radiation

MDPI

Article

Deep Texture Analysis—Enhancing CT Radiomics Features for Prediction of Head and Neck Cancer Treatment Outcomes: A Machine Learning Approach

Aryan Safakish [1,2], Lakshmanan Sannachi [1], Amir Moslemi [1], Ana Pejović-Milić [2] and Gregory J. Czarnota [1,2,3,4,5,*]

[1] Physical Sciences, Sunnybrook Research Institute, Toronto, ON M4N 3M5, Canada; aryan.safakish@sri.utoronto.ca (A.S.)
[2] Department of Physics, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada
[3] Department of Radiation Oncology, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada
[4] Departments of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
[5] Departments of Radiation Oncology, University of Toronto, Toronto, ON M5T 1P5, Canada
* Correspondence: gregory.czarnota@sunnybrook.ca

Simple Summary: Cancer treatment is a physically and emotionally stressful experience for patients. Some patients benefit from treatment, whereas others do not. In order to predict a variety of biological endpoints, radiomics features can be determined from biomedical images and used to train predictive machine learning (ML) models. In this work, treatment-planning computed tomography (CT) scans of head and neck (H&N) cancer patients were used to identify radiomics features and train ML models to predict binary treatment response as determined clinically three months post-treatment. By providing insights about potential treatment response, reliable predictive models would benefit patients by giving clinicians a useful tool in delivering personalized medical care. Furthermore, in this work, deeper layer texture features were investigated, and the results suggest that the inclusion of deeper layer radiomics features enhanced the predictive value in training ML models.

Abstract: (1) Background: Some cancer patients do not experience tumour shrinkage but are still at risk of experiencing unwanted treatment side effects. Radiomics refers to mining biomedical images to quantify textural characterization. When radiomics features are labelled with treatment response, retrospectively, they can train predictive machine learning (ML) models. (2) Methods: Radiomics features were determined from lymph node (LN) segmentations from treatment-planning CT scans of head and neck (H&N) cancer patients. Binary treatment outcomes (complete response versus partial or no response) and radiomics features for $n = 71$ patients were used to train support vector machine (SVM) and k-nearest neighbour (k-NN) classifier models with 1–7 features. A deep texture analysis (DTA) methodology was proposed and evaluated for second- and third-layer radiomics features, and models were evaluated based on common metrics (sensitivity (%S_n), specificity (%S_p), accuracy (%Acc), precision (%Prec), and balanced accuracy (%Bal Acc)). (3) Results: Models created with both classifiers were found to be able to predict treatment response, and the results suggest that the inclusion of deeper layer features enhanced model performance. The best model was a seven-feature multivariable k-NN model trained using features from three layers deep of texture features with %S_n = 74%, %S_p = 68%, %Acc = 72%, %Prec = 81%, %Bal Acc = 71% and with an area under the curve (AUC) the receiver operating characteristic (ROC) of 0.700. (4) Conclusions: H&N Cancer patient treatment-planning CT scans and LN segmentations contain phenotypic information regarding treatment response, and the proposed DTA methodology can improve model performance by enhancing feature sets and is worth consideration in future radiomics studies.

Keywords: radiomics; head and neck cancer; deep texture analysis; texture of texture; response prediction; texture features

Citation: Safakish, A.; Sannachi, L.; Moslemi, A.; Pejović-Milić, A.; Czarnota, G.J. Deep Texture Analysis—Enhancing CT Radiomics Features for Prediction of Head and Neck Cancer Treatment Outcomes: A Machine Learning Approach. *Radiation* **2024**, *4*, 50–68. https://doi.org/10.3390/radiation4010005

Academic Editor: Sam Payabvash

Received: 15 December 2023
Revised: 9 February 2024
Accepted: 16 February 2024
Published: 28 February 2024

