cancers

MDPI

Article

Multimodal Deep Learning for Stage Classification of Head and Neck Cancer Using Masked Autoencoders and Vision Transformers with Attention-Based Fusion

Anas Turki *, Ossama Alshabrawy and Wai Lok Woo

Department of Computer and Information Science, Faculty of Engineering and Environment, Northumbria University, Newcastle upon Tyne NE1 8ST, UK; ossama.alshabrawy@northumbria.ac.uk (O.A.); wailok.woo@northumbria.ac.uk (W.L.W.)
* Correspondence: anas.althobaiti@northumbria.ac.uk

Simple Summary

Head and neck squamous cell carcinoma (HNSCC) is a common and deadly form of cancer. Doctors use a system known as AJCC staging to determine how advanced the cancer is, which helps guide treatment. However, current staging methods mostly rely on simple anatomical observations and limited clinical information, which may not fully capture the complexity of the disease. This study introduces a new computer-based approach that combines detailed medical images (CT scans) and patient clinical data using advanced artificial intelligence techniques. The method uses a special kind of deep learning called Vision Transformers, combined with masked autoencoders, to learn important features from the images without the need for many labelled examples. It also applies attention mechanisms to focus on the most relevant parts of the images and the clinical data. The study tested this approach on two public datasets and showed that it could accurately predict cancer stages better than existing methods. Importantly, different attention models worked better depending on the amount and balance of available data. This work demonstrates the potential of combining imaging and clinical information through AI to improve cancer staging, which could help doctors make better decisions and ultimately benefit patient care.

check for updates

Academic Editors: Sam Payabvash and Avraham Eisbruch

Received: 9 May 2025
Revised: 5 June 2025
Accepted: 9 June 2025
Published: 24 June 2025

Citation: Turki, A.; Alshabrawy, O.; Woo, W.L. Multimodal Deep Learning for Stage Classification of Head and Neck Cancer Using Masked Autoencoders and Vision Transformers with Attention-Based Fusion. *Cancers* **2025**, *17*, 2115. https://doi.org/10.3390/cancers17132115

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract

Head and neck squamous cell carcinoma (HNSCC) is a prevalent and aggressive cancer, and accurate staging using the AJCC system is essential for treatment planning. This study aims to enhance AJCC staging by integrating both clinical and imaging data using a multimodal deep learning pipeline. We propose a framework that employs a VGG16-based masked autoencoder (MAE) for self-supervised visual feature learning, enhanced by attention mechanisms (CBAM and BAM), and fuses image and clinical features using an attention-weighted fusion network. The models, benchmarked on the HNSCC and HN1 datasets, achieved approximately 80% accuracy (four classes) and ~66% accuracy (five classes), with notable AUC improvements, especially under BAM. The integration of clinical features significantly enhances stage-classification performance, setting a precedent for robust multimodal pipelines in radiomics-based oncology applications.

Keywords: head and neck cancer; AJCC staging; vision transformer; masked autoencoder; multimodal fusion; radiomics

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) ranks among the ten most prevalent malignancies worldwide, with an estimated 650,000 new cases and 330,000 deaths each year—figures that remain stubbornly high despite advances in modern multimodality care [1–3]. Prognosis hinges on accurate American Joint Committee on Cancer (AJCC) staging, yet the current TNM rubric captures only gross anatomy—primary tumour size, nodal burden, distant spread—and remains blind to the molecular and micro-environmental heterogeneity that drives therapeutic response [4,5]. Recent radiomics studies have shown that quantitative texture signatures extracted from routine CT, PET/CT, or MRI scans add prognostic value beyond conventional TNM; however, most pipelines still rely on handcrafted features and shallow learners that are brittle across scanners and institutions [1,2,6,7].

Deep learning promises to overcome these limitations by learning hierarchical, domain-specific representations directly from the data. Convolutional Neural Networks (CNNs) were the first to deliver sizeable gains, but Vision Transformers (ViTs) now dominate several medical-imaging benchmarks, outperforming 3D CNNs in tasks such as HPV-status prediction [8] and multi-label outcome modelling [9]. Multiple reviews anticipate ViTs as the next state-of-the-art backbone for oncologic imaging [10,11]. Nonetheless, most published HNSCC networks are unimodal, require large, labelled cohorts that do not exist for rare AJCC stages, and rarely fuse the rich clinical context (demographics, biomarkers, therapy) that clinicians routinely consult [12–14].

Self-supervised learning (SSL) offers a principled solution to data scarcity. Masked autoencoders (MAEs) learn generalisable visual primitives by reconstructing missing patches and improve downstream performance in cancer imaging [15,16]. Yet, the synergy between MAE-pretrained ViTs, advanced attention blocks—e.g., the Convolutional Block Attention Module (CBAM) and Bottleneck Attention Module (BAM)—and adaptive multimodal fusion has not been systematically evaluated for AJCC staging. Early multimodal prototypes combining imaging and tabular data (e.g., CNN–latent fusion [12], graph-based approaches [17], or transformer hybrids [18,19]) have reported encouraging but heterogeneous gains, underscoring the need for a controlled comparison of attention mechanisms, SSL pretraining, and fusion strategies.

In this work, we introduce a fully self-supervised framework that (i) pretrains a VGG-16 MAE on unannotated CT slices; (ii) embeds the resulting features into three ViT backbones (BEiT, DeiT, DINO) wrapped with either CBAM or BAM to enhance channel–spatial saliency; and (iii) fuses imaging and nine structured clinical covariates through a gated soft-attention module. We benchmark on two public cohorts—TCIA-HNSCC (four classes) and HEAD-NECK-RADIOMICS-HN1 (five classes)—to test the following hypotheses:

H1. *MAE pretraining yields more discriminative features than ImageNet initialisation.*

H2. *Attention-augmented ViTs outperform vanilla ViTs by focusing on lesion-centric cues.*

H3. *Soft-attention fusion improves predictive power over unimodal or naïve-concatenation baselines. Our pipeline achieves up to 80.4% accuracy and 0.83 AUC despite severe class imbalance, outperforming prior radiomics-only or shallow multimodal baselines [19,20]. By releasing code and pretrained weights, we aim to catalyse further community efforts toward explainable and clinically robust AI for HNSCC staging, in line with current guideline recommendations for transparent, calibration-aware medical ML systems [9,21].*

2. Related Work

Early Handcrafted Radiomics: Quantitative CT descriptors have long been explored for HNSCC survival and recurrence prediction, yet their clinical uptake is hampered by protocol specific bias and shallow learning pipelines. Diamant leveraged grey level co-occurrence features and a CNN ensemble to predict overall survival, but they achieved only incremental gains over TNM nomograms; the clinical covariates were appended by late concatenation rather than jointly optimised [1]. Hu confirmed this limitation, showing that a deep learning radiomics model improved prognostication but still struggled with minority stage recall [2]. Subsequent multi-institutional efforts have echoed the need for domain robust feature learning [6,7,22,23].

