[3], Laurent Letourneau-Guillon, MD, MSc[2,3], Houda Bahig, MD, PhD[2,3], Samuel Kadoury, PhD[1,2]

[1] *MedICAL Laboratory, Polytechnique Montréal, Montreal, Canada*
[2] *Centre de recherche du CHUM (CRCHUM), Montréal, Canada*
[3] *Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, Canada*

6.1 Publication

This paper has been submitted to Medical Image Analysis on the 24rd June 2025 and is pending reviews.

6.2 Contribution

My contributions to this work covers all the coding, planning, experiments, and writing.

6.3 Abstract

Extranodal extension (ENE) is an emerging prognostic factor in human papillomavirus (HPV)-associated oropharyngeal cancer (OPC), albeit it is currently omitted in clinical staging criteria. Recent works have advocated for the inclusion of iENE as a prognostic marker in HPV-positive OPC staging. However, several practical limitations continue to hinder its clinical integration, including inconsistencies in segmentation, low contrast at the periphery of metastatic lymph nodes in CT imaging, and the labor-intensive manual annotations. To address these limitations, we propose a fully automated end-to-end pipeline leveraging computed tomography (CT) images with clinical data to evaluate nodal ENE status and predict the treatment outcomes. Our approach begins with a semi-supervised segmentation model designed to detect and delineate relevant iENE from radiotherapy planning CT scans.

From these segmentations, a set of radiomics and deep features are extracted to train an imaging-detected ENE grading classifier. The predicted ENE status is then evaluated for its prognostic value and compared against existing staging criteria. Furthermore, we integrate these nodal features with primary tumor characteristics in a multi-modal, attention-based outcome prediction model, providing a dynamic framework for outcome prediction. Our method is validated on an internal cohort of 397 HPV-positive OPC patients treated with radiotherapy or chemoradiotherapy between 2009 and 2020. The segmentation model achieved a mean Dice score of 74.4% (± 25.0), which improved to 80.5% (± 8.1) following component selection. The ENE classification model yielded an AUC of 79.92 % (± 2.51) for general iENE detection, and 89.87% (± 5.24) for identifying high- grade (grade 3) ENE. For outcome prediction at the 2-year mark, our pipeline surpassed baseline models with AUCs of 86.7% (± 7.4) for metastatic recurrence, 72.6% (± 9.6) for overall survival, and 78.1% (± 8.6) for disease-free survival. We also obtain a concordance index of 83.3% (± 6.5) for metastatic recurrence, 70.2% (± 10.9) for overall survival, and 70.0% (± 8.1) for disease-free survival.

Keywords : Oropharyngeal Cancer Extra Nodal Extension Lesion segmentation Survival Analysis Attention-based multi-omics fusion

6.4 Introduction

Extranodal extension (ENE) in oropharyngeal carcinoma (OPC) refers to the spread of malignant cells beyond the capsule of metastatic lymph nodes [64], and is an indicator of tumor aggressiveness, as well as of a greater risk for invasion into adjacent tissues. Traditionally, ENE is confirmed histopathologically (pENE) through detailed examination following lymph node dissection, a time-intensive process. Presence of ENE is associated with escalated therapeutic regimens, including intensification of adjuvant and chemo-radiation protocols, as well as closer surveillance protocols [65], which, while aiming to curb disease progression, also increases the burden of treatment-related toxicity on patients. However, a significant proportion of patients with OPC are currently treated with upfront chemoradiation without undergoing neck dissection, resulting in the absence of formal pathological lymph node evaluation, thus motivating the use of imaging as a surrogate marker of ENE.

To expedite the evaluation process, imaging-based gradation systems for imaging detectable ENE (iENE) on head and neck CT scans have been developed in recent years, such as the five grade scale proposed by [65–67]. However, these lack standardization, and their clinical utility is hindered by inter-observer variability and inconsistent segmentation boundaries. Figure 6.1 illustrates the four-tier iENE classification framework: grade 0—no radiologic evidence of ENE; grade 1—minor extension (through the capsule into surrounding fat); grade 2—coalescing matted nodal mass with disappearance of inter-nodal planes; and grade 3—gross invasion into surrounding anatomical structures (adapted from [66, 67]).

Therefore, despite increasing evidence advocating for the incorporation of ENE status into the staging of HPV-associated OPC, numerous challenges ranging from diagnostic inconsistency to a lack of universal criteria continue to impede its widespread and reliable integration into oncologic practice.

In recent years, there has been growing interest in leveraging machine learning (ML) techniques to extract valuable prognostic insights from medical imaging, providing a way to extract predictive information beyond subjective image interpretation and potentially enhancing clinical guidelines with more precise and reliable image analysis. By leveraging the complementary predictive value of multiple modalities and their interactions, patient-specific treatment strategies could be integrated and result in reduced toxicities ([68]).

In this paper, we propose AMO-ENE a fully automated pipeline using an attention-based multi-organ fusion model for iENE evaluation based on planning radiotherapy head and neck CT images. Specific subtasks include nodal segmentation, grade classification and treatment outcome prediction.

This study's contributions include the following:

- Automation of pathological node segmentation on radiation planning CTs, comparing state-of-the-art methods, and subsequent iENE grade classification based on complementary feature sets.
- Introduction of an attention-based multi-omics fusion model for 2-year risk assessment in HPV-associated OPC treated with radiation therapy.
- Adaptation of the proposed model to multi-bin risk modeling for a deep learning based-longitudinal risk estimation of outcomes in HPV-associated OPC.
- Evaluation of extra nodal extension prognostic value using statistical survival analysis on a cohort of 397 OPC patients.

6.5 Related Works

6.5.1 Lesion Segmentation

Medical image segmentation is a popular task in oncological research, as robust tumor detection and delineation may lead the way for automated screening and treatment planning.

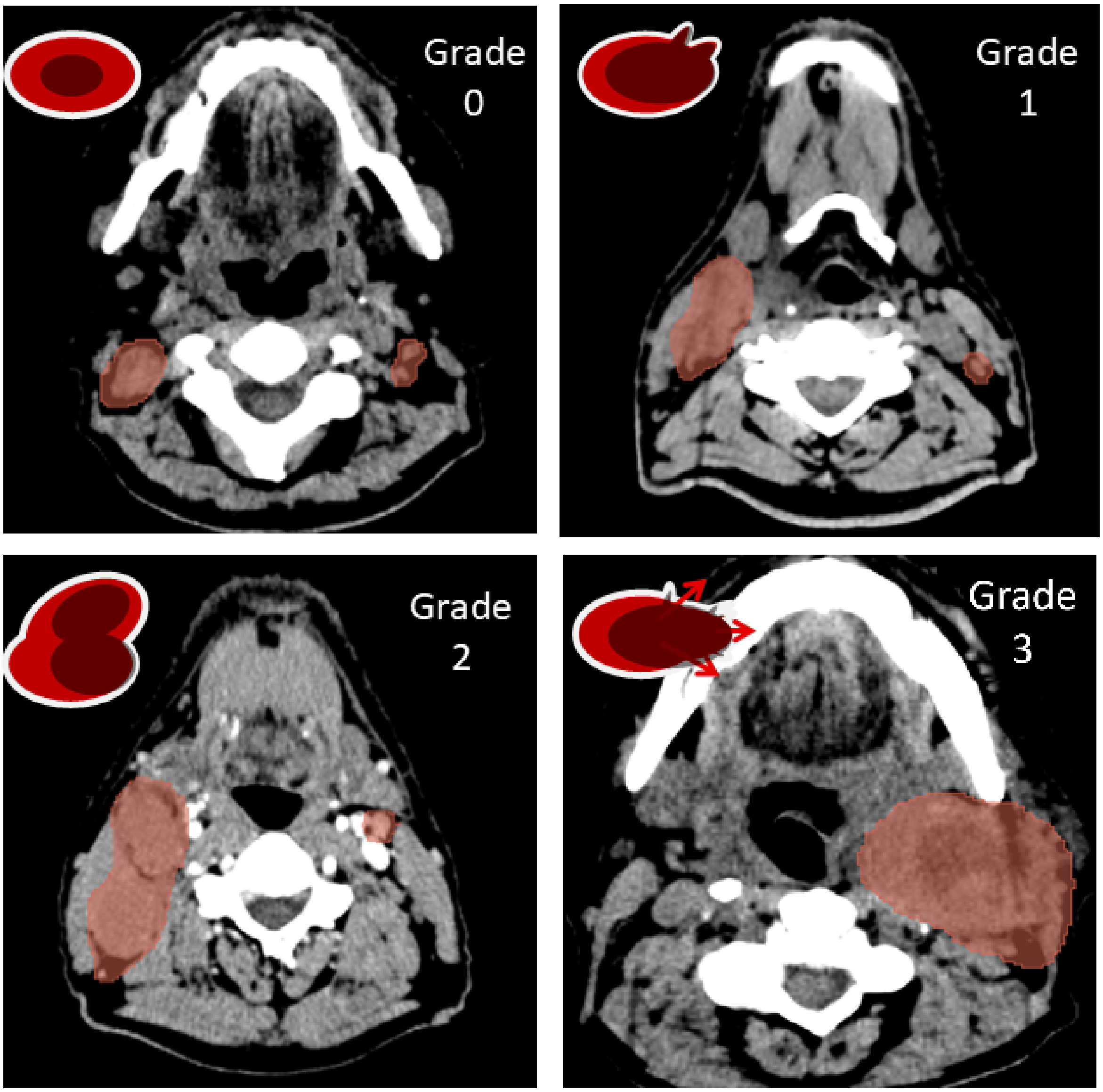


Figure 6.1 Illustration for iENE status grading system. Nodes are presented with white borders, red healthy tissue background and black tumoral content. Grade 0 indicates metastatic pathological node with no signs of extension. Grade 1 indicates extension through the nodal capsule into surrounding fat. Grade 2 characterizes coalescent nodal masses with loss of internodal planes. Grade 3 is defined by overt invasion of surrounding structures and tissues. In case of multiple ENE, the maximum grade is reported.

In recent years, several approaches were proposed to automatically retrieve organs at risk (OARs) from planning images [37], however they often exclude nodal structures in this category. In addition, there are no methods to our knowledge for automated extra nodal extension segmentation in the literature, which may be explained by the lack of iENE status integration in clinical guidelines for HPV-associated OPC and by the labour intensive annotation task it demands.

Since its introduction in 2012 [2], deep convolutional neural networks (CNN) have established themselves as a standard for image analysis. Adapted to the popular encoder-decoder "U" style architecture [11], it varies to different field of views by harnessing the strong spatial awareness of convolution masks in an encoder-decoder fashion for 3D segmentation prediction. The 2022 HECKTOR MICCAI challenge [28] demonstrated the value of CNN models for nodal lesion segmentation, [69] proposed a residual CNN for this task and achieved the best overall results. In addition, [70] provided a nnUnet pipeline [18] for gross tumor volume (GTV) and nodal lesion segmentation. This self - configuring U-shaped convolutional neural network, known for its reliability and state-of-the-art performances, allowed for automated parameter tuning and preprocessing, often serving as a strong segmentation baseline in comparative studies.

Vision Transformer (ViT) [71] [72] segmentation architectures, with their ability to model long-range dependencies, have been proposed as an alternative to convolution networks, leveraging attention between input image patches treated as tokens. SwinUNETR [19] has been proposed as a U-shaped iteration of the ViT, integrating a shifted windowed attention mechanism allowing multi-scale refinement of this mechanism to capture finer details. SwinUNETRV2 [12] is an incremental modification of the aforementioned network, with an hybrid CNN-Vit encoder, combining the strength of both mechanisms with reduced data requirements.

Datasets with segmentation annotations from large cross institutional cohorts have recently led to the popularity of foundation models, which can leverage large scale pre-training datasets for efficient fine-tuning capability. The generalist segment anything model (SAM) [40] was adapted to a medical imaging oriented 3D version in SamMed3D [20], a prompt based 3D Foundation model pre-trained on 143 thousand segmentation masks. Foundation models offer the ability to segment organs accurately via the given prompt [**?**] [**?**], even for previously unseen structures. However SamMed3D requires location prompts as input, effectively skipping the detection task inherent to segmentation.

6.5.2 Node and extranodal extension grade classification

Several approaches have explored the characterization of lymph node lesions in OPC, albeit not considering iENE as a primary target. In addition to the lack of ENE segmentation models, there are still no imaging based extra nodal extension classification model in the literature.

Kann et al. [73] developed a three-dimensional convolutional neural network (3D-CNN) to detect pathologically confirmed extranodal extension on head and neck CT scans across various head and neck cancer (HNC) subtypes. While their work also focused on the automated ENE prediction, authors utilized pENE—determined via histopathological analysis as the ground truth, whereas current efforts have explored the consensus of radiologists for imaging-detected ENE (iENE) as the target variable, as neck dissection are seldomly performed as the initial therapy. Although pENE offers a more definitive standard by capturing microscopic disease indicators not always apparent on imaging, its clinical relevance in HPV-positive oropharyngeal squamous cell carcinoma (OPSCC) is often debated, with evidence suggesting that microscopic pENE may not significantly influence outcomes [74, 75]. In addition Kann et al. did not assess whether the proposed AI-predicted pENE correlated with clinical endpoints.

The Hecktor 2022 challenge [28] catalyzed further exploration into the integration of nodal radiomics and multi-omic data for enhanced outcome prediction. Notably, the top-performing team [70] achieved a concordance index (C-index) of 0.68 for progression-free survival prediction in HNC. Their approach involved the extraction of 1,209 features from both primary and nodal tumor segmentation, subsequently analyzed through a bagged binary weighted model. Other contributions in the challenge investigated the fusion of segmentation data with radiomic descriptors [76]; however, these methods did not surpass the predictive performance of radiomics alone. While these studies confirm the predictive benefit of combining features from primary and nodal lesions, they notably omit extra nodal extension, a clinically relevant factor in prognosis.