1. Introduction

Undergoing cancer treatment can be a taxing process for patients. In addition to the physical and emotional toll the cancer presents, patients disrupt their day-to-day routine by attending treatment (chemotherapy/radiation) and managing potential unwanted physical side-effects of treatment. Improvements in cancer treatment outcomes have come in large part due to a patient-centric approach, based on a plethora of factors (tumour size, location, stage, patient age, and other underlying conditions, to name a few). With an emphasis on personalized care and the growing popularity of machine learning (ML) applications, there has been a push to incorporate biomarkers (genetic, clinical, and imaging) to train and create models capable of predicting various biological endpoints. This is ultimately in order to permit the customization of care based on prognostic factors. Typically, biomedical imaging allows physicians to gain qualitative insight regarding patient's conditions. Whereas qualitative analysis of images is useful, it is also dependent on individual physicians and their interpretation of the images. In 1973, Haralick et al. pioneered the field of radiomics, which involves "mining" biomedical images for quantitative insights (textural features), based on the assumption that textural information may be represented by the overall or "average" spatial relationship of the pixels within the images [1]. In order to study these spatial relationships, Haralick et al. proposed the concept of a gray level co-occurrence matrix (GLCM), which is a matrix based on relationships between neighbouring pixels in an image [1]. To quantify texture features like contrast, homogeneity, and entropy, calculations are defined for the GLCM and more recently other similar matrices (gray level run length matrix (GLRLM) [2], gray level size zone matrix (GLSZM) [3], and gray level dependence matrix (GLDM) [4]). Textural analysis can be based on an entire image or on a specific region of interest (ROI) to create a set of imaging biomarkers. When these features are labelled retrospectively, with known biological endpoints, they can be used to train ML classifiers to create predictive models.

In this work, the possibility of predicting binary head and neck (H&N) cancer treatment outcomes from treatment-planning computed tomography (CT) scans was investigated. H&N cancers are a broad category of epithelial malignancies originating in the oral cavity, pharynx, larynx, paranasal sinuses, nasal cavity, and salivary glands [5]. According to the World Health Organization's International Agency for Research on Cancer, in 2020, there were an estimated 933,000 new cases of H&N cancers and some 460,000 persons who died as a result of H&N cancer complications, globally [6]. Approximately 90% are squamous cell carcinomas (SCCs) [7], with risk factors including tobacco [8] and alcohol consumption [9], p53 [10] and p16 gene mutations [11], and the presence of human papillomavirus (HPV) genomic DNA [12]. Although distant metastasis is rare at the time of diagnosis (10%), the majority of patients experience lymph node (LN)-related symptoms associated with regional spread of cancerous cells [5].

Treatment approaches include a combination of surgery, radiotherapy (XRT), and systemic therapy, and are individualized factoring in the patient's overall health and tumour stage. For up-front XRT, standard treatment objectives include 70 Gy in 33–35 fractions to high dose target volume for gross disease and 63/56 Gy in 33–35 fractions to intermediate and low dose (risk) target volumes, respectively [13]. Globally, 5-year mortality rates for H&N cancers are around 50% but vary based on factors like tumour stage and location (~90% for lip cancers, but <40% for cancer of hypopharynx), as well as geographic and socioeconomic considerations regarding access to healthcare [14,15]. Despite advances in personalized patient care, including newer treatment-planning software and innovations like intensity-modulated radiation therapy (IMRT) [16] and volumetric modulated arc therapy (VMAT) [17], there are always some patients who do not exhibit the desired response to treatment.

Studying tumour compositions and microenvironments are of particular interest within cancer research, with the understanding that tumour heterogeneity plays a very important role in treatment outcomes, disease progression, metastasis, and/or recurrence [18–21]. Understanding that genomic heterogeneity could translate to heterogeneous

tumour metabolism and eventually anatomy, radiomics analysis presents a hypothetically feasible quantitative signature profiling method. Furthermore, current profiling of cancerous tumours involves the acquisition of a biopsy sample, which, while very useful, has two major limitations: (i) acquiring a biopsy sample is an invasive procedure, and (ii) a subsample of the cancerous tissue, which may or may not be representative of the whole tumour, particularly with more heterogeneous tumours. Radiomics analysis of cancerous tumours presents a few noteworthy advantages; mainly, (i) it is non-invasive, (ii) it provides analysis of the whole tumour, and because of the previously mentioned non-invasiveness, (iii) it allows for longitudinal analysis and monitoring of changes. Finally, (iv) radiomics allows for maximizing the utility of CT scans which are routinely acquired as part of treatment planning and dosimetry.

In recent years, ML applications have gained popularity in several fields, including but not limited to finance, e-commerce, security authentication, autonomous driving, and medicine [22]. Medical applications include efforts to discriminate between metastatic and non-metastatic disease [23], as well as the prediction of cancer treatment response [24–26] and likelihood of recurrence [27], just to name a few. Mining phenotypic information from images with radiomics analysis in conjunction with analytical ML algorithms presents a promising field of study to address countless clinical outcomes [28–30].