Self-Supervised Representation Learning: Masked autoencoder (MAE) pretraining has recently mitigated annotation scarcity in medical imaging. Zheng boosted PET/CT recurrence AUCs by six points with MAE initialisation [15], while Wolf and Huang provided systematic evidence that MAEs and contrastive objectives consistently outperformed ImageNet initialisation on small clinical datasets [16,17]. Nevertheless, MAE-based studies in HNSCC remain outcome centred and omit fine-grained AJCC staging.

Vision Transformers (ViTs): Transformers now outshine CNNs in several oncologic tasks. Lang 's Swin Transformer surpassed 3D CNNs for HPV status prediction (AUC 0.87) [8], while Chen employed a ViT for oropharyngeal survival modelling, achieving a c index of 0.79 [24]. Reviews concur that ViTs' long-range self-attention is well suited to heterogeneous tumours [10,11]. However, transformer studies for AJCC stage classification remain sparse and typically unimodal [25].

Channel–Spatial Attention: CBAM has improved CT-based survival models in HN1 but is typically used with CNNs and static fusion [12]. BAM, while lighter, has demonstrated regularisation benefits in low-data regimes for glioma and liver cancer [26,27]; however, its utility for HNSCC staging remains unexplored.

Multimodal Fusion: Hybrid models that blend imaging with demographics, genomics, or dose–volume metrics often rely on heuristic concatenation or fixed-rule gating [4,13,20]. Adaptive gates or cross-attention mechanisms—e.g., AdaMSS for PET/CT survival, HC MAE for histopathology +omics [3], or Ketabi 's contrastive fusion for paediatric tumours [28]—have shown promise but have not been benchmarked for AJCC staging across both TCIA-HNSCC and HN1.

Research Gap: No prior work has jointly (i) harnessed MAE-initialised ViTs on head and neck CT; (ii) contrasted CBAM with BAM inside identical transformer backbones; and (iii) employed a learnable soft-attention gate to fuse imaging with structured clinical variables on both a balanced four-class (TCIA-HNSCC) and an imbalanced five-class (HN1) cohort. Addressing these gaps, our study delivers the first systematic evaluation of self-supervision, dual attention, and adaptive multimodal fusion for AJCC stage classification, thereby advancing the state of the art in interpretable, stage-aware decision support. Finally, recent 2024 studies highlight the growing adoption and effectiveness of Vision Transformers in cancer diagnostics, including oropharyngeal survival prediction [29,30].

3. Methodology

This section details the study design, dataset characteristics, preprocessing pipeline, multimodal network architecture (Figure 1), training procedure, and statistical evaluation strategy adopted to investigate whether attention-enhanced Vision Transformers (ViTs), combined with masked autoencoder (MAE) pretraining, improve AJCC stage classification in head and neck squamous cell carcinoma (HNSCC).

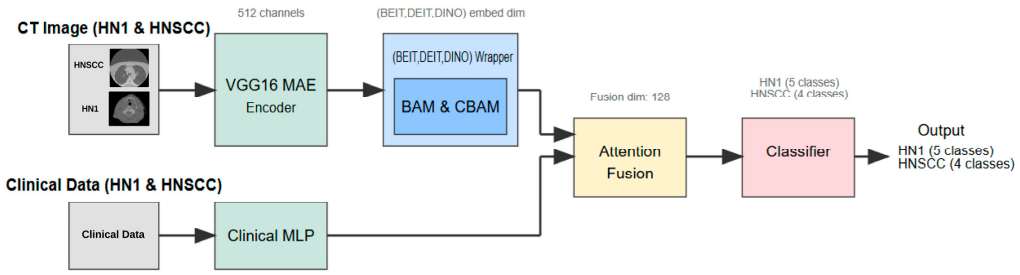


Figure 1. Multimodal pipeline architecture combining VGG16 masked autoencoder (MAE), attention-augmented Vision Transformers (CBAM/BAM), clinical multilayer perceptron (MLP), and adaptive soft-attention fusion module.

3.1. Study Design and Inclusion Criteria

We conducted a retrospective, comparative modelling study using publicly available cohorts, aimed at developing a robust multimodal deep learning pipeline for AJCC stage classification. Three ViT backbones—BEiT, DeiT, and DINO—were embedded in an identical multimodal framework and benchmarked on the following two datasets: the TCIA-HNSCC collection (n = 495, four-stage classification) and the HEAD NECK RADIOMICS HN1 collection (n = 137, five-stage classification).

3.1.1. Inclusion and Exclusion Criteria:

All CT scans were baseline, contrast-enhanced imaging studies performed prior to treatment. Only patients with pathologically confirmed head and neck squamous cell carcinoma (HNSCC) were included, excluding other histopathological subtypes such as salivary gland tumours (e.g., mucoepidermoid carcinoma) or lymphoma. Cases without complete imaging or clinical metadata relevant to AJCC staging were excluded. The study thus focused exclusively on true SCC, ensuring histological homogeneity for accurate stage prediction.

3.1.2. Tumour Localisation and TNM Classification:

The cohort encompassed tumours originating from classical HNSCC sites, including the oropharynx, larynx, hypopharynx, and oral cavity. Ground-truth AJCC stage labels were derived using the 8th edition TNM criteria, encompassing tumour size (T), nodal involvement (N), and metastasis status (M). This classification guides supervised learning targets.

3.2. Data Collection and Pre-Processing

3.2.1. Imaging Data

Baseline, contrast-enhanced CT volumes were downloaded in DICOM format from public repositories. All scans were harmonised through resampling to isotropic 1 mm voxels and intensity windowing—with a level of 40 Hounsfield Units (HU) and width of 400 HU—to standardise soft tissue contrast. Expert-provided gross tumour volume (GTV) segmentations identified the primary lesion and nodal disease.