6.5.3 Survival Analysis for OPC

Survival analysis is a statistical framework aimed at estimating risk scores associated with patient outcomes, primarily by forecasting time-to-event variables, such as disease progression or mortality.

In the application to HNC, radiomic features have demonstrated significant prognostic value in survival prediction tasks [27]. These characteristics, extracted from medical images, have proven effective in characterizing the morphological and textural attributes of primary tumor volumes, thereby offering predictive insights into adverse treatment outcomes. Several methods have been proposed to harness the integration of radiomic data, baseline logistic regression models [77] with traditional lasso-based feature selection, non parametric bagged estimators with sign attribution to each feature [70] or other random forest algorithms with recursive feature selection [27]. More recent approaches have integrated radiomic features in deep networks [35], alleviating the curse of dimensionality with strong regularization and higher non-linear parameter spaces.

Emerging methodologies have begun to shift focus away from handcrafted radiomic features, exploring the latent potential of non-explicit representations.

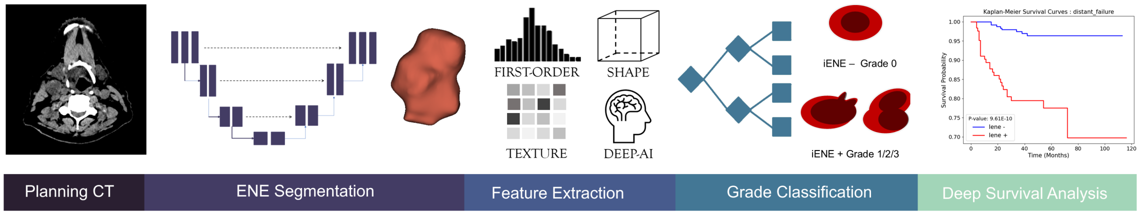
Vision-based deep learning models have been explored as salient features extractors from images [43], allowing for a deeper exploration of hierarchical biomarkers. In particular, foundation models using self-supervised learning (SSL) strategies [10] have been proposed. These models leverage vast amounts of unannotated data to learn robust, high dimensional representations without the need for manual feature engineering. While they offer a promising avenue for generalizable and multi-task frameworks in biomarker discovery, their opaque nature as “black boxes” makes the interpretability of their outputs limited. This is of primary concern in safety critical medical applications. Importantly, their application in integrating nodal characteristics, particularly in HPV-associated oropharyngeal carcinoma (OPC), remains unexplored.

Modeling time-to-event outcomes is often performed using Cox proportional hazards regression, where hazard ratios derived from input features serve as predictors of event timing. While Cox regression remains a widely used and effective method for survival analysis, it relies on the proportional hazards assumption and can be sensitive to high-dimensional or correlated features [78], which may limit its applicability in complex clinical datasets.

To address these limitations, Multi-Task Logistic Regression (MTLR) was introduced by [46] as a discrete-time survival modeling framework that enables flexible risk estimation without relying on the proportional hazards assumption. MTLR models the probability of surviving each time interval as a sequence of logistic regressions, thereby allowing to adapt to various heterogeneous data types. Its effectiveness has been demonstrated in clinical contexts, including head and neck cancer prognosis [47], where it outperformed traditional survival analysis methods.

6.6 Materials and Methods

In this section, we describe the patient cohort characteristics, the overall prediction framework for iENE and its individual components. Figure 6.2 presents the overall CT to outcome pipeline. In Sec. 6.6.1 we report the detailed patient characteristics and inclusion/exclusion criteria. Section 6.6.2 describes the iENE segmentation module implementation and training parameters, while in Sec. 6.6.3 we list the different feature extraction methodologies and the subsequent ENE grade classification experiments, followed by the survival analysis methodology (Sec. 6.6.3) and the multi-organ attention based outcome prediction model. Finally, in Sec. 6.6.4 we report on the imaging and experimental setup used in each step of this study.



Proposed pipeline for lymph node segmentation, classification of iENE status and outcome prediction using pre-radiation therapy planning CT scans

Figure 6.2 Overall schematic diagram of the proposed framework for CT-based iENE status and prognosis prediction. First, an automated segmentation of pathological node allows for the retrieval of the largest metastatic lymph node. Afterwards, radiomics and deep features are extracted from the structure to classify iENE grades in a binary problem. Finally, from the obtained predictions, the survival characteristics of predicted iENE groups are compared to demonstrate the influence of extra nodal extensions in HPV-associated OPC

6.6.1 Clinical Dataset

Following institutional review board approval, we collected a retrospective cohort treated at Centre Hospitalier de l'Université de Montréal (CHUM) for HPV-associated OPC between 2009 and 2020. The study cohort comprises 397 patients diagnosed with oropharyngeal cancer, with a median age at diagnosis of 62 years (range: 39–84). The HPV status was clinically established with a positive p16 status upon biopsy of either the primary tumor or cervical lymph node. Exclusion criteria includes non curative intent chemo-radiation, negative HPV status, non cN+ staging and transoral robotic surgery (TORS) treated patients.

A vast majority of patients were male (317 males vs. 80 females). Median number of smoking pack-years was 8 (range: 0–100), with 46 % current smokers and 67% reporting prior exposure to tobacco. Tumor staging, based on the AJCC 8th edition [64], indicates that 25.6 % of patients had T1 tumors, 38 % had T2, 24 % had T3, and 12.4 % had T4. Nodal staging showed a predominance of N1 disease (311 patients), followed by N2 (67 patients) and N3 (19 patients). Concurrent systemic therapy was administered to the majority of the cohort (333 patients).

Most patients had between 1 to 4 abnormal lymph nodes (298 patients), while 99 patients presented with 5 or more. Abnormal retro pharyngeal lymph nodes larger than 8 millimeters were observed in 44 cases. Radiological assessment of extra nodal extension (iENE) showed that 271, 26, 75, and 25 patients were scored as iENE 0, 1, 2, and 3 respectively.

For all patients, the following information was collected and represented the clinical variables incorporated in the prognosis prediction model:

> [noitemsep]Age at diagnosis: Patient age before treatment planning. Patient sex: Identified sex at diagnosis. ECOG [79] score: A standardized measure of cancer's impact on patient functional status. Smoking pack-years: Measures lifetime tobacco exposure TNM 8th: Clinical AJCC staging [64] for Tumoral, nodal and metastatic components. Concurrent Systemic Chemotherapy: If the patient undergoes chemo-radiation. Type of Concurrent chemotherapy: Protocol used in the chemotherapy.

For each patient, the planning pre-treatment CT scan and the associated radiation oncologist segmentations of gross tumor volume and iENE were collected.

Therapeutical outcomes were collected at study date, leading to the report on the population observed events in the following Table 6.1:

Table 6.1 Outcome representation in the patient cohort for Uncensored (Event presence at the latest available follow up) and censored patients.

Outcome	Event type	Male	Female	Total
OS	Uncensored	27	11	38
	Censored	290	69	359
DM	Uncensored	30	5	35
	Censored	287	75	362
DFS	Uncensored	52	16	68
	Censored	265	64	329

Here, overall survival (OS) was defined by mortality from any cause, distant metastasis (DM) by the failure at distant sites from the head and neck region, disease-free survival (DFS) by any recurrence or death at follow up. Time to event was calculated from date of diagnosis. The mean and median follow-up were 47.0 (±22.3) months and 44.4 (Inter-quartile range 32.7-61.1).

iENE Annotation Protocol Ground truth imaging-detected extra nodal extension (iENE) was graded and segmented independently and blinded to clinical outcomes by two board-certified neuroradiologist. In addition, fifty five head and neck CT cases were independently assessed by three raters to rate agreement on decisions.

The distribution of graded iENE presents as a heavily unbalanced problem, with 69.4%, 5.9%, 18.8% and 5.9%, proportion for grades 0 to 3 respectively.

6.6.2 Supervised ENE Segmentation

To perform the segmentation of ENE's on CT scans, we employed the nnUnet framework [18], a collection of U-shaped CNN networks with automated processing and augmentation transformations selection. We selected the 3D residual encoder model (M version) as the backbone, and trained it on the entire head and neck region of interest (ROI), processed in a sliding patch window process, in order to segment all pathological node extensions.

By leveraging 3D convolutions at varying depths, the encoder-decoder CNN architectures leverages local and global features, enabling precise object detection and border delineation. This process is particularly relevant for nodal extensions as precise local decisions are required to accurately evaluate the spread of tumoral content.

We trained the binary semantic segmentation model using the soft Dice Score (DSC) loss as described in Eq. 6.1:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{I} Y_i \hat{Y}_i}{\sum_{i=1}^{I} Y_i + \sum_{i=1}^{I} \hat{Y}_i + \epsilon} \tag{6.1}$$

where Y_i represents the ground truth, $\hat{Y}_i$ the predicted voxels for each patient $i \in I$, and ϵ (1×10^{-8}) a stability term.

6.6.3 Feature extraction and classification

From the generated segmentation masks for extra nodal extension, we aim to extract specific features allowing the automatic and robust identification of the extension status in patients.

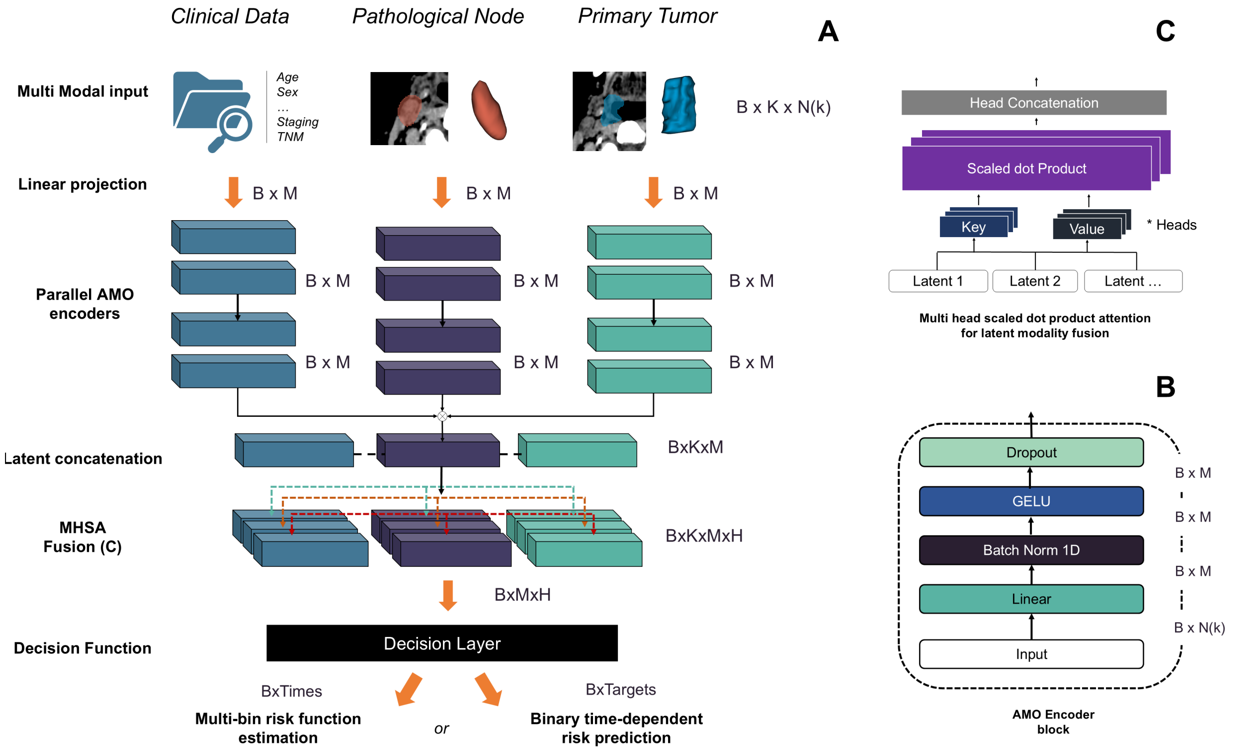


Figure 6.3 Graphical representation of the proposed attention based modality fusion pipeline. In (A), multiple modality-specific feature encoders are trained in parallel, to extract features from each specific modality. The latent representation of each modality is concatenated and a multi-head scaled dot product attention is performed to evaluate interactions and combine latent representations. A final decision layer is adaptable to either predict a time dependent outcome target or multi-bin risk modeling. Refer to 6.6.3 for the relevant notations. In (B), the proposed modality encoder module is shown, which is composed of multiple blocks containing parametric layers. (C) Detailed description of the operation performed in the attention layer.

Dichotomization schemes ENE grading presents itself as a heavily unbalanced problem with 4–10 times more class 0 than the other three and edge cases presenting mostly between classes 0-1 and 2-3 further increase potential variability in the dataset. These edge cases result mainly from ambiguous findings. As such, we propose three distinct dichotomization schemes, as opposed to a multi-class problem.

We refer to the 0 vs 1-2-3 scheme as iENE^{-} vs iENE^{+}, and it is the basis to all further model development with outcomes. This case maximizes the retention of ENE nodes, regardless of their severity. We also provide results for the 0-1 vs 2-3 classification scheme, which may represent a more conservative gradation of ENE, especially relevant in potentially indecisive radiologist annotations. Finally, we also classify grades 3 against other grades, putting greater emphasis towards detecting aggressive features and nodal extensions.