In this study, phenotypic radiomics signals from XRT-targeted LN segmentations of H&N cancer patients were investigated, and in tandem, with retrospective treatment outcomes, used to train predictive ML models. A predictive model that could accurately and reliably predict treatment outcomes from pre-treatment phenotypic imaging features, would greatly improve standard cancer treatment protocols. For example, if a patient is predicted to respond well to treatment, they could be given reassurances about the predicted benefits and encouraged to overcome fears they may have. Alternatively, such models would also serve to benefit patients predicted to not achieve desired outcomes by allowing for treatment interventions such as changes in radiation dose or fractionation (e.g., dose escalation) or perhaps the avoidance of unnecessary and ineffective treatment and thereby sparing the patient from associated unwanted side effects.

Moreover, in this study, a novel methodology we named deep texture analysis (DTA) was investigated. DTA is an iterative process developed from the hypothesis that the examination of the spatial distribution of insightful features within an ROI can enhance the phenotypic insights for training predictive ML models. Visually summarized in Figure 1, DTA involves (i) identification of promising features, (ii) analyzing the spatial distribution of said features by creating texture feature maps, and subsequently (iii) mining radiomics features from deeper layers to (iv) recreate predictive models trained with newer feature sets that should theoretically demonstrate better with a superior balanced accuracy compared to models in previous layers due to the retention of top features at each layer. In the past, for the same patient cohort, quantitative ultrasound (QUS) features of index LNs were found to be useful for training predictive models, and "deeper" layer features (referred to as "texture-of-texture" features) enhanced predictive classification performance [24]. In this work, features were evaluated from treatment-planning CT (as opposed to QUS) and up to three layers deep (as opposed to two layers in the QUS study) [24].