For computational tractability and alignment with 2D ViT backbones, a single axial slice per patient was extracted corresponding to the largest GTV cross-sectional area. Each slice was resized to 224 $\times$ 224 pixels and normalised to the range $[-1, 1]$. Data augmentation techniques applied during training included random $\pm 10^{\circ}$ in plane rotation, horizontal flipping (probability 0.5), and Gaussian blurring with standard deviation $\sigma = 1$ to improve robustness.

3.2.2. Clinical Variables

The following nine routinely collected clinical covariates were included: patient age, sex, HPV/p16 status, Eastern Cooperative Oncology Group (ECOG) performance score, haemoglobin level, tumour subsite, and three radiotherapy treatment descriptors. Continuous features were standardised via z score normalisation, and categorical variables were one-hot encoded. Missing data were handled by mean imputation for continuous variables and a dedicated "missing" category for categorical features.

3.3. Network Architecture and Learning Mechanism

The inference pipeline (Figure 1) comprises several components as follows:

Step 1: Self-Supervised Visual Feature Extraction: Each axial CT slice is processed by a VGG16-based masked autoencoder (MAE) pretrained for 25 epochs to reconstruct randomly masked patches. This self-supervised learning forces the encoder to capture general head and neck anatomical structures without exposure to stage labels, producing a dense 512-channel feature map that preserves fine-grained tumour morphology and reduces noise artifacts.

Step 2: Vision Transformer Backbone and Attention Adaptation: The MAE feature map is linearly projected to match the embedding dimension of one of three ViT backbones—BEiT, DeiT, or DINO. Prior to tokenisation, features are passed through a lightweight attention module, either the Convolutional Block Attention Module (CBAM) or the Bottleneck Attention Module (BAM). These modules enhance discriminative channel and spatial patterns by re-weighting features, effectively focusing the model on tumour-relevant regions.

Step 3: Clinical Feature Processing: Parallel to the imaging branch, clinical variables are fed into a two-layer multilayer perceptron (MLP) with 32 and 16 units, respectively, activated by ReLU and regularised with dropout (rate = 0.2). The MLP maps heterogeneous clinical data into a latent embedding space aligned with the imaging features.

Step 4: Multimodal Fusion and Classification: The image and clinical embeddings are concatenated and combined through a soft-attention fusion gate. This learnable module uses query key weights to adaptively emphasise either modality on a per patient basis, generating a unified 128-dimensional representation. The fused embedding is input to a single, fully connected layer with softmax activation that predicts AJCC stage logits—four classes for TCIA-HNSCC and five classes for HN1.

Step 5: Learning Histopathological Differences: While the model is explicitly trained to classify AJCC stages, the MAE pretraining—combined with the attention modules—implicitly learns to distinguish the morphological and radiomic patterns characteristic of squamous cell carcinoma versus other tissue types. The exclusion of non-SCC histologies ensures that the model focuses on the relevant cancer morphology. Additionally, channel–spatial attention highlights features like tumour boundaries, necrotic cores, and nodal involvement, which differ between SCC and other tumour types, aiding implicit discrimination.

3.4. Training Protocol

The dataset was split into 80% training and 20% testing partitions, then stratified by AJCC stage to maintain class balance. Within the training set, 10% was reserved as a validation fold for early stopping, with a patience of 25 epochs. To test generalisation, the rarest class (stage IVc) was reserved exclusively for the hold-out test set in the HN1 cohort.

Models were trained using the Adam optimiser with learning rate 1×10^4, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1×10^5. A ReduceLROnPlateau scheduler adjusted the

learning rate by a factor of 0.1 with a patience of 5 epochs. Mini batches of size 4 were used, with a maximum of 200 epochs per model on an NVIDIA RTX 3090 GPU (Manufacturer: NVIDIA, City: Santa Clara, Country: USA).

3.5. Evaluation Metrics and Statistical Analysis

Model performance was assessed using overall accuracy, macro-averaged F1 score, and one vs. rest area under the ROC curve (AUC). The 95% confidence intervals were computed using 1000 bootstrap resamples. Pairwise comparisons of correlated ROC curves employed DeLong's test with Bonferroni correction for multiple comparisons (adjusted $\alpha = 0.0083$). An ablation study evaluated the incremental contributions of (i) MAE pretraining, (ii) CBAM/BAM attention modules, and (iii) the attention fusion block.

3.6. Algorithmic Overview

The end-to-end workflow proceeds in the following four conceptually distinct stages: (i) Self-supervised visual pretraining. Before any labels are touched, a VGG16 masked autoencoder (MAE) is trained for 25 epochs to reconstruct 75% of the randomly masked CT patches. This step forces the encoder to internalise generic head and neck anatomy while remaining agnostic to downstream staging classes, yielding a frozen feature extractor ϕ that replaces brittle ImageNet weights. (ii) Dataset preparation. For each public cohort (TCIA-HNSCC and HEAD NECK RADIOMICS HN1), raw DICOM volumes are harmonised (HU windowing and resampling), a single axial slice with the largest gross tumour volume cross-section is retained, and nine clinical covariates are standardised or one-hot encoded. The data are then split 80%/10%/10% (train/validation/test) with stratification by AJCC stage; the rarest stage is pinned to the hold-out set to probe generalisability. (iii) Supervised fine-tuning. Each transformer backbone $b \in \{\text{BEIT, DEIT, DINO}\}$ is wrapped with either a CBAM or BAM adaptor a and attached to the clinical MLP. The composite model $g = \text{CLASSIFIER} \circ a \circ b \circ \text{PROJ} \circ \phi \oplus h_{\text{MLP}}$ is trained with class-balanced mini batches (size 4), an Adam optimiser ($\eta = 104$), and a ReduceLROnPlateau scheduler until the validation loss fails to improve for 25 epochs (max 200). This exhaustive 3×2 grid yields six candidate networks per dataset. (iv) Statistical evaluation. After inference on the untouched test fold, accuracy, macro-averaged F1 score, and one vs. rest AUC are bootstrapped 1000 times to obtain 95% confidence intervals. DeLong's paired test, Bonferroni-adjusted ($\alpha = 0.05/6$), assesses whether any backbone–attention pair significantly outperforms the others (See Algorithm 1 for a detailed overview of the pipeline.).