Feature extraction From the segmented ENE node and ground truth GTV masks, we propose two distinct feature extraction approaches based on hand-engineered characteristics and deep self-supervised features to classify and anticipate outcomes. For the hand-engineered approach, we extract radiomic features from the CT with the Python PyRadiomics [48] package. We set intensity bins to 10HU and resample instances to a fixed isotropic $1 \times 1 \times 1$ mm spacing. Computed features are obtained by selecting the first-order, shape, and gray-level co-occurence matrix (GLRLM, GLSZM, GLDM and GLCM) features (sum average disabled). A total of a 100 features were computed describing the shape of the predicted node, the intensity statistics, and high level texture maps. For the self-supervised approach, we extract deep features obtained from a pre-trained foundation model [10] for cancer imaging biomarkers (referred to as FMCIB). This method extracts 4096 features in the latent space of the self-supervised model which was pre-trained on 11 000 lesions. A fixed region of interest of size $50 \times 50 \times 50$ mm is extracted around the center of mass of the predicted iENE segmentation and fed to the model. Prior to inference, all scans are resampled to isotropic $1 \times 1 \times 1$ mm spacing and normalized by subtracting -1024 and dividing by 3072 (lower and upper bounds provided by the FMCIB model [10]).

Grade classification Using the extracted features, we construct a modular binary classification pipeline with components that were systematically evaluated through a grid search (see Section 6.6.4 for all evaluated hyperparameters).

All tabular data underwent standard scaling based on the respective training set distributions. We trained machine learning classifiers such as the Random Forest (RF), the XGBoost algorithm [80] and a deep learning based multi-layer perceptron (MLP).

For each method, we tested out principal component analysis (PCA) and Lasso regression for feature selection. This step is crucial as both radiomic and FMCIB features present high co-linearity, increasing noise during model training.

To handle class imbalance, we employ SmoteTomek [49] generation of training samples, which allows to balance both classes while preserving variance in the under-represented class. Grid search parameter evaluation was performed for each combination of classifier and selection algorithm, where parameters from both components in each cross-validation fold were optimized to prevent data leakage.

Attention-based Multi-modal Outcome (AMO) risk prediction model

In order to combine our multi-modal radiomic feature sets, we propose AMO-ENE, a scalable multi-branch attention-based deep classifier module, allowing for the independent unimodal representation extraction of biomarkers paired with the fusion of these modalities in order to assess interactions and their relative importance with respect to progonostic prediction.

Furthermore, to explore and integrate all available modalities, namely clinical features and nodal radiomic/deep features extracted in Section 6.6.3, we integrate primary tumor characteristics using the same radiomic and foundation model pipeline. This evaluation is motivated by the hypothesis that nodal features are complementary with primary tumor characteristics in their discriminative ability to assess the metastatic potential of HPV positive OPC.

Let $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(K)}\}$ denote a set of K input modalities, such as radiomic features extracted from multiple organs or lesions. Each modality $X^{(k)} \in \mathbb{R}^{n_k}$ represents a feature vector of dimensionality n_k. The multimodal representation is derived via the following steps:

1. Biomarker Extraction Each modality is processed by a modality-specific neural encoder $f^{(k)}$ to project it into a shared latent space of dimension M:

$$C^{(k)} = f^{(k)}(X^{(k)}) \in \mathbb{R}^{M}, \quad \forall k \in \{1, \ldots, K\}. \tag{6.2}$$

Each encoder $f^{(k)}$ is implemented as a two-layer feedforward neural network. The architecture consists of a first linear transformation mapping from $\mathbb{R}^{n_k}$ to $\mathbb{R}^{M}$, followed by batch normalization, a GELU (Gaussian Error Linear Unit) activation, and dropout for regularization. This consistent structure ensures each modality is independently projected into a common latent space while mitigating overfitting and capturing non-linear relationships. These encoders serve the dual role of feature selection and dimensionality reduction, yielding modality-specific latent embeddings $C^{(k)}$.

2. Latent Concatenation The latent embeddings from all K modalities are then vertically stacked to form a matrix $C \in \mathbb{R}^{K \times M}$:

$$C = \begin{bmatrix} (C^{(1)})^T \\ (C^{(2)})^T \\ \vdots \\ (C^{(K)})^T \end{bmatrix}. \tag{6.3}$$

Here, each row of the matrix C corresponds to the latent embedding of a specific modality, where $C^{(k)} \in \mathbb{R}^M$ is transposed to align with row-major stacking. The resulting matrix C thus encapsulates modality-wise biomarker representations, where each modality contributes one row of M latent features.

3. Multi-Head Self-Attention Fusion (MHSA) To capture dependencies between modalities and enhance contextual interactions, we employ a self-attention fusion mechanism based on multi-head attention followed by residual connections and layer normalization.

Given the stacked modality representations $C \in \mathbb{R}^{K \times M}$ where each of the K rows corresponds to a latent embedding of a modality in the M dimensional latent space the self-attention layer operates as follows. The input C is treated as the set of queries, keys, and values:

$$\hat{C},\ A = \text{MultiHeadAttention}(Q = C, K = C, V = C), \tag{6.4}$$

where $A \in \mathbb{R}^{K \times K}$ denotes the attention weight matrix capturing pairwise dependencies among modalities, and $\hat{C}$ the attention output of shape $\mathbb{R}^{K \times M}$.

A residual connection is applied between the attention output and the original modality embeddings, followed by layer normalization to stabilize training. The resulting representation is further refined using a two-layer feedforward network with GELU activation and dropout.

4. Global Fusion for Downstream Tasks After attention-based interaction and refinement, we aggregate the modality-enhanced embeddings to produce a single unified representation via average pooling across modalities:

$$\bar{Z} = \frac{1}{K} \sum_{k=1}^{K} \hat{C}_k \in \mathbb{R}^M, \tag{6.5}$$

where $\hat{C}_k$ is the k-th row of the final attention-refined matrix C. The resulting vector $\bar{Z}$ captures the integrated multimodal information and serves as input to downstream predictive modules (e.g., classification or regression heads).

Figure 6.3 presents an illustration of the proposed module. Each feature-set is given as input to its dedicated encoding branch, composed of two successive extractor blocks that combine high level features to a task specific biomarker latent. Each modal representation is concatenated and fused through a multi-head scaled dot product attention layer, allowing the retrieval of task specific interactions and serving as an attention weighted feature selection mechanism.

The decision layer of the model can be adapted to multiple prognosis prediction paradigms, we thus propose two distinct cases in the following paragraphs describing binary time assessment, and multi-bin risk modeling.

Multi-modal 2-year risk/prognosis prediction The model was trained to predict the 2-year landmark, where time to outcome was determined from the treatment start date.

Using as input to the three distinct branches the nodal extension radiomics, the GTV radiomics and the clinical data, we predicted the presence (1), absence or censoring (0) of OS, DM, and DFS as binary oncological outcomes.

We optimize our model with a weighted binary right censored cross entropy loss as described in Eq. 6.6:

$$L_{cc} = \sum_i \left[\begin{array}{l} \mathbb{1}_T \cdot w_0 \cdot \ln \left(f \left(X_i \right) \right) \\ + \left(1 - \mathbb{1}_T \right) \cdot w_1 \cdot \ln \left(f \left(X_i \right) \right) \end{array} \right] \tag{6.6}$$

where the indicator function $\mathbb{1}_T$ represents outcome presence at time T (two year in our case). Weighting parameters w_0 are computed by obtaining the inverse population ratios relative to each class on the relative training set of each instance.

Multi-bin risk modeling for survival analysis Beyond binary classification, we extend our multi-modal model to perform survival analysis through a discretized, multi-bin risk regression framework. This is achieved by replacing the classifier head with a Multi-Task Logistic Regression (MTLR) module [46]. Patient survival times are discretized into T quantile-based intervals derived from the empirical survival distribution of the training cohort. The model then outputs a sequence of time-specific logits, each representing the log-risk score for a given time interval.

Training is performed by minimizing the MTLR negative log-likelihood loss, defined as:

$$\begin{aligned} \mathcal{L}_{\text{MTLR}} = &- \sum_{i \in \mathcal{I}_c} \log \left(\sum_{t : y_{it} = 1} e^{z_{it}} \right) \\ &- \sum_{i \in \mathcal{I}_u} \sum_{t=1}^{T} y_{it} z_{it} \\ &+ \sum_{i=1}^{N} \log \left(\sum_{t=1}^{T} e^{z_{it}} \right), \end{aligned} \tag{6.7}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times T}$ are the model's predicted logits over T time bins for N patients, $\mathbf{Y} \in \{0,1\}^{N \times T}$ are the encoded survival targets, and $\mathcal{I}_c$, $\mathcal{I}_u$ represent the sets of censored and uncensored samples, respectively. This formulation enables the model to learn a smooth, time-dependent hazard function, capturing the changing risk of event occurrence over time. During inference, the predicted logits can be converted into cumulative survival probabilities, enabling individualized survival curve estimation.

6.6.4 Experimental Setup

All models and experiments were carried out on a single Nvidia A6000 GPU with 48GB of available VRAM. Segmentation experiments were performed on a 12GB RTX Titan card with batch size scaling. Experiments were carried out in pytorch version 2.1.0 and python 3.11.5. Size occupied on disk by the dataset of 397 cases was 20 GiB.

iENE Segmentation

We initially trained the nnUnet [18] model for 750 training epochs, which allowed for convergence of the model using the internal multimodal our dataset. We trained the model with a stochastic gradient descent optimizer with an initial learning rate of 0.01, 1×10^{-5} weight decay factor and the polynomial scheduler. Given the characteristics of our dataset, namely the varying number of nodes to segment and the size discrepancy, we adapted the performance evaluation in our segmentation task to account for possible bias towards larger nodes.

The largest node is the main region of interest in the imaging grading process and the most relevant for prognosis evaluation, and as such, we report segmentation results based on the largest predicted and annotated nodes.

Results were evaluated using the mean Dice score obtained over 5 fold cross validation splits. Given Y the ground truth and $\hat{Y}$ the predicted voxels, we extract the largest components from each masks and compute their Dice as:

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{6.8}$$

where TP, FP and FN are the true positive, false positive and false negative number of voxels.

Segmentation models were implemented within the Monai [81] framework, with Pytorch-lightning for acceleration [82].

iENE classification

The classifier for nodal extension grades were evaluated in a 5-fold stratified cross-validation scheme, where each split obtains similar grade distributions as the overall population. We evaluated the different models with the mean area under the receiver operating characteristic curve (AUC) metric computed over the test folds, as it allows for a fine logit threshold selection, maximizing the respective recall and specificity contribution of each model.

Grid search parameter evaluation was performed for each combination of classifier and selection algorithm. Searched parameters included: number of estimators $\in \{50, 100, 200, 500, 1000\}$; tree depth $\in \{3, 5, 7\}$; initial learning rates $\in \{0.01, 0.1, 0.2\}$; L1 ratio $\in \{0.001, 0.01, 0.1, 1\}$; number of PCA components $\in \{5, 10, 15, ..., 100\}$.

Clinical value of iENE as a prognostic biomarker

To evaluate the iENE's predictive performance for outcome prediction, we conduct the following experiments.

We first analyze the predicted iENE grade as a stand-alone clinical biomarker for outcome prediction. Stratification of the study population and comparison with current AJCC reported criteria is reported. We then present a 2-year prediction experiment for each of the OS, DM and DFS binary outcome, introducing a novel attention based multi-modal fusion predictor as the backbone. Interactions between iENE and primary tumor volumes are evaluated through this Deep Learning based model. Finally we explore the use of this model on a longer follow-up period with multi-bin risk modeling.

Kaplan-Meier iENE score univariate estimation As iENE presence is not currently used in the current AJCC recommendations for HPV positive OPCs, we aim to demonstrate that iENE presence is associated with worsened outcomes. From the previously obtained grade classification, we preserve the testing logits from the ENE- vs ENE+ dichotomization scheme as a specific feature. This pipeline based on synthesized deep feature descriptor is compared in uni- and multi-variate models for outcome prediction in OPCs. First as a standalone variate, we fit a Kaplan-Meier estimator to stratify outcome based on the predicted grade probability. Additionally, we perform a log-rank test between the predicted score and outcome presence to determine the outcomes associated with the iENE features.

iENE vs ENE inter-rater variability As the proposed iENE classification model was trained on established neuro-radiology guidelines, we compared the obtained prognostic value of the model predictions with individual raters annotation in order to assess the clinical performance.

We compared the obtained prognostic value of our model predictions with individual raters annotations on a test set of 55 patients. These were independently scored by three clinicians via a Log-rank test for the outcomes DM, DFS and OS, using the annotation as an univariate predictor. All scores used are dichotomized in classes iENE$^-$ vs iENE$^+$.

Additionally, we report on a simulated reader by selecting a random annotation from the set of three clinicians for each patient, then bootstrapping this operation over 10 000 sets, effectively estimating the performance of a trained neurologist by providing an average of the annotation styles. This experiment allows us to estimate a standardized error rate across annotating practices, as well as to assess how a single reader's iENE grading would correlate with outcomes.

Two year risk prediction All outcome prediction models were evaluated in a 5-fold stratified cross-validation scheme, penalized by the negative binary cross entropy log loss function defined as follows:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{6.9}$$

where y represents the ground truth label and p the estimated predicted probability. Models were trained and evaluated using the mean AUC metric.

Baseline model parameters were as follows: two extractor blocks, 500 feature latent space size, two attention heads with hidden size 256, ReLU non linearity activation function, dropout ratio of 0.3, Adam optimizer with a batch size of 32 and starting learning rate of 1×10^{-3}.

To determine the optimal parameter selections, we performed a grid search on the following parameter ranges of values: learning rate $\in \{0.01, 0.001, 0.007\}$; batch size $\in \{8, 16, 32\}$; hidden size of latent space $\in \{256, 512, 1024\}$; dropout ratio $\in \{0, 0.3, 0.4\}$; number of attention heads $\in \{1, 2, 3\}$.

Multi-bin risk outcome modeling Finally, the multi-bin risk prediction performance was assessed using the C-index obtained over the same 5 folds as the previous experiment on binary prediction.