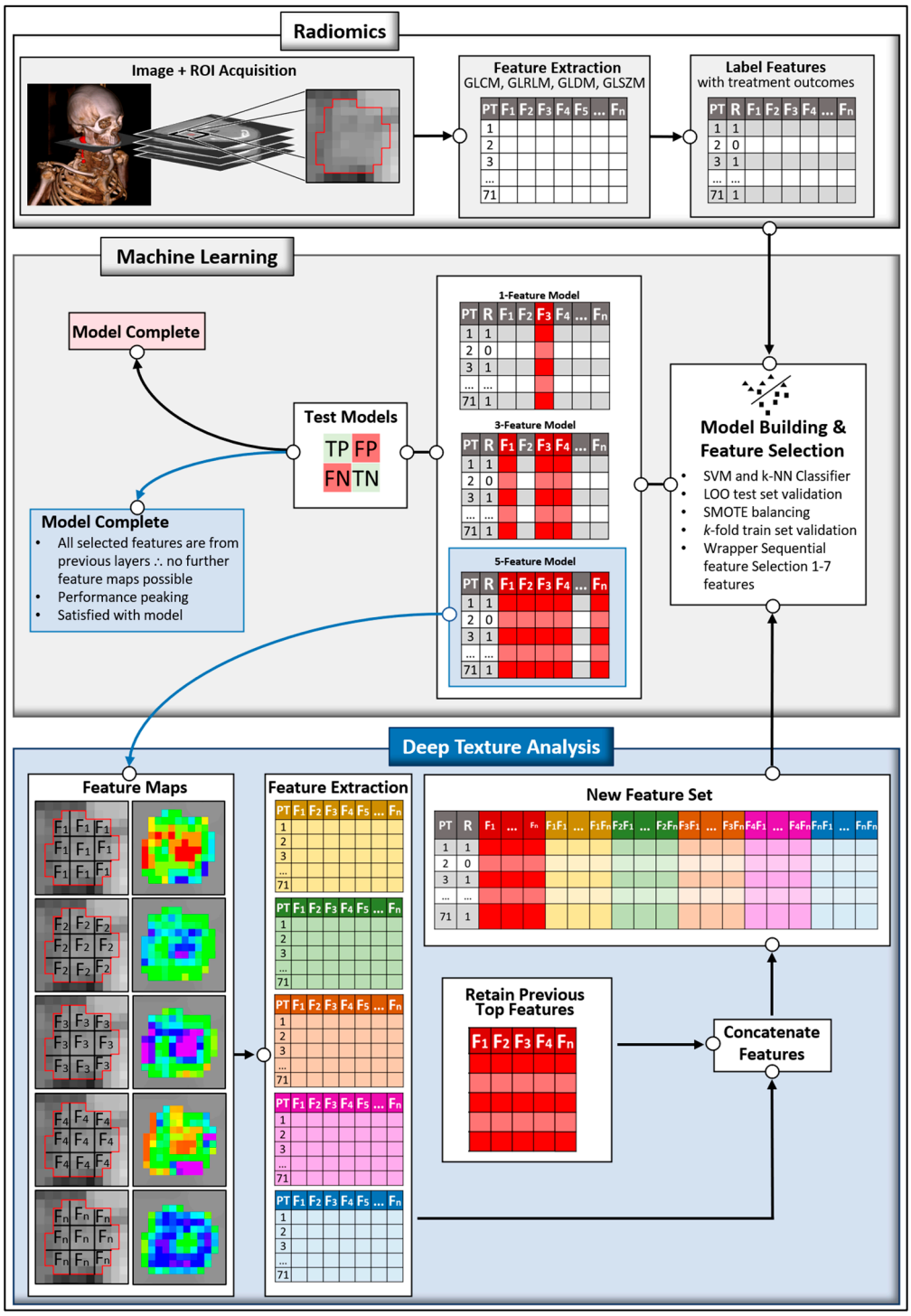

Figure 1. Visual summary of deep texture analysis. Note that models were made with 1-7 features but in this figure only the 1-, 3-, and 5-feature models are shown.

2. Materials and Methods

This study was conducted at Sunnybrook Health Sciences Centre and approved by the Sunnybrook Research Institute Ethics Board (SUN-3047). Data were included for $n = 71$ patients. Patients included in the study had a biopsy-confirmed diagnosis of H&N cancer, were to be treated with radiotherapy for gross disease, and had pathologically enlarged and measurable LN involvement (≥15 mm in "short axis" when assessed by CT scan). Nodal size is normally reported as two dimensions in the plane in which the image is obtained (for a CT scan, this is almost always in the axial plane). The smaller of these two

dimensions is called the "short axis". Patients were labelled as either complete or partial responders (CRs or PRs) based on clinical follow-up using contrast-enhanced MR imaging (based on Response Evaluation Criteria in Solid Tumours (RECIST) guidelines) conducted in the first 3 months after completion of treatment [31]. Through visual inspection, with disappearance of the primary disease and reduction of the index LN to <10 mm, patients were categorized as CRs. The remaining patients demonstrated at least a 30% reduction in the sum of diameters of tumours compared to baseline measurements, thus categorized as PRs. Criterion for stable and progressive disease are outlined in the protocol as well, but these patients were not identified in this study. Standard treatment follow-up protocols included additional follow-ups every 3–6 months for the first two years and every 6–12 months thereafter; however, the goal of this study was to predict treatment outcomes within the first three months of finishing treatment.

Gross tumour volume (GTV) segmentations were expanded by 5 mm for the high-dose clinical target volume on the primary and nodal volume. Furthermore, a 1 cm margin was added to the GTV to create the clinical tumour volume (CTV56). XRT administration was carried out using IMRT or VMAT techniques available at Odette Cancer Centre, Sunnybrook Health Sciences Centre, in Toronto, Ontario, Canada.

Treatment plans—including treatment-planning CT scans, segmentations, and transformations—were gathered from an institutional database. DICOM files were opened with open-source 3D Slicer (slicer.org), and treatment plans were registered to the associated CT scan using the transformation matrix [32]. Once the position was confirmed, LN segmentations were isolated and saved as a .nrrd file. If the treatment-plan segmentations delineated multiple LN segmentations, they were added together to create a single LN segmentation file. CT scans were saved as .nii files. An example can be seen in Figure 2.