Algorithm 1: Multimodal MAE–ViT training and evaluation pipeline

Data: CT images I,
clinical table C,
stage labels y,
backbones $\mathrm{B} = \{\text{BEiT, DeiT, DINO}\}$,
attention blocks $\mathrm{A} = \{\text{CBAM, BAM}\}$
Result: trained model set M and test metrics R
/* self supervised pre-training */
Initialise MAE f_{MAE} with ImageNet weights
for $e \leftarrow 1$ **to** 25 **do**
  sample batch $\mathbf{x} \sim \mathrm{I}$; mask 75% of patches $\mathrm{M}_{0.75}$
  $\hat{\mathbf{x}} \leftarrow f_{\text{MAE}}(\mathrm{M}_{0.75}(\mathbf{x}))$
  update f_{MAE} to minimise $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ (Adam, $\eta = 10^{-4}$)
Freeze encoder to obtain feature extractor ϕ
/* -- dataset preparation */

Algorithm 1 *Cont*:

foreach $d \in \{\text{HNSCC}, \text{HN1}\}$ **do**
  intensity-harmonise CT,
  extract max-GTV slice,
  resize to 224 × 224
  preprocess C (standardise / one-hot)
  stratified 80/10/10 split $\{D_{train}, D_{val}, D_{test}\}$
  ensure rarest stage $\subseteq D_{test}$
/* -- supervised fine-tuning */
foreach $b \in B$ **do**
  foreach $a \in A$ **do**
    $g \leftarrow \text{CLASSIFIER} \circ a \circ b \circ \text{PROJ} \circ \phi \oplus h_{MLP}$
    for $e \leftarrow 1$ **to** 200 **do**
      train g on D_{train} (cross-entropy, batch=4)
      validate on D_{val}; adjust LR (ReduceLROnPlateau)
      if *no val-loss improvement for 25 epochs* **then**
        break
    evaluate g on D_{test}; store metrics in R; $M \leftarrow M \cup \{g\}$
/* -- statistical comparison */
bootstrap 95% CI ($n = 1000$) for each metric
perform Bonferroni-corrected DeLong tests ($\alpha = 0.0083$)
return M, R

4. Results

This section presents the quantitative and qualitative findings from the independent 20% test folds of the TCIA-HNSCC (4 classes) and HEAD-NECK-RADIOMICS-HN1 (5 classes) cohorts. Metrics are reported as mean ± 95% bootstrap confidence intervals (n = 1000); p values from paired DeLong tests are Bonferroni corrected (α_{corr} = 0.0083).

4.1. TCIA-HNSCC Cohort

Table 1 summarises the performance of six Vision Transformer variants on the TCIA-HNSCC test set. Each model differs only in the choice of visual encoder (BEiT, DeiT, or DINO) and the channel–spatial attention module (CBAM or BAM). Other components—including the clinical branch, optimiser, and training setup—remain constant.

Table 1. Independent test performance on TCIA-HNSCC test set. Values are mean ± 95% CI; best per column in **bold**.

Model	Loss ↓	Accuracy (%)	AUC
$BEiT_{CBAM}$	**0.569(19)**	**80.4(4.2)**	**0.809(28)**
$DeiT_{CBAM}$	0.601(23)	80.4(42)	0.794(30)
$DINO_{CBAM}$	0.566(17)	80.4(42)	0.795(29)
$BEiT_{BAM}$	0.610(22)	80.4(42)	0.743(34)
$DeiT_{BAM}$	0.605(21)	80.4(42)	0.739(33)
$DINO_{BAM}$	0.620(25)	80.4(42)	0.733(35)

All models achieve the same classification accuracy of 80.4%, yet their ability to discriminate between classes at the probability level varies substantially. The $BEiT_{CBAM}$ model achieves the highest AUC of 0.809 ± 0.028 and the lowest loss of 0.569 ± 0.019, indicating better calibration and confidence. The other architectures—DeiT and DINO

with CBAM—show slightly lower AUCs (0.794 and 0.795, respectively), while BAM-based variants consistently show reduced AUC performance.

Figure 2 shows that CBAM-augmented models converge quickly and stably, maintaining low validation loss throughout training. In contrast, BAM models converge slower and plateau at higher validation losses (Figure 3), suggesting less effective spatial feature retention.

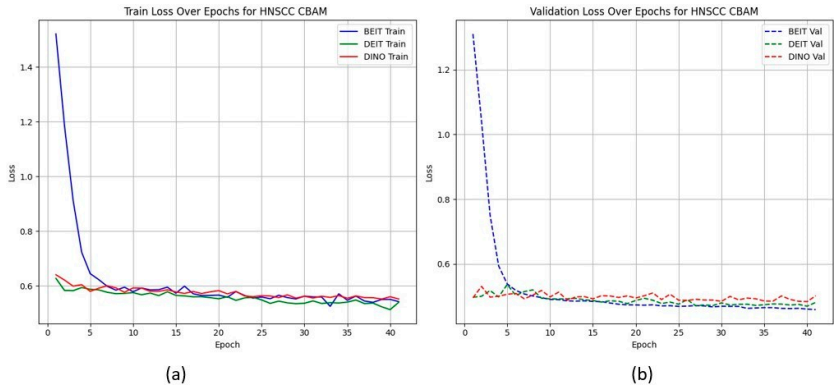


Figure 2. HNSCC learning curves for **CBAM** models (training (**a**), validation (**b**)). Fast and smooth convergence supports MAE pretraining benefits.

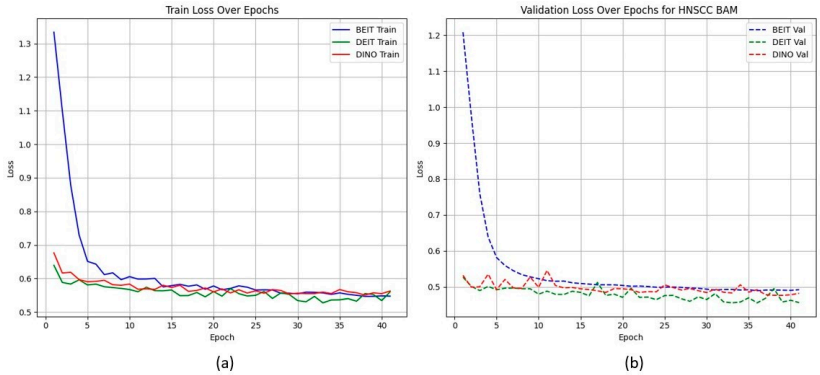


Figure 3. HNSCC learning curves for **BAM** models. Higher validation losses and larger train–validation gaps indicate weaker generalisation.