Bins were computed as described in the MTLR paper [46], yielding $\sqrt{|observations|}$ bins corresponding to the quantiles of the survival time distribution.

The C-index provides an assessment of the ranking ability of their survival times based on predicted risks and is calculated as follows:

$$Cindex = \frac{\sum_{i \neq j} 1\left\{r_i < r_j\right\} 1\left\{T_i > T_j\right\} \mathbb{1}_j}{\sum_{i \neq j} 1\left\{T_i > T_j\right\} \mathbb{1}_j}. \tag{6.10}$$

with r_i the predicted risk score for a patient i and T_i their time to event or censoring, effectively reporting the proportion of concordant pairs of risk ranking between two distinct patients over all possible pairings.

We performed the same grid search as described in the binary risk assessment, carried out on the Pytorch lightning [82] framework, using the lifelines [83] package in order to compute the C-index.

6.7 Results

6.7.1 iENE segmentation

Table 6.2 presents the results from the iENE segmentation module. Overall, the proposed nnUnet-derived model achieves the best mean performances across all ENE grades with a DSC of 74.4 ± 24.9, followed by SwinUNETRV2 [12] (70.1 ± 27.7) and SwinUNETR [19] (69.5 ± 28.3). All finetuned models present a high variability in the performance of the segmentation, with standard deviations above 25 % for all of them.

ENE grade stratification shows that smaller nodes (grades 0 and 1) have the highest standard deviations (29.0 and 28.2). Grade 2 nodes, on the other hand, show the best segmentation performance across all nodes and trained models, with a mean DSC of 81.1 ± 16.2 achieved by SwinViT. Grade 3 nodal extensions generally present worse segmentation results with high variance (68.0 ± 24.5 for the proposed pipeline), demonstrating the added difficulty in identifying the exact nodal extension margin on lower contrast regions where tissue and tumor content are of similar densities (refer to the last row of 6.4 for a quantitative result).

SamMed3D [20] performed the worst out of all models, with a Dice score of 46.4 ± 18.4 for the single prompt version. Increasing the number of prompt points to 10 did not improve performance, which lowered the mean Dice score to 39.4 ± 18.1.

Qualitative results for the trained models are shown in Figure 6.4. We can observe that nnUnet presents more precise and detailed margins in comparison to the vision transformers, which can be explained by differences in receptive fields, with the local precision of convolution in comparison to the global context of the attention mechanism. For all these cases, the

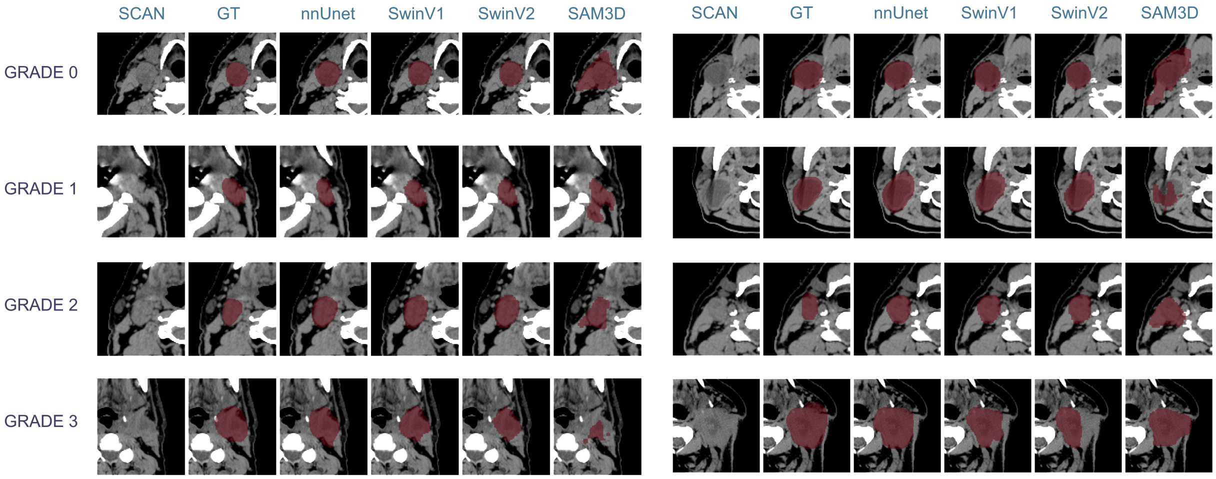


Figure 6.4 Qualitative results for the iENE segmentation task on radiotherapy planning CT images. Each row presents the raw CT scan and the output segmentation maps for two distinct patients from test folds over a 5-fold cross validation. Comparative methods are shown in their respective columns : nnUnet [18] ; SwinUNETR-SSL [19] ; SwinUNETRV2 [12] and SamMed3D [20]. GT: ground truth mask.

obtained segmentation detected the pathological nodule and provided out of object margins, as expected from annotations originating from radiotherapy planning. For the presented grade 3 case, we can observe that the proposed extremities of the extension vary significantly from the ground truth to models, showing the annotation difficulty of this advanced stage.

Table 6.2 Quantitative segmentation results obtained in the largest iENE component segmentation task (n=397). Each result presents the mean DSC (± standard deviation) over the 5 cross-validation folds. * SamMed3D is not trained and only used for inference with a selected number of interest point prompts; all other models are fine-tuned. *Vit*: Vision image transformer architecture [1]. *CNN*: convolutional neural network architecture [2].

Model	Type	Pre-training	Parameters	Grade 0	Grade 1	Grade 2	Grade 3	Average
SwinUNETR	Vit	✓	62.2 M	67.3 ± 29.9	63.1 ± 33.0	80.0 ± 16.6	67.3 ± 25.9	69.5 ± 28.3
SwinUNETRV2	Vit + CNN		72.8 M	67.8 ± 30.0	**69.3 ± 25.9**	**81.1 ± 14.4**	66.1 ± 26.1	70.1 ± 27.7
SamMed3D (1 point) *	Vit + Prompt	✓	100.0 M	43.4 ± 18.6	55.1 ± 19.4	55.1 ± 10.7	58.3 ± 20.8	46.4 ± 18.4
SamMed3D (10 points) *	Vit + Prompt	✓	100.0 M	35.5 ± 18.0	37.9 ± 16.4	47.8 ± 13.6	54.9 ± 17.0	39.4 ± 18.1
Proposed pipeline	CNN		101.9 M	**68.8 ± 29.0**	68.1 ± 28.2	80.3 ± 16.2	**68.0 ± 24.5**	**74.4 ± 24.9**

6.7.2 Automatic iENE grading from segmented CT

Table 6.3 presents the binary iENE grade classification results. The highest AUC of 79.92 ± 2.51 was achieved using radiomic features with the XGBoost algorithm and Lasso feature selection. For the dichotomization case 0-1 vs. 2–3, the best performance was obtained using principal component analysis (PCA) with XGBoost, yielding an AUC of 80.6 ± 4.52. For grade 3 identification, an AUC of 89.93 ± 5.07 was achieved using the XGBoost algorithm with PCA feature selection on the combined radiomic and FMCIB feature set. The choice of feature selection method appeared to depend on the dichotomization scheme, with Lasso performing better for iENE- vs. iENE+, and PCA showing superior performance in the other cases. The corresponding ROC curves for all dichotomization schemes are shown in Figure 6.5.

The inclusion of FMCIB features alone generally hindered performance. Indeed, these features alone led to significantly lower results across all schemes, with an AUC of 66.82 ± 2.32 for iENE classification. However, the combination of FMCIB features to radiomics produced the best result for grade 3 identification, the standard deviation was too large to confidently attribute the improvement to their inclusion.

Figure 6.6 presents the SHapley Additive exPlanations (SHAP) analysis of feature importance for iENE- vs. iENE+ classification. Higher levels of texture non-uniformity were most predictive of a positive ENE case. While most contributing features were texture-related, greater elongation distance of the node also contributed to iENE+ classification.

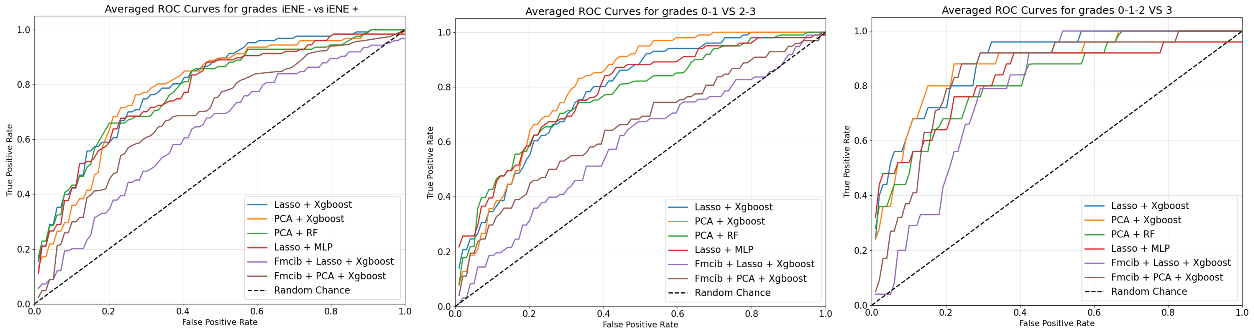


Figure 6.5 ROC curves obtained for the binary classification of dichotomized iENE classes. Individual patient logits were obtained in a 5-fold cross validation scheme with the L2 norm penalized XGBOOST classifier.

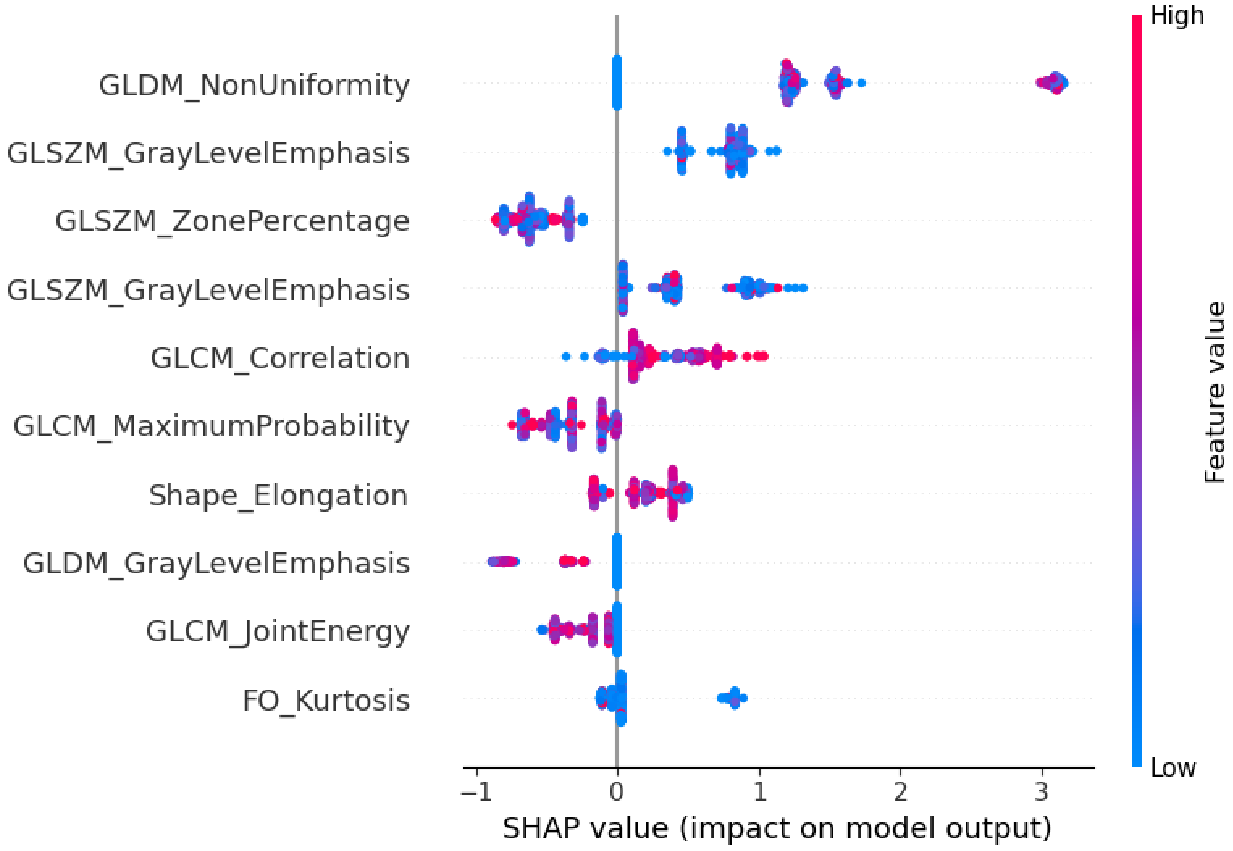


Figure 6.6 SHAP analysis for the prediction of iENE- vs iENE + cases.

Table 6.3 iENE grade classification results for different dichotomization schemes. Figures are the mean AUC (± standard deviation) over the 5 tests fold obtained in a stratified cross-validation. Features indicate the CT feature extraction strategy used as inputs to the classification Method.