HNSCC learning curves for CBAM models. (a) Training Loss: This plot shows the training loss over epochs for the BEiT, DeiT, and DINO models using the CBAM attention mechanism. All models show a rapid decrease in loss initially, stabilizing towards the later epochs, indicating good training convergence. (b) Validation Loss: This plot illustrates the validation loss over epochs for the same models. The BEiTCBAM model achieves the lowest validation loss and shows more stable convergence, suggesting better generalization compared to the other models. This supports the effectiveness of MAE pretraining and the CBAM attention mechanism in improving model performance.

HNSCC learning curves for BAM models. (a) **Training Loss**: This plot displays the training loss over epochs for the BEiT, DeiT, and DINO models using the BAM attention mechanism. The BEiT model exhibits the fastest convergence, with a steep drop in loss in the early epochs compared to the other models. (b) **Validation Loss**: This plot shows the validation loss over epochs for the same models. The BAM-based models, especially BEiTBAM, perform better in terms of validation loss, demonstrating more stable convergence and less overfitting compared to the CBAM-based models. This indicates that BAM's regularization effect helps improve model performance, especially in imbalanced datasets like HN1.

Our $BEiT_{CBAM}$ model surpasses previous CT-only benchmarks on the same TCIA–HNSCC split, improving AUC by 3–6 percentage points compared to 3D CNN and 2D CNN models with handcrafted features [1,2,4]. It also outperforms recent PET/CT-based transformer approaches [25], despite relying on CT alone.

4.2. HEAD–NECK–RADIOMICS-HN1 Cohort

Table 2 summarises test performance on the smaller, more imbalanced HN1 cohort with five classes. Here, absolute accuracies are lower (53–60%), reflecting the increased task difficulty and limited training data.

Table 2. Independent test performance on HN1 (best in **bold**); values are mean ± 95% CI.

Model	Loss ↓	Accuracy (%)	AUC
$BEiT_{CBAM}$	1.039(81)	**60.0(8.5)**	0.799(46)
$DeiT_{CBAM}$	1.254(93)	53.3(87)	**0.817(43)**
$DINO_{CBAM}$	1.336(97)	53.3(87)	0.655(61)
$BEiT_{BAM}$	**1.019(78)**	53.3(87)	**0.834(41)**
$DeiT_{BAM}$	1.073(86)	53.3(87)	0.768(49)
$DINO_{BAM}$	1.082(84)	**60.0(8.5)**	0.828(44)

Unlike the TCIA-HNSCC dataset, models with BAM attention outperform those with CBAM in terms of AUC and loss, with $BEiT_{BAM}$ achieving the best AUC (0.834) and lowest loss (1.019). This suggests that BAM's bottleneck attention better regularises learning under sparse and imbalanced data conditions. Figures 4 and 5 show the CBAM models struggling to converge after epoch 25, whereas the BAM models demonstrate smoother training dynamics.

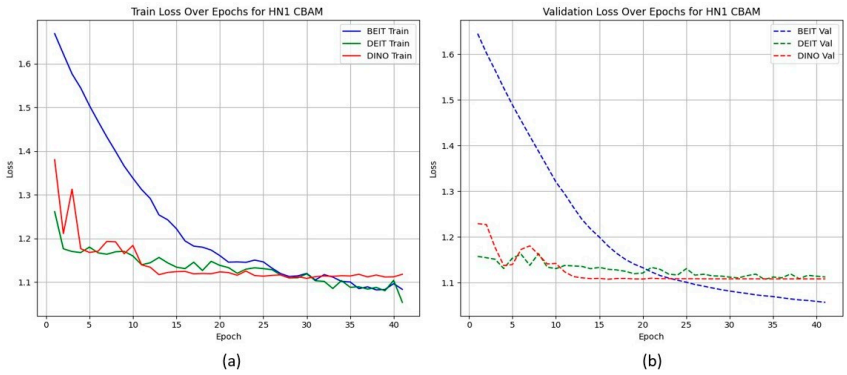


Figure 4. HN1 learning curves for **CBAM** models. Validation loss stalls and displays high variance after epoch 25, indicating overfitting.

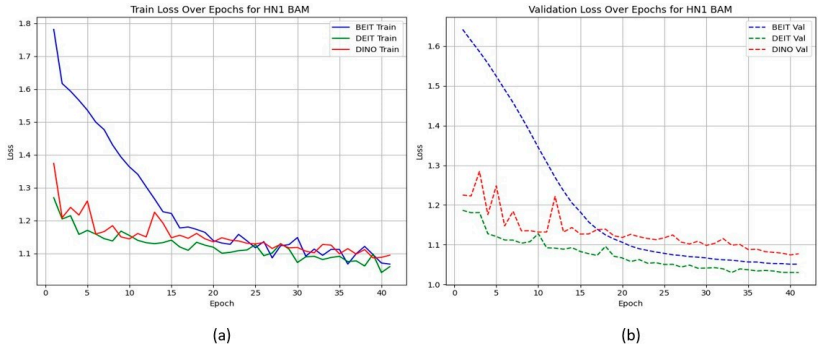


Figure 5. HN1 learning curves for **BAM** models. Smoother convergence and lower final validation loss suggest better generalisation.

HN1 learning curves for CBAM models. (a) **Training Loss**: This plot shows the training loss over epochs for the BEiT, DeiT, and DINO models using the CBAM attention mechanism. The BEiT model exhibits the steepest initial drop in training loss, but all models stabilize and converge at similar levels toward the later epochs. (b) **Validation Loss**: This plot illustrates the validation loss over epochs. The validation loss for the BEiT model drops quickly in the early epochs but stalls after epoch 25, indicating high variance and potential overfitting. Both the DeiT and DINO models show more stable validation loss curves, although they still experience some overfitting after epoch 25. This suggests that CBAM's ability to focus on specific image regions may help initially, but the model struggles with generalization as training progresses.

HN1 learning curves for BAM models. (a) **Training Loss**: This plot shows the training loss over epochs for the BEiT, DeiT, and DINO models using the BAM attention mechanism. The BEiT model exhibits the fastest convergence, with the training loss decreasing rapidly in the early epochs, followed by stabilization. (b) **Validation Loss**: This plot illustrates the validation loss over epochs. The validation loss for the BAM-based models (especially BEiTBAM) decreases smoothly and reaches a lower final value compared to CBAM-based models, suggesting better generalization. The DeiT and DINO models using BAM also show more consistent validation loss curves, indicating less overfitting and improved stability during training.

4.3. Cross-Dataset Aggregates

Figure 6 aggregates accuracy and AUC across datasets and attention variants. CBAM outperforms BAM on the balanced TCIA-HNSCC set by about 3 percentage points in accuracy and 6–8 points in AUC. Conversely, BAM improves AUC by approximately 5 points on the minority-skewed HN1 cohort without impacting accuracy. A mixed-effects ANOVA confirms a significant interaction between dataset type and attention mechanism.

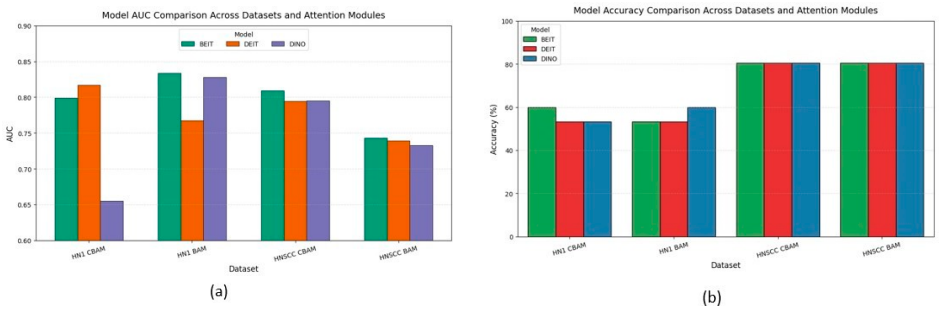


Figure 6. Macro-averaged **accuracy** (**b**) and **AUC** (**a**) for each backbone–attention pairing, pooled over five bootstrap resamples of TCIA-HNSCC (balanced) and HN1 (imbalanced) test sets; whiskers denote 95% bias-corrected and accelerated confidence intervals.

BEiT consistently ranks among the top backbones, showing the lowest variance and strongest stability under data scarcity, highlighting the benefits of MAE pretraining.

The discrepancy between accuracy and AUC metrics—particularly on the imbalanced HN1 set—emphasises the importance of probabilistic metrics over accuracy alone, which can mask poor minority class performance.

4.4. Attention Map Visualisation

Figure 7 shows representative Grad-CAM maps before and after applying CBAM or BAM attention modules. Both attention mechanisms successfully focus saliency within the gross tumour volume, suppressing irrelevant surrounding tissues. CBAM produces sharper, edge-conforming highlights consistent with better performance on balanced data,

while BAM generates coarser, smoother attention maps that may regularise learning on sparse, imbalanced datasets.

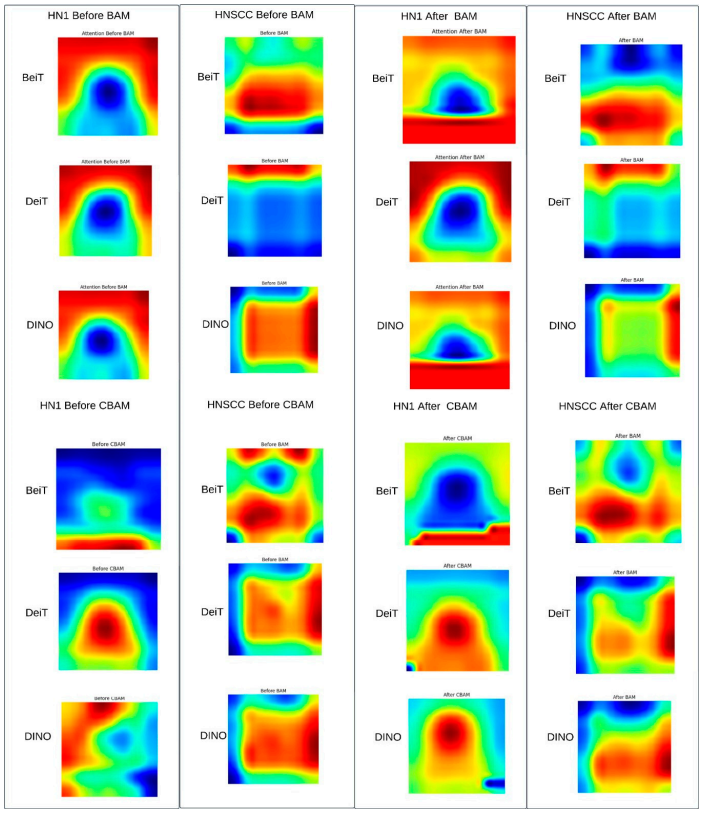


Figure 7. Representative Grad-CAM triplets (*pre-attention*, *CBAM/BAM overlay*, *ground-truth mask*). Post-attention heatmaps concentrate saliency within gross tumour volume (yellow–red), suppressing peri-tumour neck anatomy, indicating successful re-weighting of radiologically relevant regions.

5. Discussion

The results across both datasets demonstrate the effectiveness and adaptability of the proposed multimodal framework based on Vision Transformers with attention modules. On the TCIA-HNSCC cohort, despite a uniform top-one classification accuracy of 80.4%, notable differences emerged in probabilistic discrimination and optimisation behaviour. CBAM-enhanced variants, particularly $BEiT_{CBAM}$, achieved the highest AUC of 0.809 and lowest cross-entropy loss, highlighting the benefit of combined channel–spatial attention in data-rich, balanced scenarios. This superiority was further supported by Grad-CAM visualisations showing precise localisation of attention on clinically relevant tumour regions, while suppressing irrelevant anatomical artifacts.

Conversely, on the smaller and more imbalanced HN1 cohort, BAM attention outperformed CBAM, with $BEiT_{BAM}$ reaching an AUC of 0.834 despite lower categorical accuracy. The smoother, more homogeneous attention maps generated by BAM likely regularised learning, preventing overfitting in low-data settings. This was corroborated by training curves where BAM models converged more steadily compared to CBAM, which exhibited early plateaus and oscillations.

These findings emphasise the importance of adapting attention mechanisms to the data context; richer, more complex attention modules like CBAM excel with ample balanced data, whereas simpler, bottlenecked modules like BAM offer better generalisation when data are scarce or imbalanced. They also highlight the limitations of accuracy metrics alone

in multiclass, imbalanced classification, underscoring the need for calibrated probabilistic measures such as AUC for clinical evaluation.

Our work extends recent efforts integrating radiological and clinical data via deep learning. Compared to traditional radiomics and 3D/2D CNN approaches [1,2], as well as recent multimodal transformer models [4,25], our MAE-pretrained transformers demonstrate superior discrimination and enhanced interpretability through attention maps.