Features	Method	iENE^{-} *vs* iENE^{+}	0-1 / 2-3	0-1-2 / 3
Radiomics	RF + PCA	78.12 ± 4.67	76.93 ± 5.98	89.87 ± 5.24
	Lasso + MLP	77.55 ± 2,65	74.98 ± 2.80	80.64 ± 17.31
	PCA + XGBOOST	76.39 ± 2.56	**80.60 ± 4.52**	86.60 ± 8.40
	Lasso + XGBOOST	**79.92 ± 2.51**	78.00 ± 1.61	88.85 ± 5.27
FMCIB	PCA + XGBOOST	65.33 ± 3.41	63.32 ± 5.51	75.52 ± 12.18
	Lasso + XGBOOST	66.82 ± 2.32	59.98 ± 5.69	63.19 ± 10.36
FMCIB &	PCA + XGBOOST	67.82 ± 7.25	66.98 ± 7.65	87.99 ± 2.58
Radiomics	Lasso + XGBOOST	76.68 ± 5.70	76.85 ± 6.03	**89.93 ± 5.07**

6.7.3 Prognostic value of ENE in HPV+ OPC

Kaplan-Meier curves Figure 6.7 presents the Kaplan Meier survival curves for the three outcomes as described in Section 6.6.4. We can observe a clear separation between groups based on the predicted iENE status from our classification pipeline. This indicates the importance of the iENE status in HPV-associated OPCs, as patients are more likely to develop worse clinical outcomes if they have higher iENE grades. With regards to the logrank statistical test, for the distant treatment failure, a p-value of $9.61 \times 10^{-10} < 0.05$ was obtained; for DFS, a p-value of $9.83 \times 10^{-5} < 0.05$; for survival, a p-value of $2.8 \times 10^{-3} < 0.05$. This indicates that all have a significant difference between survival groups. This experiment demonstrates that the iENE status may be used for outcome prediction, and that it may be an important predictive biomarker for cancer staging.

Table 6.4 presents the results comparing individual reviewers, the simulated reader (SR), and our model's predictions. For OS, DM and DFS predictions, the model demonstrated a statistically significant difference between event groups, whereas this is not the case for the SR. For all outcomes, inter-reader variability can be directly associated with event-based

Table 6.4 Log-rank statistical test for 55 patients annotated by three separate clinicians. $\overline{C}$: A collegial decision involved a three way consensus over the case definition, and served as the ground truth for model training. C_{1-3} refers to clinician 1, 2 or 3 respectively. SR: a single reader is simulated by randomly selecting one of the three annotations per patient, and bootstrapping the obtained results 10 000 times, giving the average score a unique reader could obtain.

Outcome	$\overline{C}$	C_1	C_2	C_3	SR	Proposed
DM	2.34×10^{-2}	1.76×10^{-2}	3.13×10^{-2}	1.61×10^{-1}	7.78×10^{-2}	8.02×10^{-5}
DFS	1.98×10^{-1}	2.16×10^{-1}	3.03×10^{-2}	3.39×10^{-1}	1.93×10^{-1}	3.30×10^{-2}
OS	2.62×10^{-1}	1.82×10^{-1}	1.86×10^{-1}	3.32×10^{-1}	2.82×10^{-1}	2.90×10^{-2}

discrimination, as indicated by the provided p-values.

Table 6.5 Binary 2-year classification algorithms for distant metastasis (DM), disease-free survival (DFS) and overall survival (OS). Proportion of (uncensored) events at the end of the study is reported in parentheses. Figures are the mean AUC (± standard deviation) over the 5 cross-validation folds. All presented algorithms incorporate the primary tumor, predicted nodal and clinical characteristics. For every parametric model we present the best performance obtained via grid-search as proposed by the respective authors.

Model	Algorithm	DM 2 Y (6%)	DFS 2Y (12%)	OS 2Y (5%)
Cox	Proportional Hazard Regression	69.7 ± 10.7	67.0 ± 8.5	63.4 ± 17.9
[77]	Logistic Regression	68.2 ± 6.4	59.4 ± 17.9	57.1 ± 11.6
[70]	Bagged risk estimation	70.7 ± 5.0	62.9 ± 3.2	60.9 ± 12.7
[27]	Mrmr Random Forest	78.2 ± 7.2	67.2 ± 5.9	61.4 ± 12.4
[77]	Fuzzy Logistic Mrmr	80.5 ± 7.9	63.6 ± 6.3	57.9 ± 15.1
Ours	Multi-Head Attention Classifier	**88.2 ± 4.8**	**78.1 ± 8.6**	**72.6 ± 9.6**

2-Year outcome risk prediction results Table 6.5 presents the results for the 2-year outcome prediction pipeline. Several multi-modal binary classifiers were compared to the proposed approach. Our multi-modal attention based classifier obtains the highest performance for all three reported outcomes, with 88.2 ± 4.8 for 2-year DM, 76.2 ± 9.6 for DFS and 76.9 ± 10.5 for OS prediction.

Table 6.6 Ablation study for the modality fusion module of AMO-ENE, tested on the two year DM classification task.

Fusion Type	Params (M)	Memory (MB)	Recall	Specificity	AUC
Early Fusion	1.95	15.4	60.3±19.2	87.5±3.5	76.4±7.5
Late Concatenation	0.86	15.3	55.0±28.6	**91.2±6.2**	80.1±7.1
Late Soft Attention	1.17	12.7	75.2±23.2	88.4±5.2	86.7±6.6
Ours (MHSA)	2.37	15.4	**75.4±19.5**	86.6±5.2	**88.2±4.8**

Table 6.7 Ablation study for the number of heads used in our multi-head attention fusion module of AMO-ENE, tested on the two year DM classification task.

Model	Number of Heads	Memory (MB)	Recall	Specificity	AUC
AMO-ENE	2	15.62	71.2±27.0	74.3±21.5	75.4±10.1
	4	15.65	62.0±12.9	83.2±7.6	80.8±7.0
	8	15.68	**75.4±19.5**	86.6±6.5	**88.2**±**4.8**
	16	15.71	59.3±27.6	**87.2**±**4.9**	80.1±13.4

Figure 6.8 presents the comparative experiments for the outcome prediction model with all possible combinations of input features provided. We observe that imaging based characteristics allow for an improved outcome prediction over the clinical baseline (88.2% > 73.5% AUC on DM prediction), which includes actual AJCC cancer staging statuses. For metastatic involvement, the combination of all three modalities yielded the best result (88.2%), with a major component residing in the iENE characteristics, scoring 82.1 % when considered as a sole input. DFS also benefited from the inclusion of all three components reaching 78.1%± 8.6.

Multi-bin risk modeling outcome estimation results Table 6.8 presents the C-indexes of trained survival models on the multi-omics dataset. The results for the Cox regression model only include the clinical metadata as they could not converge on multi-omics. Inclusion of the nodal component through the predicted iENE score or the multi-omics approach improved risk assessment for all three outcomes (80.1±6.9 > 66.0±7.7 for DM). Performance did not improve for OS by integrating MTLR fusion (70.2 ± 10.9 AUC) over a Cox model integrating the iENE score (70.39 ± 8.9 AUC).

Table 6.8 Survival risk estimation for metastatic failure (DM), disease-free survival (DFS) and overall survival (OS). Figures are the mean concordance index (C-index) (± standard deviation) over the folds from the same 5-fold stratified cross-validation. For every parametric model we present the best performance obtained via grid-search as proposed by the respective authors.

Outcome	Features		Algorithm	DM	DFS	OS
	Clinical	Radiomics				
Cox	✓		Cox Regression	66.0 ± 7.7	66.5 ± 6.3	64.4 ± 18.5
[70]	✓	✓	Bagged risk estimation	81.5 ± 8.1	68.8 ± 4.2	62.89 ± 8.1
Cox + XGBOOST iENE score	✓	✓	Cox Regression	80.1 ± 6.9	68.6 ± 5.1	**70.39 ± 8.9**
Ours (MHSA)	✓	✓	AMO-ENE + MTLR	**83.3 ± 6.5**	**70.0 ± 8.1**	70.2 ± 10.9

6.7.4 Ablation Experiments

Finally, we present the results of ablation studies designed to evaluate the relative importance of key components in AMO-ENE for integrating multi-modal input.

Table 6.6 reports performance on the DM prediction task using different modality fusion strategies, while keeping all other components of AMO-ENE unchanged. Late fusion approaches all performed better than early fusion, with significant improvements in performance with comparable or lower parameter and memory requirements. Additionally, we observed that attention-based fusion outperformed latent concatenation, underscoring that enabling cross-modal interactions through attention mechanisms further enhances predictive performance, albeit at the cost of increased parameter count.

Table 6.7 presents the effect of varying the number of attention heads in the MHSA fusion module. Using 8 heads yielded the best classification performance, with notable gains over using just 2 heads. However, increasing the number of heads to 16 did not result in further improvements. These results underscore the importance of hyperparameter tuning when integrating MHSA for modality fusion.

6.8 Discussion

In this study, we proposed a framework to automatically detect and characterize extra nodal extensions in HPV positive OPC patients, obtaining their grading score which is used as a prognostic imaging marker of radiotherapy outcomes. We first proposed an automated nodal extension segmentation method from CT scans, which avoids time-consuming and interpretation-subjective tasks by physicians. From the obtained segmentation masks, the model extracts specific biomarkers to automatically assess the iENE grade. We linked the obtained score to DM, DFS and OS outcomes and compared its prognostic value against single annotators and current clinical guidelines. Finally, we proposed a multi-omic attention based network to predict these three outcomes at two years, combining and assessing the role of lesions for this survival analysis task.

Segmentation of ENE remains a challenging task, with ground truth annotation margins significantly affecting both the training and evaluation of prognostic models. The increased variability in Dice scores underscores this issue, particularly for small nodes, which tend to have a higher surface-to-volume ratio. We achieved a maximum overall Dice of 74.4 ± 24.9 using nnUNet, followed by SwinUNETRv2, indicating that the size of the dataset was better suited to CNN-based methods. The use of a foundation prompt model proved detrimental to segmentation performance. We hypothesize that SamMed3D was not pre-trained on HNC nodal lesions and, as a result, struggled to accurately identify segmentation boundaries—even with 10 points provided. It should be noted that segmentation performance only reflects a model's ability to replicate the provided annotations, which may suffer from inter-annotator variability as is our case. Segmentation performance also does not necessarily reflect the prognostic relevance of the predicted masks.

From the extracted masks, specific biomarkers were subsequently extracted, allowing for the automated grading of an iENE score. To that end, we compared radiomics against deep foundation model features [10] and obtained the best result with the former, achieving 79.92 ± 2.51 AUC score for iENE$^-$ (0) against iENE$^+$ (1/2/3) classification. We hypothesize that radiomic features harness strong information for ENE discrimination, as the size of the object is indicative of higher grades, the texture indicative of tumoral spread and the sphericity of the node indicates how likely one is to have its capsule ruptured. Deep features performed worse than the radiomics approach: we hypothesize that non handcrafted FMCIB features may be less explicit in their ability to model nodes and the imposed fixed region of interest of the FMCIB model hindered uniformity of context for our variable collection of object sizes. We confirm this hypothesis with an explainability analysis with SHAP values, and observe that textural intensity features such as GLDM non uniformity contribute highly to the suspicion of iENE+ if the metastatic node has high variance of its HU values. Once again, classification performance needs to placed in the context of the provided annotations, as predicted iENE scores may be more or less indicative of the actual status of the disease.

To evaluate the prognostic relevance of the predicted iENE score, we compared its stratification performance against both ground truth annotations and individual clinician assessments across OS, DM and DFS. The groupings generated by the model's predicted iENE score consistently achieved statistical significance across all outcomes, with p-values of 8.02×10^{-5} for DM, 3.30×10^{-2} for DFS, and 2.90×10^{-2} for OS based on the log-rank test.

These values are notably more discriminative than the collegial decision ($\overline{C}$), which yielded less significant p-values: 2.34×10^{-2} (DM), 1.98×10^{-1} (DFS), and 2.62×10^{-1} (OS). When compared to individual raters and against the simulated single reader (SR) scenario, which mimics clinical uncertainty by bootstrapping individual annotations, weaker association with outcomes was found, failing to reach statistical significance in all cases. These results indicate that the model not only surpasses individual expert assessments but also outperforms a robust consensus label in stratifying patients by risk. We hypothesize that this is due to the inherent subjectivity of iENE grading, which may limit human-level reproducibility. In contrast, the model leverages standardized, data-driven features, allowing for a more robust and precise interpretation of underlying imaging biomarkers.

Finally, from the nodal characteristics, we introduced an outcome prediction model and compared it to other multi-omic fusion methods in the literature. Our model achieved the best performance for all three outcomes at the two year landmark, with 88.2 ± 4.8, 78.1 ± 8.6 and 72.6 ± 9.6, for DM, DFS and OS, respectively. For DM, the characterization of the nodal extension was the main contributor to the prediction (83.6 % as standalone), but was complemented by both the clinical and tumoural primary information, increasing the AUC by 4.6 %. This suggests that extra nodal extension expresses the ability of an HPV-associated OPC to metastasize, and that its characteristics contain biomarkers predictive of the outcome. DFS also showed the best results with the combination of all three modalities. Patients' survival at two years was mainly dictated by the primary tumor characteristics (82.1 % AUC > 66.2 for node radiomics), suggesting that metastatic failure through nodal extensions is not the main life threat at this time point.

From separate multi-modal inputs, we show that deep attention based modality fusion allows to improve the combination and exploration of interactions than the proposed selection algorithms.

We also observed an improvement in performance in survival estimation across the full follow-up period (2 months - 7 years) by integrating the MTLR framework [46] into our multi-omic fusion model. The most notable improvement was seen in the distant metastasis (DM) prediction, where the model achieved a C-index of 83.3 ± 6.5, substantially outperforming the Cox regression baseline trained on clinical features alone (66.0 ± 7.7) as well as the Cox model integrating the predicted iENE score in addition to the clinical criteria (80.1 ± 6.9). This highlights the significant impact of extra-nodal extension features in modulating metastatic risk over time. Disease-free survival (DFS) estimation also benefited from this approach, reaching a C-index of 70.0 ± 8.1. Overall survival (OS) prediction improved compared to the clinical baseline (70.2 ± 10.9 vs. 64.4 ± 18.5), although it did not surpass the Cox model that incorporated the iENE score. This may be attributed to the relatively small number of uncensored OS events (n=38), which limits the model's ability to learn from death-associated patterns. Furthermore, OS is inherently complex to model, as it may be influenced by multiple competing risk inducing conditions, leading to unreported causes of death not directly related to OPC [84].