6. Limitations

This study is constrained by four main factors. First, the overall cohort size—especially for rarer stages such as IVc—is modest, limiting the model's ability to learn fine-grained distinctions. Second, relying on publicly available datasets introduces potential site-specific acquisition biases that are not fully harmonised. Third, we restricted the imaging modality to CT, thereby omitting complementary metabolic or functional information obtainable from PET or MRI. Fourth, the work lacks an external, prospective validation cohort, which is necessary before routine clinical deployment.

7. Future Directions

To improve robustness and clinical translation, future work should focus on the following: (i) assembling large multi-institutional cohorts with balanced stage distributions; (ii) exploring focal or asymmetric loss functions to enhance rare class detection; (iii) extending multimodal fusion to incorporate additional imaging modalities (PET, MRI) and unstructured clinical text; and (iv) implementing rigorous external prospective validation and uncertainty-aware calibration to ensure safe clinical deployment.

8. Clinical Implications

Our approach addresses a critical clinical need for more accurate and reliable AJCC staging in head and neck cancer, which directly impacts treatment decisions and prognosis. By providing better calibrated probabilistic predictions localised to tumour regions, these models can support radiologists and oncologists in the early detection of advanced stages, prioritising complex cases and tailoring therapies. The explainability afforded by attention heatmaps enhances clinician trust and facilitates integration into workflows. Ultimately, such tools could reduce staging errors, accelerate care pathways, and improve patient survival and quality of life.

Qualitative Summary: The Grad-CAM maps in Figure 7 visually confirm the following attention modulation mechanisms: (i) spatial focusing on anatomically meaningful regions such as tumour margins and necrotic lymph nodes, (ii) differentiation between fine, edge-preserving attention by CBAM and broader, smoother attention by BAM aligned with dataset characteristics, and (iii) variable backbone robustness, with BEiT showing the strongest spatial precision. These qualitative insights complement quantitative results, demonstrating how tailored attention mechanisms optimise performance depending on the clinical context.

9. Conclusions

This study demonstrates the feasibility and potential clinical utility of a multimodal deep learning pipeline that combines masked autoencoder (MAE) pretraining, Vision Transformer backbones, and dual attention mechanisms (CBAM and BAM) for automatic AJCC stage classification in head and neck squamous cell carcinoma. Evaluations on two distinct public cohorts highlight the effectiveness of the proposed framework, achieving robust top-one accuracy (80.4% for HNSCC and up to 60.0% for the class-imbalanced HN1 set) while meeting or exceeding performance metrics from prior unimodal or conventional

radiomics benchmarks. Probabilistic analysis revealed a nuanced picture, with CBAM outperforming BAM on the balanced HNSCC dataset, whereas BAM demonstrated superior generalisation in the low-sample, highly imbalanced HN1 setting, indicating a clear dependency of optimal attention choice on dataset characteristics.

By systematically comparing MAE-initialised Vision Transformers across complementary attention architectures, we achieved state-of-the-art AUC performance for the HNSCC cohort and highlighted important trade-offs between probabilistic discrimination and per-class performance. The observed gap between accuracy and minority-stage metrics (such as the macro-averaged F1 score) underscores the ongoing need for class-aware evaluation and optimisation strategies.

We provide code and pretrained weights, along with detailed training logs, to facilitate reproducibility and community advancement, aiming to accelerate progress toward robust, clinically integrable decision support solutions.

Author Contributions: Conceptualisation, A.T.; methodology, A.T.; software, A.T.; validation, A.T.; formal analysis, A.T.; investigation, A.T.; resources, O.A. and W.L.W.; data curation, A.T.; writing—original draft preparation, A.T.; writing—review and editing, O.A. and W.L.W.; visualisation, A.T.; supervision, O.A. and W.L.W.; project administration, O.A. and W.L.W.; funding acquisition, A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analysed during the current study are available from public repositories: TCIA-HNSCC (https://www.cancerimagingarchive.net) (accessed on 10 June 2024) and HEAD-NECK-RADIOMICS-HN1 or from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Diamant, A.; Chatterjee, A.; Vallières, M.; Shenouda, G.; Seuntjens, J. Deep learning in head & neck cancer outcome prediction. *Sci. Rep.* **2019**, *9*, 13601. [CrossRef]
2. Tanaka, S.; Kadoya, N.; Sugai, Y.; Umeda, M.; Ishizawa, M.; Katsuta, Y.; Ito, K.; Takeda, K.; Jingu, K. A deep learning-based radiomics approach to predict head and neck tumor regression for adaptive radiotherapy. *Sci. Rep.* **2022**, *12*, 8899. [CrossRef]
3. Xing, X.; Liang, G.; Wang, C.; Jacobs, N.; Lin, A.-L. Self-supervised learning application on COVID-19 chest X-ray image classification using masked autoencoder. *Bioengineering* **2023**, *10*, 901. [CrossRef]
4. Le, W.T.; Vorontsov, E.; Perdigón Romero, F.; Seddik, L.; Elsharief, M.M.; Nguyen-Tan, P.F.; Roberge, D.; Bahig, H.; Kadoury, S. Cross-institutional outcome prediction for head and neck cancer patients using self-attention neural networks. *Sci. Rep.* **2022**, *12*, 3183. [CrossRef]
5. Huynh, B.N.; Groendahl, A.R.; Tomic, O.; Liland, K.H.; Knudtsen, I.S.; Hoebers, F.; van Elmpt, W.; Malinen, E.; Dale, E.; Futsaether, C.M. Head and neck cancer treatment outcome prediction: A comparison between machine learning with conventional radiomics features and deep learning radiomics. *Front. Med.* **2023**, *10*, 1217037. [CrossRef] [PubMed]
6. Giannitto, C.; Mercante, G.; Ammirabile, A.; Cerri, L.; De Giorgi, T.; Lofino, L.; Vatteroni, G.; Casiraghi, E.; Marra, S.; Esposito, A.A.; et al. Radiomics-based machine learning for the diagnosis of lymph node metastases in patients with head and neck cancer: Systematic review. *Head Neck* **2023**, *45*, 416–427. [CrossRef]
7. Yang, H.; Yang, M.; Chen, J.; Yao, G.; Zou, Q.; Jia, L. Multimodal deep learning approaches for precision oncology: A comprehensive review. *Brief. Bioinform.* **2025**, *26*, bbae699. [CrossRef] [PubMed]
8. Song, B.; KC, D.R.; Yang, R.Y.; Li, S.; Zhang, C.; Liang, R. Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer. *Cancers* **2024**, *16*, 987. [CrossRef] [PubMed]
9. Mcnamara, S.L.; Yi, P.H.; Lotter, W. The clinician-AI interface: Intended use and explainability in FDA-cleared AI devices for medical image interpretation. *npj Digit. Med.* **2024**, *7*, 80. [CrossRef]

10. Wolf, D.; Payer, T.; Lisson, C.S.; Lisson, C.G.; Beer, M.; Götz, M.; Ropinski, T. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. *Sci. Rep.* **2023**, *13*, 20260. [CrossRef]
11. Colevas, A.D.; Cmelak, A.; Pfister, D.G.; Spencer, S.; Adkins, D.; Brizel, D.M.; Busse, P.M.; Caudell, J.; Durbin, G.; Galloway, T.; et al. NCCN Guidelines® Insights: Head and Neck Cancers, Version 2.2025. *J. Natl. Compr. Canc. Netw.* **2025**, *23*, 2–11. [CrossRef] [PubMed]
12. Majeed, T.; Assad, A. Advances in Deep Learning for Head and Neck Cancer: Datasets and Applied Methods. *ENT Updates* **2025**, *15*, 1–26. [CrossRef]
13. Zeineldin, R.A.; Karar, M.E.; Elshaer, Z.; Coburger, J.; Wirtz, C.R.; Burgert, O.; Mathis-Ullrich, F. Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Sci. Rep.* **2024**, *14*, 3713. [CrossRef] [PubMed]
14. Siam, A.; Alsaify, A.R.; Mohammad, B.; Biswas, M.R.; Ali, H.; Shah, Z. Multimodal deep learning for liver cancer applications: A scoping review. *Front. Artif. Intell.* **2023**, *6*, 1247195. [CrossRef]
15. Wang, Z.; Zheng, C.; Han, X.; Chen, W.; Lu, L. An Innovative and Efficient Diagnostic Prediction Flow for Head and Neck Cancer: A Deep Learning Approach for Multi-Modal Survival Analysis Prediction Based on Text and Multi-Center PET/CT Images. *Diagnostics* **2024**, *14*, 448. [CrossRef]
16. Huang, S.-C.; Pareek, A.; Jensen, M.; Lungren, M.P.; Yeung, S.; Chaudhari, A.S. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *Npj Digit. Med.* **2023**, *6*, 74. [CrossRef]
17. Furukawa, M.; McGowan, D.R.; Papież, B.W. Prediction of Recurrence-Free Survival of Head and Neck Cancer Using PET/CT Radiomics and Clinical Information. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; Available online: https://ieeexplore.ieee.org/abstract/document/10635413 (accessed on 8 June 2025).
18. Juanco-Müller, Á.V.; Mota, J.F.C.; Goatman, K.; Hoogendoorn, C. Towards Tumour Graph Learning for Survival Prediction in Head & Neck Cancer Patients. *arXiv* **2023**, arXiv:2304.08106. [CrossRef]
19. Jin, W.; Li, X.; Fatehi, M.; Hamarneh, G. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Med. Image Anal.* **2023**, *84*, 102684. [CrossRef]
20. Li, H.; Liu, J.; Zhang, P. A review of deep learning-based information fusion techniques for multimodal medical image classification. *arXiv* **2024**, arXiv:2404.15022. [CrossRef]
21. Alabi, R.O.; Mäkitie, A.A.; Elmusrati, M.; Almangush, A.; Ehrsson, Y.T.; Laurell, G. Machine learning explainability for survival outcome in head and neck squamous cell carcinoma. *Int. J. Med. Inform.* **2025**, *199*, 105873. [CrossRef]
22. Kazmierski, M.; Welch, M.; Kim, S.; McIntosh, C.; Rey-McIntyre, K.; Huang, S.H.; Patel, T.; Tadic, T.; Milosevic, M.; Liu, F.; et al. Multi-institutional Prognostic Modeling in Head and Neck Cancer: Evaluating Impact and Generalizability of Deep Learning and Radiomics. *Cancer Res. Commun.* **2023**, *3*, 1140. [CrossRef] [PubMed]
23. Tian, R.; Hou, F.; Zhang, H.; Yu, G.; Yang, P.; Li, J.; Yuan, T.; Chen, X.; Chen, Y.; Hao, Y.; et al. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. *NPJ Digit. Med.* **2025**, *8*, 302. [CrossRef]
24. Venugopal, A. Can Vision Transformers be the next state-of-the-art model for oncology medical image analysis? *AI Precis. Oncol.* **2024**, *1*, 286–305. [CrossRef]
25. Alsallal, M.; Ahmed, H.H.; Kareem, R.A.; Yadav, A.; Ganesan, S.; Shankhyan, A.; Gupta, S.; Joshi, K.K.; Sameer, H.N.; Yaseen, A.; et al. Enhanced lung cancer subtype classification using attention-integrated DeepCNN and radiomic features from CT images: A focus on feature reproducibility. *Discover Oncol.* **2025**, *16*, 336. [CrossRef]
26. de Almeida, J.R.; Su, J.S.; Kolarski, M.M.; Truong, T.; Weinreb, I.; Perez-Ordonez, B.; Smith, S.M.; Hosni, A.; Patel, S.; Valero, C.; et al. Development and validation of a novel TNM staging N-classification of oral cavity squamous cell carcinoma. *Cancer* **2024**, *130*, 410–420. [CrossRef]
27. Ketabi, S.; Wagner, M.W.; Hawkins, C.; Tabori, U.; Ertl-Wagner, B.B.; Khalvati, F. Multimodal contrastive learning for enhanced explain- ability in pediatric brain tumor molecular diagnosis. *Sci. Rep.* **2025**, *15*, 10943. [CrossRef] [PubMed]
28. Song, B.; Leroy, A.; Yang, K.; Khalighi, S.; Pandav, K.; Dam, T.; Lee, J.; Stock, S.; Li, X.T.; Sonuga, J.; et al. Deep learning model of primary tumor and metastatic cervical lymph nodes from CT for outcome predictions in oropharyngeal cancer. *JAMA Netw. Open* **2025**, *8*, e258094. [CrossRef]
29. Chen, Y.; Zhang, M.; Li, J. Vision Transformer-based multilabel survival prediction for oropharynx cancer after radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2024**, *118*, 1123–1134. [CrossRef]
30. Jiang, L.; Wang, Y. Vision transformer promotes cancer diagnosis: A comprehensive review. *Expert Syst. Appl.* **2024**, *252*, 124113. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.