Ablation studies on the proposed AMO-ENE model confirmed that multi-head self attention improved classification results in comparison to other modality fusion methods and other attention mechanisms. They also showed that the number of attention heads used in multi-head self attention was a crucial hyperparameter, needing to be tuned accordingly in order to allow the attention mechanism to evaluate a sufficient set of different representations between tokens while not trying to search too many.

This study has some limitations that would need to be addressed in future works. First, the model was trained using data from a single center, and would benefit from an external validation set to evaluate generalizability, choice of scanning protocols and treatment variations. Fluctuations inter institutions for iENE annotation practices would be specifically important to assess. In addition, our study population is predominantly male, leading to potentially unexplored sex specific discrepancies in the obtained results. The current dataset lacks information on race and ethnic background, preventing the evaluation of potential performance disparities across under-represented groups. Finally, the timing disparity in the follow up exams may affect risk estimation. The inclusion of larger datasets with long-term follow-ups would improve robustness of presented results.

6.9 Conclusion

In conclusion, we propose a fully automated end-to-end framework for the staged detection and classification of iENE followed by outcome prediction using head and neck planning CT scans and clinical data in HPV-associated OPC patients. This pipeline could be deployed in a clinical setting with the potential to alleviate the burden of the nodal assessment task and provide a more robust classification system for cancer staging. Notably, this standardized approach could help mitigate inter-rater variability in iENE grading, with the added benefit of promoting the access of prognostic tools for centers without specialized expertise. The findings of this work support the inclusion of the iENE status in cancer staging evaluation for HPV positive OPCs. Further research will be conducted on integrating a larger cohort of patients, with longer follow up evaluations and multi-site cross institutional databases. Additionally, we will further evaluate our outcome prediction model by including a larger cohort of organs at risk and lesions in order to enhance predictions by mapping the complex head and neck structure. Finally, a clinical deployment may be warranted to evaluate the clinical impact of such model on treatment decision making.

Credit authorship contribution statement

- Manuscript redaction: Gautier Hénique
- Main technical contribution: Gautier Hénique, Gabriel Dayan
- Dataset collection and clinical practice: Kristoff Nelson Apostolos Christopoulos , Edith Filion, Phuc-Felix Nguyen-Tan , Laurent Letourneau , Houda Bahig.
- Reviewing and supervision: Laurent Letourneau , Houda Bahig, Samuel Kadoury

Declaration of Interests

The authors have no relevant financial or non-financial interest to disclose.

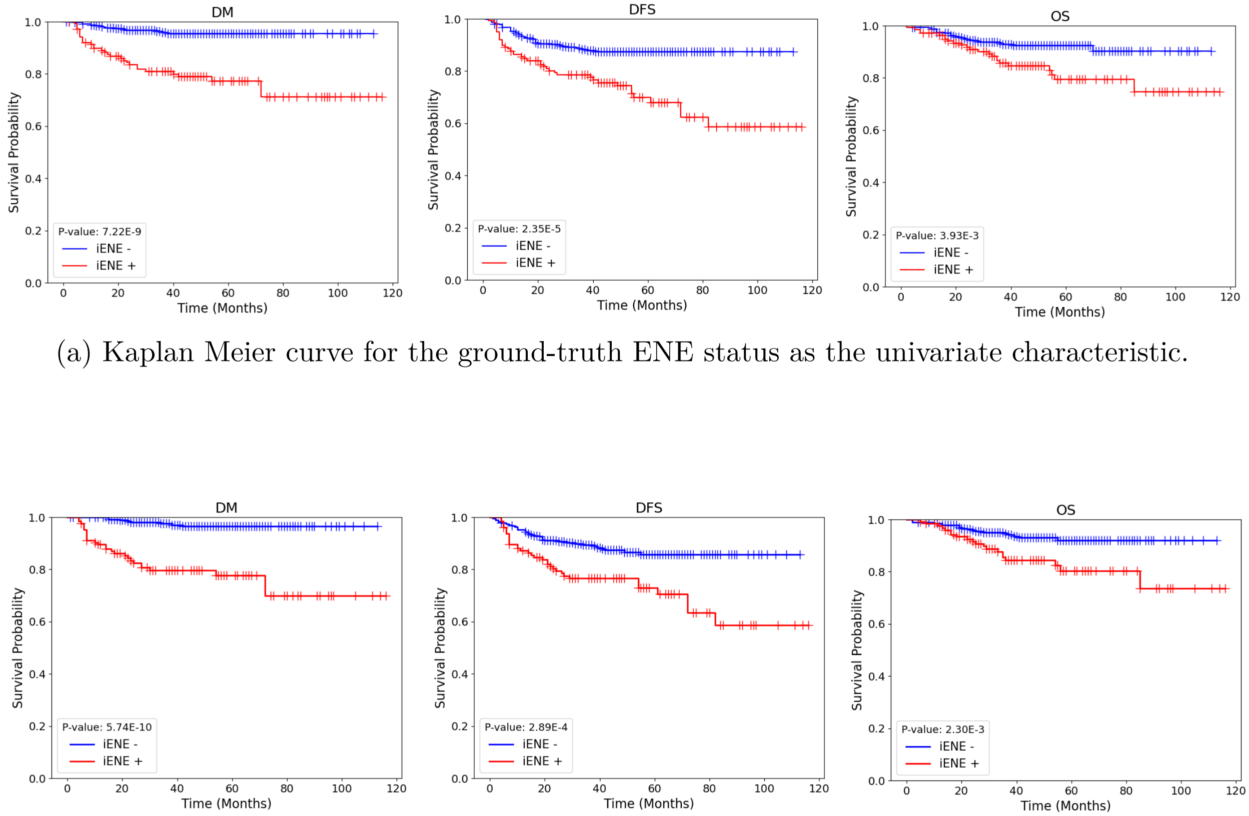
Compliance to ethical standards

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the CHUM's Ethics Committee (IRB 19.026).

Data availabiltiy Data may be made available upon request. Address to the relevant authors for additional information.

Acknowledgements. Funding was provided by:

This research has been funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). Laurent Letourneau-Guillon: Fonds de Recherche Quebec Sante (FRQ-S)/Fondation de L'Association des Radiologistes du Quebec (FARQ), clinical research scholarship Junior 1 salary award (311203).



(a) Kaplan Meier curve for the ground-truth ENE status as the univariate characteristic.

(b) Kaplan Meier curve for the predicted iENE score as the univariate characteristic.

Figure 6.7 Kaplan Meier curves obtained by stratifying all 397 patients with the predicted iENE$^+$ score. Blue lines constitute the real or predicted iENE$^-$ groups, red lines show the real or predicted iENE$^+$ population. The selected node for grade classification is obtained as the largest predicted volume. The Kaplan Meier fit function is unimodal and takes as input the thresholded (from AUC) binary prediction value. Each curve displays as legend the p-value obtained in a logrank test (DM: $p = 9.6 \times 10^{-10}$, DFS: $p = 9.83 \times 10^{-5}$; OS: $p = 2.8 \times 10^{-3}$).

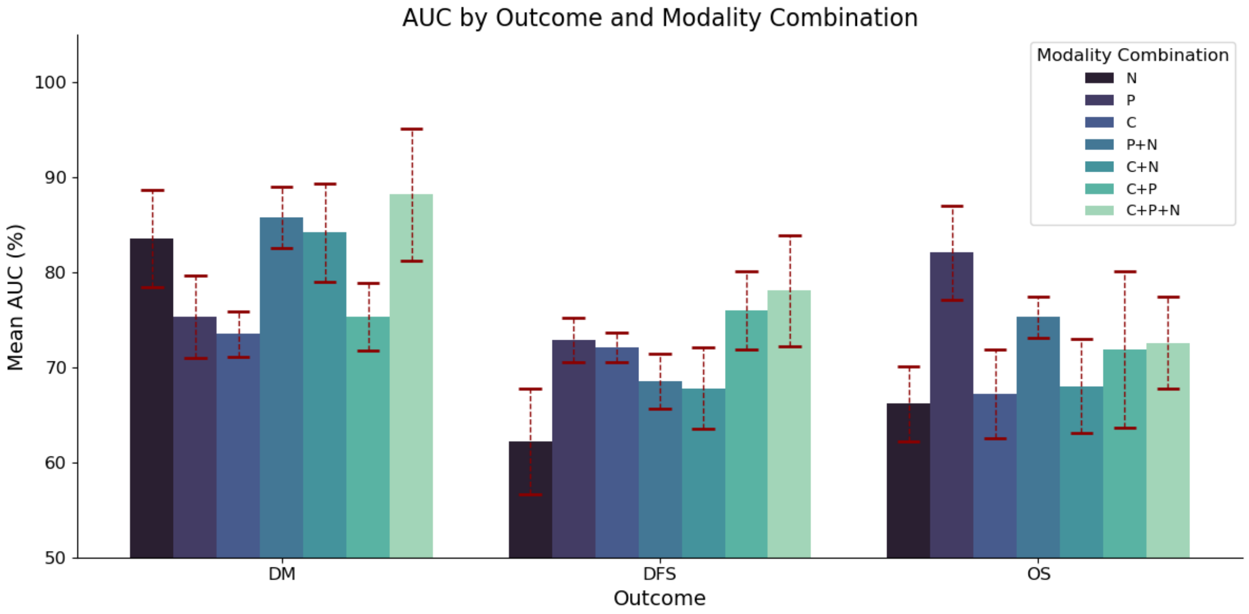


Figure 6.8 Modality ablation experiment results for the 2-year therapeutic outcome prediction model. For readability, C indicates the clinical variables (including the TNM staging), P represents the primary tumor value (GTV radiomic characteristics) and N the predicted nodal volume (iENE) radiomics. Each result presents the mean AUC over the 5 tests fold obtained in a stratified k-fold validation. For each modality and combination of modality, we present the best mean AUC obtained over a grid search as described in 6.6.4, as well as the standard deviation plotted in red dashed lines.

CHAPTER 7 GENERAL DISCUSSION

In the preceding sections, we presented our use of biomedical imagery and deep learning approaches in order to enhance radiation-therapy outcome and toxicity control.

7.1 Summary of Works

7.1.1 Dose-Aware Toxicity Prediction

Our initial study investigated the interactions between anatomical shifts and dosimetry in radiotherapy planning, with the aim of assessing treatment-induced toxicities. This research enabled us to explore the following key aspects:

- The integration of dose plans with daily anatomical deformations
- Temporal stratification of toxicity prediction results through fraction-wise analysis

Extracting toxicity-relevant biomarkers from dose-deformation maps proved to be a challenging task, heavily influenced by pre-processing procedures and the selection of neural network hyperparameters. To support this effort, we expanded our in-house dataset by acquiring imaging data from over 700 additional patients. Consequently, the pre-processing pipeline played a critical role in ensuring the generation of accurate and reliable dose-deformation maps.

Our analysis revealed modality-specific predictive dynamics. The incorporation of dosimetric data alongside deformation maps was found to be significant primarily in the prediction of radionecrosis. In contrast, anatomical deformations alone remained the dominant factor in predicting the need for a feeding tube. Interestingly, while the predictive performance for radionecrosis using dose-deformation maps did not vary significantly across time points, suggesting that these maps might reflect treatment planning risk factors more than actual dose delivery to organs-at-risk (OARs), the improved overall performance observed when combining both modalities suggests otherwise. This indicates that the model is capable of capturing nuanced interactions between dose distributions and anatomical changes relevant to toxicity prediction. However, predicting hospitalization remained particularly challenging, with the integration of imaging data failing to yield significant improvements in performance.

Limitations of the Toxicity Prediction Pipeline

A primary limitation identified early in our pipeline was the accuracy of deformation quantification between CBCT (cone-beam computed tomography) fractions. This issue stems from multiple factors, most notably the lack of annotated CBCT scans. Although global image similarity metrics were used to evaluate the quality of deformable image registration, it is plausible that the registration framework was effective only at capturing large-scale anatomical shifts, rather than subtle tissue- or tumor-specific deformations that may carry critical information for predicting toxicity and treatment outcomes.

Moreover, our approach to longitudinal analysis relied on binary fraction comparison. While this simplified the modeling task and reduced computational complexity, it also introduced a smoothing effect across daily anatomical variations. Such daily fluctuations may be highly relevant for characterizing actual dose delivery and, by extension, toxicity outcomes.

Another limitation lies in the absence of interpretability metrics within our current model. This restricts the clinical applicability of the results, as the decision-making processes underlying toxicity predictions remain opaque. Consequently, these outputs cannot be easily used by medical physicists to adjust treatment plans or by radiation oncologists to assess the likelihood of toxicity. Future integration of interpretability mechanisms such as attention maps or Gradient weighted Class Activation Mapping (Grad CAM) could enhance model transparency and facilitate clinical decision-making.

Lastly, the performance of the vision-based models employed in this study was constrained by the computational resources available at the time. Notably, significant improvements could likely be achieved by leveraging large-scale pre-training frameworks now available for CT imaging, which may enhance feature extraction and model generalization.

7.1.2 AMO-ENE

The second study presented in this thesis evaluated the role of imaging-visible extranodal extension in predicting oncological outcomes for patients with HPV-positive head and neck cancer. This work allowed us to explore the following key components:

- Automatic segmentation of nodal extensions using Convolutional Neural Networks, Vision Transformers, and foundation models
- Radiomics-based classification of iENE grades
- Introduction of the AMO network for attention-based fusion of radiomic features

- Risk stratification for oncological outcomes based on the integration of nodal and tumoral radiomics

This work aimed to re-center our research attention on more clinically actionable targets with defined roles in treatment selection and dose escalation strategies. To this end, we developed a new pipeline focused on a clinically relevant subgroup of our dataset: patients with HPV-positive oropharyngeal cancer.

Our pipeline demonstrated robust performance in segmenting iENE by accurately delineating both pathological nodes and their interactions with surrounding tissues. This enabled a high-fidelity anatomical representation of extranodal spread.

We further demonstrated the feasibility of automated iENE grade classification from the segmented regions, leveraging radiomic features extracted from the detected nodes. The proposed classification model achieved strong performance, with an AUC exceeding 80% for distinguishing iENE-negative from iENE-positive cases. Notably, the model performed even better in extreme grade 3 cases, achieving AUCs of approximately 90%.

Our findings also confirmed that a positive iENE status is significantly associated with poorer oncological outcomes, reinforcing the clinical relevance of this feature for pre-treatment risk stratification. Additionally, when comparing the stratification ability of our model to that of clinician assessments, we observed that radiomics-based classification provided a more discriminative separation of outcome groups. This suggests that iENE grading is challenging in the clinical setting, and that deep learning models may effectively aggregate inter-reader variability to produce more reliable predictions.

The AMO-ENE model outperformed other radiomic-based pipelines for head and neck outcome prediction on our dataset. This result emphasizes the importance of capturing deep, cross-modal interactions and highlights the value of radiomics in oncological risk assessment. The inclusion of late-stage multimodal attention mechanisms significantly enhanced prediction performance, offering notable gains over baseline feature selection methods. However, these improvements came at the cost of an extensive hyperparameter optimization process. Specifically, model performance was highly sensitive to the number of attention heads and the dimensionality of latent feature representations.

Limitations of the AMO-ENE Pipeline

Despite its promising results, the AMO-ENE pipeline has several limitations that warrant further investigation and refinement.

Firstly, the relatively limited size of the dataset (397 patients) may have constrained the generalizability and robustness of the segmentation model. As with most deep learning approaches, larger datasets tend to yield better performance. In this context, CNN architectures outperformed ViTs, likely due to the latter's greater data requirements. Unfortunately, publicly available datasets annotated for extranodal extension are scarce. Thus, expanding the dataset will likely require alternative strategies such as semi-supervised learning (e.g., mean teacher methods) or self-supervised learning on available HNC datasets to enhance feature representation without the need for manual annotations.

Secondly, our classification pipeline was designed under the assumption that the largest detected node is the most clinically relevant for iENE grading. While this is consistent with clinical practice, it may oversimplify patient stratification, potentially overlooking smaller but more aggressive metastatic nodes that contribute significantly to oncological risk.

We also explored the use of deep feature representations beyond radiomics, including those derived from foundation models. However, these models underperformed relative to expectations. A likely explanation is the lack of pretraining on tasks that explicitly involve extranodal extension as a target feature, limiting the models' ability to extract meaningful representations. Future work could address this limitation by developing node-specific masks and incorporating them into the training process, which may allow more complex feature descriptors to surpass the predictive power of traditional radiomics.

Lastly, the AMO-ENE pipeline did not incorporate radiotherapy planning parameters, such as dose-volume histogram characteristics or additional OARs. These data could be integrated into the existing framework to provide a more holistic view of treatment-related risks and potentially improve outcome prediction performance.

7.2 Discussion on Datasets Used

In our studies, we focused on in-house datasets collected at CRCHUM. This was necessary for both projects, as there is currently no public dataset reporting radionecrosis, nasogastric tube usage, or hospitalizations in HNC patients, nor any dataset providing extranodal segmentation annotations. The closest available resource is the HEKtor dataset [28], which includes pathological node segmentations on PET-CT images, but lacks iENE scores and extranodal extension segmentations.

Using in-house datasets, was not without challenges. Both projects required extensive data extraction, conversion, preprocessing, and quality control.

For the toxicity prediction task, over 750 new cases had to be extracted from the clinical

Table 7.1 Simple description of the datasets used in the two studies. * Distribution of outcomes is reported for the whole available follow up time at the date of study.

	OBJECTIVE/PAPER : 1	OBJECTIVE/PAPER : 2
Datasets	1012 patients	398 patients
Cancer types	HNC : All sub-sites	HNC : Oropharyngeal HPV +
Target end points	RD , NG , HOSPI	iENE grade, DM , DFS, OS
Distribution of events	RD : 4 % NG : 17 % HOSPI : 5%	iENE + : 35 % DM : 8 % * DFS : 17 % * OS : 9 %
Mean Follow up Time	N/A	47.0 (±22.3) months
Median Age	64 years (range: 18–90)	62 years (range: 39–84)
Sex Distribution	70 % Male	80 % Male

system. This required the development of an automated querying tool to interface with the hospital storage system. From the raw DICOM files, metadata headers were anonymized to ensure patient privacy, and the data were converted into formats suitable for machine learning. The most time-consuming step was the development of the registration pipeline for CT to CBCT0, and CBCT0 to CBCTt, with more than 35,000 image pairs processed.

For AMO-ENE, we relied on a subset of the previously described dataset. However, several previously excluded cases had to be reintegrated and re-exported. To expedite the ground-truth segmentation process, automated baseline proposals were generated using an in-house developed technical solution.

Ideally, the methods developed would be validated against publicly available datasets to improve the generalizability and robustness of our results. Our datasets, as summarized in 7.1, reflects the patient population at our institution and may carry inherent biases related to gender, socio-economic status, or specific local care practices. Moreover, differences in imaging equipment and acquisition protocols across institutions pose a significant challenge to model generalization, as image quality, voxel resolution, and contrast profiles factors that directly impact the performance of deep learning models trained on imaging data.

In addition, treatment protocols themselves can differ significantly, such as variations in fractionation schemes (no daily CBCT acquisition in some centers), dose prescriptions (dose levels vary on provided guidelines), immobilization techniques, and concurrent systemic therapies. These procedural differences may lead to variability in both anatomical deformations and dose distributions, affecting both model inputs and ground truth labels. Similarly, population-level differences such as HPV prevalence, comorbidity profiles, and tumor staging at diagnosis can influence both clinical outcomes and the manifestation of treatment-related toxicities. One major example of this fact in HNC is the role of the betel nut in Asia [85], which the consumption reveals to be a region specific risk enhancing practice.

All of these factors underscore the need for external validation on diverse, multi-institutional datasets to assess the robustness, portability, and clinical reliability of AI-driven approaches in head and neck cancer care.

7.3 Future Research

Future work in toxicity prediction should aim to more thoroughly integrate patient-specific anatomical characteristics, with a particular focus on improving the spatial and temporal modeling of treatment dynamics. A promising direction involves analyzing the characteristics of delivered dosimetry during treatment in contrast to the initially intended treatment planning. This approach may yield clinically relevant intermediate representations that better reflect actual exposure patterns and thus improve toxicity prediction accuracy.

Additionally, characterizing dose-volume histogram features longitudinally could offer valuable insights into the evolving radiation exposure experienced by organs-at-risk throughout the treatment course. Such a temporal analysis may facilitate more direct and physiologically grounded modeling of toxicity outcomes.

Regarding the modeling of spatial-temporal interactions, preliminary investigations into the use of four-dimensional (4D) Swin Transformers were conducted, though not completed. Future studies may benefit from exploring more expressive and computationally efficient representations of anatomical deformations using CBCT sequences, coupled with advanced 4D modeling frameworks. These approaches may enable richer and more informative toxicity prediction pipelines.

For the AMO-ENE framework, external validation is essential. This includes extending the analysis to cohorts with non-HPV-related head and neck cancers and to patients without nodal involvement, to evaluate model generalizability and robustness. Given the heterogeneity in the number and distribution of pathological nodes across patients, graph neural networks (GNNs) present a promising direction for future work. GNNs could facilitate the modeling of complex nodal architectures and enhance the integration of cancer-specific variations into the predictive framework.

Finally, future research must prioritize model interpretability and clinical usability. The development of highly complex, opaque models, referred to as black boxes, risks limiting clinical adoption. For toxicity prediction models to be actionable in radiotherapy planning, they must offer transparent and interpretable outputs that enable radiation oncologists to make informed treatment adjustments and ultimately improve patient care.

CHAPTER 8 CONCLUSION

Radiotherapy is a cornerstone of HNC cancer treatment, offering high precision through the integration of advanced medical imaging. This opens the door for the integration of deep learning applications to personalize care and tailor treatment strategies to patient-specific anatomical and biological characteristics. However, for such technologies to be adopted in clinical practice, they must demonstrate not only high predictive performance but also interpretability and clinical actionability. Treatment decisions can only be safely modified when prediction models are robust, explainable, and aligned with clinical reasoning.

The objective of this work was to investigate methodologies for predicting adverse treatment responses by integrating available clinical imaging with automated deep learning pipelines. Our aim was to contribute to the development of reliable tools that may one day support clinical decision-making in radiotherapy.

To this end, we first explored deep learning vision models as a means to analyze both pre-treatment anatomical data and in-treatment imaging responses. This allowed us to identify potential imaging-derived biomarkers associated with increased toxicity risk. Our initial study focused on evaluating the interaction between dose distributions and anatomical deformations as captured through longitudinal CBCT imaging. This analysis provided valuable insight into the role of spatial and temporal anatomical shifts in toxicity development and highlighted the potential of dose-deformation maps as predictive tools.

In a second line of investigation, we returned to treatment planning data to assess the prognostic significance of extranodal extension as visualized on planning CT images. We introduced a fully automated pipeline for iENE detection and grading, culminating in the development of the AMO-ENE model. This architecture employed a multi-head attention mechanism to fuse multi-organ radiomic features and demonstrated strong performance in stratifying outcomes for HPV-positive OPC patients. Our findings emphasized the clinical relevance of iENE and the added value of integrating radiomic features from multiple anatomical structures for outcome prediction.

Taken together, the findings from both lines of research suggest that further optimization and integration of spatial-temporal imaging data, radiomic signatures could significantly enhance the personalization of radiotherapy. These advances hold promise for the development of adaptive treatment strategies aimed at improving patient outcomes and minimizing treatment related toxicities.

Continued work in this direction particularly with attention to external validation, interpretability, and clinical integration may ultimately pave the way toward more responsive, data-driven radiotherapy planning.

REFERENCES

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, arXiv:2010.11929 [cs]. [Online]. Available: http://arxiv.org/abs/2010.11929

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira *et al.*, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[3] "Head And Neck Cancer Treatments, Causes & Symptoms - NHCancerClinics," Feb. 2024. [Online]. Available: https://nhcancerclinics.com/cancer-types/head-and-neck-cancer/

[4] C. C. S. . S. c. d. cancer, "External radiation therapy." [Online]. Available: https://cancer.ca/en/treatments/treatment-types/radiation-therapy/external-radiation-therapy

[5] C. H. McCollough and P. S. Rajiah, "Milestones in CT: Past, Present, and Future," *Radiology*, vol. 309, no. 1, p. e230803, Oct. 2023, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/full/10.1148/radiol.230803

[6] A. Courville, "IFT6135 Winter 2020 session: Apprentissage de Représentations," https://sites.google.com/mila.quebec/ift6135/lectures?authuser=0, 2020, [PowerPoint slides].

[7] "CNN in Deep Learning: Algorithm and Machine Learning Uses." [Online]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network

[8] P. Giraud *et al.*, "Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers," *Frontiers in Oncology*, vol. 9, Mar. 2019, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2019.00174/full

[9] K. Men *et al.*, "A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial," *International Journal of Radiation Oncology*Biology*Physics*, vol. 105, no. 2,

pp. 440–447, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S036030161930834X

[10] S. Pai *et al.*, "Foundation model for cancer imaging biomarkers," *Nature Machine Intelligence*, vol. 6, no. 3, pp. 354–367, Mar. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-024-00807-9

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

[12] Y. He *et al.*, "SwinUNETR-V2: Stronger Swin Transformers with Stagewise Convolutions for 3D Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, H. Greenspan *et al.*, Eds. Cham: Springer Nature Switzerland, 2023, pp. 416–426.

[13] A. V. Dalca *et al.*, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.

[14] S.-C. Huang *et al.*, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–9, Oct. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-020-00341-z

[15] G. Zhang *et al.*, "Multimodal Deep Learning for Cancer Survival Prediction: A Review," *Current Bioinformatics*, vol. 20, no. 4, pp. 299–322. [Online]. Available: https://www.eurekaselect.com/article/140707

[16] S. Wiegrebe *et al.*, "Deep learning for survival analysis: a review," *Artificial Intelligence Review*, vol. 57, no. 3, p. 65, Feb. 2024. [Online]. Available: https://doi.org/10.1007/s10462-023-10681-3

[17] K. He *et al.*, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03385

[18] F. Isensee *et al.*, "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation," Jul. 2024, arXiv:2404.09556 [cs]. [Online]. Available: http://arxiv.org/abs/2404.09556

[19] A. Hatamizadeh *et al.*, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," Jan. 2022, arXiv:2201.01266 [eess]. [Online]. Available: http://arxiv.org/abs/2201.01266

[20] H. Wang *et al.*, “SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images,” Sep. 2024, arXiv:2310.15161 [cs]. [Online]. Available: http://arxiv.org/abs/2310.15161

[21] C. C. Society, “head and neck cancers | Canadian Cancer Society.” [Online]. Available: https://cancer.ca/en/cancer-information/resources/glossary/h/head-and-neck-cancers

[22] A. Barsouk *et al.*, “Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma,” *Medical Sciences*, vol. 11, no. 2, p. 42, Jun. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10304137/

[23] S. Cros *et al.*, “Combining dense elements with attention mechanisms for 3D radiotherapy dose prediction on head and neck cancers,” *Journal of Applied Clinical Medical Physics*, vol. 23, no. 8, p. e13655, Aug. 2022.

[24] S. Cros, “Méthodes d’apprentissage profond 3d en radiothérapie pour la segmentation d’organes et la prédiction de distributions de dose,” Master’s thesis, Polytechnique Montreal, Nov. 2021, thèse de master soutenue en novembre 2021.

[25] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sep. 2020, arXiv:1905.11946 [cs]. [Online]. Available: http://arxiv.org/abs/1905.11946

[26] S. Xie *et al.*, “Aggregated Residual Transformations for Deep Neural Networks,” Apr. 2017, arXiv:1611.05431 [cs]. [Online]. Available: http://arxiv.org/abs/1611.05431

[27] M. Vallières *et al.*, “Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer,” *Scientific Reports*, vol. 7, no. 1, p. 10117, Aug. 2017, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-017-10371-5

[28] V. Andrearczyk *et al.*, “Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT,” *Head and neck tumor segmentation and outcome prediction: third challenge, HECKTOR 2022, held in conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings. Head and Neck Tumor Segmentation Challenge (3rd: 2022: Singapor...*, vol. 13626, pp. 1–30, 2023.

[29] B. Khajetash *et al.*, “Ensemble learning approach for prediction of early complications after radiotherapy for head and neck cancer using CT and MRI radiomic features,”

Scientific Reports, vol. 15, no. 1, p. 14229, Apr. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-025-93676-0

[30] M. E. A. Elforaici *et al.*, "Semi-supervised ViT knowledge distillation network with style transfer normalization for colorectal liver metastases survival prediction," *Medical Image Analysis*, vol. 99, p. 103346, Jan. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524002718

[31] K. Men *et al.*, "A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial," *International Journal of Radiation Oncology*Biology*Physics*, vol. 105, no. 2, pp. 440–447, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S036030161930834X

[32] W. T. Le *et al.*, "Comparing 3D deformations between longitudinal daily CBCT acquisitions using CNN for head and neck radiotherapy toxicity prediction," Mar. 2023, arXiv:2303.03965 [cs]. [Online]. Available: http://arxiv.org/abs/2303.03965

[33] S. Starke *et al.*, "Multitask Learning with Convolutional Neural Networks and Vision Transformers Can Improve Outcome Prediction for Head and Neck Cancer Patients," *Cancers*, vol. 15, no. 19, p. 4897, Jan. 2023, number: 19 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-6694/15/19/4897

[34] Y. Tang *et al.*, "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis," Mar. 2022, arXiv:2111.14791 [cs]. [Online]. Available: http://arxiv.org/abs/2111.14791

[35] R. Saber *et al.*, "Feature Tokenizer-Transformers with Self-Training for The Prediction of PD-L1 Expression of Non-Small Cell Lung Cancer from CT," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2024, pp. 1–5, iSSN: 1945-8452. [Online]. Available: https://ieeexplore.ieee.org/document/10635812

[36] G. Podobnik *et al.*, "HaN-Seg: The head and neck organ-at-risk CT and MR segmentation challenge," *Radiotherapy and Oncology*, vol. 198, p. 110410, Sep. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167814024006807

[37] K. Peng, D. Zhou, and S. Gong, "OAR-UNet: Enhancing Long-Distance Dependencies for Head and Neck OAR Segmentation," *Electronics*, vol. 13, no. 18, p. 3771, Jan.

2024, number: 18 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/13/18/3771

[38] Z. Zhang *et al.*, “SegReg: Segmenting OARs by Registering MR Images and CT Annotations,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2024, pp. 1–5, iSSN: 1945-8452. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10635437

[39] Y. Tang *et al.*, “Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis,” Mar. 2022, arXiv:2111.14791 [cs]. [Online]. Available: http://arxiv.org/abs/2111.14791

[40] A. Kirillov *et al.*, “Segment Anything,” Apr. 2023, arXiv:2304.02643 [cs]. [Online]. Available: http://arxiv.org/abs/2304.02643

[41] J. Wu *et al.*, “Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation,” Dec. 2023, arXiv:2304.12620 [cs]. [Online]. Available: http://arxiv.org/abs/2304.12620

[42] J. Chen *et al.*, “TransMorph: Transformer for unsupervised medical image registration,” *Medical Image Analysis*, vol. 82, p. 102615, Nov. 2022, arXiv:2111.10480 [eess]. [Online]. Available: http://arxiv.org/abs/2111.10480

[43] W. T. Le *et al.*, “Cross-institutional outcome prediction for head and neck cancer patients using self-attention neural networks,” *Scientific Reports*, vol. 12, no. 1, p. 3183, Feb. 2022.

[44] R. Vanguri *et al.*, “Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l)1 blockade in patients with non-small cell lung cancer,” *Nature Cancer*, vol. 3, pp. 1–14, 08 2022.

[45] J. L. Katzman, “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network | BMC Medical Research Methodology | Full Text.” [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1

[46] C.-N. Yu *et al.*, “Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors,” in *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011. [Online]. Available: https://papers.nips.cc/paper_files/paper/2011/hash/1019c8091693ef5c5f55970346633f92-Abstract.html

[47] M. Chen, K. Wang, and J. Wang, “Advancing Head and Neck Cancer Survival Prediction via Multi-Label Learning and Deep Model Interpretation,” *ArXiv*, p. arXiv:2405.05488v1, May 2024.

[48] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Research*, vol. 77, no. 21, pp. e104–e107, Nov. 2017.

[49] N. V. Chawla *et al.*, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, arXiv:1106.1813 [cs]. [Online]. Available: http://arxiv.org/abs/1106.1813

[50] B. Lacas *et al.*, “Role of radiotherapy fractionation in head and neck cancers (march): an updated meta-analysis,” *The Lancet Oncology*, vol. 18, no. 9, pp. 1221–1237, 2017.

[51] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Dec. 2015, arXiv:1511.08458 [cs]. [Online]. Available: http://arxiv.org/abs/1511.08458

[52] C. Ordóñez-Sanz *et al.*, “Cbct imaging: a simple approach for optimising and evaluating concomitant imaging doses, based on patient-specific attenuation, during radiotherapy pelvis treatment,” *The British Journal of Radiology*, vol. 94, no. 1124, p. 20210068, 2021.

[53] Z. Zhang *et al.*, “Advantages and robustness of partial vmat with prone position for neoadjuvant rectal cancer evaluated by cbct-based offline adaptive radiotherapy,” *Radiation Oncology*, vol. 18, no. 1, pp. 1–11, 2023.

[54] P. Buranaporn *et al.*, “Relation between dir recalculated dose based cbct and gi and gu toxicity in postoperative prostate cancer patients treated with vmat,” *Radiotherapy and Oncology*, vol. 157, pp. 8–14, 2021.

[55] A. D. Yock *et al.*, “Initial analysis of the dosimetric benefit and clinical resource cost of cbct-based online adaptive radiotherapy for patients with cancers of the cervix or rectum,” *Journal of Applied Clinical Medical Physics*, vol. 22, no. 10, pp. 210–221, 2021.

[56] Q. Qin *et al.*, “Cone-beam ct radiomics features might improve the prediction of lung toxicity after sbrt in stage i nsclc patients,” *Thoracic Cancer*, vol. 11, no. 4, pp. 964–972, 2020.

[57] B. S. Rosen *et al.*, “Early changes in serial cbct-measured parotid gland biomarkers predict chronic xerostomia after head and neck radiation therapy,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1319–1329, 2018.

[58] W. T. Le *et al.*, “Comparing 3D deformations between longitudinal daily CBCT acquisitions using CNN for head and neck radiotherapy toxicity prediction,” Mar. 2023, arXiv:2303.03965 [cs]. [Online]. Available: http://arxiv.org/abs/2303.03965

[59] A. Wentzel *et al.*, “Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy,” *Radiotherapy and Oncology*, vol. 148, pp. 245–251, Jul. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167814020302796

[60] K. Men *et al.*, “A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 105, no. 2, pp. 440–447, Oct. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S036030161930834X

[61] S. Chen, K. Ma, and Y. Zheng, “Med3D: Transfer Learning for 3D Medical Image Analysis,” Jul. 2019, arXiv:1904.00625 [cs]. [Online]. Available: http://arxiv.org/abs/1904.00625

[62] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, “Early vs Late Fusion in Multimodal Convolutional Neural Networks,” in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, Jul. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9190246

[63] L. N. Smith and N. Topin, “Super-convergence: Very fast training of neural networks using large learning rates,” in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[64] M. B. Amin *et al.*, “The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging,” *CA: a cancer journal for clinicians*, vol. 67, no. 2, pp. 93–99, Mar. 2017.

[65] T. Hiyama *et al.*, “Extra-nodal extension in head and neck cancer: how radiologists can help staging and treatment planning,” *Japanese Journal of Radiology*, vol. 38, no. 6, pp. 489–506, Jun. 2020.

[66] F. Hoebers *et al.*, “Augmenting inter-rater concordance of radiologic extranodal extension in HPV-positive oropharyngeal carcinoma: A multicenter study,” *Head & Neck*, vol. 44, no. 11, pp. 2361–2369, 2022,

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hed.27130. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hed.27130

[67] ——, "Augmenting inter-rater concordance of radiologic extranodal extension in HPV-positive oropharyngeal carcinoma: A multicenter study," *Head & Neck*, vol. 44, no. 11, pp. 2361–2369, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hed.27130. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hed.27130

[68] K. M. Boehm *et al.*, "Harnessing multimodal data integration to advance precision oncology," *Nature Reviews. Cancer*, vol. 22, no. 2, pp. 114–126, Feb. 2022.

[69] A. Myronenko *et al.*, "Automated head and neck tumor segmentation from 3D PET/CT," Sep. 2022, arXiv:2209.10809 [eess]. [Online]. Available: http://arxiv.org/abs/2209.10809

[70] L. Rebaud *et al.*, "Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT," in *Head and Neck Tumor Segmentation and Outcome Prediction: Third Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings.* Berlin, Heidelberg: Springer-Verlag, Sep. 2022, pp. 121–134. [Online]. Available: https://doi.org/10.1007/978-3-031-27420-6_13

[71] X. Xiong *et al.*, "Head and Neck Cancer Segmentation in FDG PET Images: Performance Comparison of Convolutional Neural Networks and Vision Transformers," *Tomography*, vol. 9, no. 5, pp. 1933–1948, Oct. 2023, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2379-139X/9/5/151

[72] M. Wodzinski, "Benchmark of Deep Encoder-Decoder Architectures for Head and Neck Tumor Segmentation in Magnetic Resonance Images: Contribution to the HNTSMRG Challenge," in *Head and Neck Tumor Segmentation for MR-Guided Applications*, K. A. Wahid *et al.*, Eds. Cham: Springer Nature Switzerland, 2025, pp. 204–213.

[73] B. H. Kann *et al.*, "Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks," *Scientific Reports*, vol. 8, no. 1, p. 14036, Sep. 2018, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-018-32441-y

[74] N. Kharytaniuk *et al.*, "Association of Extracapsular Spread With Survival According to Human Papillomavirus Status in Oropharynx Squamous Cell Carcinoma and Carcinoma of Unknown Primary Site," *JAMA otolaryngology– head & neck surgery*, vol. 142, no. 7, pp. 683–690, Jul. 2016.

[75] P. Sinha *et al.*, "High metastatic node number, not extracapsular spread or N-classification is a node-related prognosticator in transorally-resected, neck-dissected p16-positive oropharynx cancer," *Oral Oncology*, vol. 51, no. 5, pp. 514–520, May 2015.

[76] M. Meng *et al.*, "Radiomics-enhanced Deep Multi-task Learning for Outcome Prediction in Head and Neck Cancer," Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.05409v1

[77] M. Kazmierski *et al.*, "Multi-institutional Prognostic Modeling in Head and Neck Cancer: Evaluating Impact and Generalizability of Deep Learning and Radiomics," *Cancer Research Communications*, vol. 3, no. 6, pp. 1140–1151, Jun. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10309070/

[78] J. Peng *et al.*, "The prognostic value of machine learning techniques versus cox regression model for head and neck cancer," *Methods*, vol. 205, pp. 123–132, Sep. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1046202322001554

[79] M. M. Oken *et al.*, "Toxicity and response criteria of the Eastern Cooperative Oncology Group," *American Journal of Clinical Oncology*, vol. 5, no. 6, pp. 649–655, Dec. 1982.

[80] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, arXiv:1603.02754 [cs]. [Online]. Available: http://arxiv.org/abs/1603.02754

[81] M. J. Cardoso *et al.*, "MONAI: An open-source framework for deep learning in healthcare," Nov. 2022, arXiv:2211.02701 [cs]. [Online]. Available: http://arxiv.org/abs/2211.02701

[82] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019. [Online]. Available: https://github.com/Lightning-AI/lightning

[83] C. Davidson-Pilon, "lifelines: survival analysis in python," *Journal of Open Source Software*, vol. 4, no. 40, p. 1317, 2019. [Online]. Available: https://doi.org/10.21105/joss.01317

[84] B. S. Rose *et al.*, “Population-Based Study of Competing Mortality in Head and Neck Cancer,” *Journal of Clinical Oncology*, vol. 29, no. 26, pp. 3503–3509, Sep. 2011, publisher: Wolters Kluwer. [Online]. Available: https://ascopubs.org/doi/10.1200/JCO.2011.35.7301

[85] M.-J. Su, C.-H. Ho, and C.-C. Yeh, “Association of alcohol consumption, betel nut chewing, and cigarette smoking with mortality in patients with head and neck cancer among the Taiwanese population: A nationwide population-based cohort study,” *Cancer Epidemiology*, vol. 89, p. 102526, Apr. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877782124000055