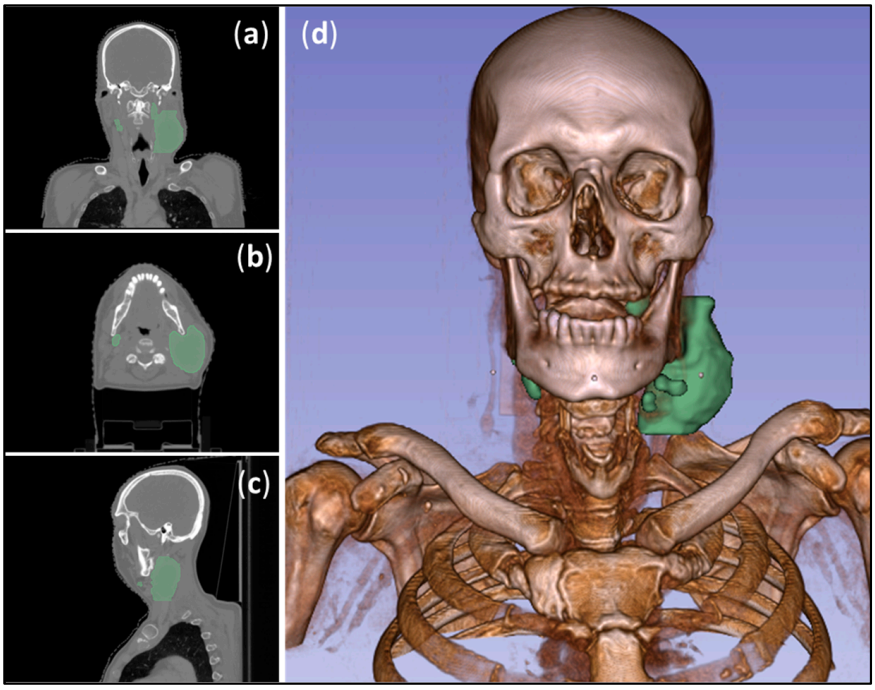

Figure 2. (**a**) Coronal, (**b**) axial, and (**c**) sagittal view of treatment-planning CT scan with lymph node segmentation highlighted in green. (**d**) Three-dimensional reconstruction of the scan and segmentation. Images created with 3D Slicer.

Next, 24 GLCM, 16 GLRLM, 16 GLDM, and 14 GLSZM 2-dimensional texture features were determined from the axial slices for each patient and associated LN ROI, using Pyradiomics (v3.0.1), an open-source Python (v3.7.10) (Python Software Foundation, Version 3.7.10, Delaware, USA) package [33]. Finally, patient features were labelled with the bi-

nary treatment response before moving on to ML analysis. This was called the S_1 (first stage) dataset.

Models were built using two well-established classifiers: *k*-nearest neighbour (*k*-NN) and support vector machine (SVM) classifiers. To create the models, an iterative leave-one-out test validation method was implemented whereby each sample was left out while the remaining samples were used to train and validate the models, before finally testing on the left-out sample. After leaving out the test sample, to account for the imbalance in data (25 CR/46 PR) and to avoid "anomaly-type" classification problems, the synthetic minority oversampling technique (SMOTE) was applied to the training set [34]. Using an iterative 5 *k*-fold split, the training set was further divided into 80% training and 20% validation sets to train and tune models. Because the number of acquired radiomics features (n = 70) $\approx$ number of patients (n = 71), to avoid the curse of dimensionality, which increases susceptibility to overfitting and reduces model generalizability, feature selection was carried out [35]. Feature extraction is another method to reduce dimensionality, for example through Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA); however, these processes involve transforming the original features to create a new (smaller) set of features [35]. Feature selection, however, is a method of reducing dimensionality through the identification of the most important and informative features. Identification of the most important radiomics features can hypothetically allow for anatomical and physiological interpretations in an effort to better understand the disease and its effects on treatment outcomes. Feature selection was carried out by an iterative sequential forward selection (SFS) method in a wrapper framework based on balanced accuracy for each of the training folds, and the most frequently selected features were used to train models and for testing on the left-out sample [35]. Model performance was evaluated based on sensitivity (%S_n), specificity (%S_p), accuracy (%Acc), precision (%Prec), balanced accuracy (%Bal Acc) and the area under the curve (AUC) of the receiver operating characteristic (ROC) for single-variable and multi-variable models including up top seven features.

$$\%S_n = \frac{TP}{TP + FN} \tag{1}$$

$$\%S_p = \frac{TN}{TN + FP} \tag{2}$$

$$\%Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\%Per = \frac{TP}{TP + FP} \tag{4}$$

$$\%Bal\ Acc = \frac{Sensitivity + Specificity}{2} \tag{5}$$

where *TP*, *TN*, *FP*, and *FN* indicate true positive (true response), true negative (true non-response), false positive, and false negative, respectively.

Within the field of deep learning, the attention mechanism stands out as a cornerstone of the transformer network [36]. This mechanism amplifies the impact of crucial information (characteristics), thereby enhancing differentiation between labels. Drawing inspiration from the attention mechanism, in this work, novel features were derived from a set of top-*k* (k = 5) selected features, which encapsulated the most discriminative information, in a method we called deep texture analysis.

To investigate the spatial distribution of the important features, the features identified in the 5-feature multivariable model were used to create texture feature maps by calculating the value of the identified features for sub-ROI (3 $\times$ 3 pixels) windows and assigning said value to the central pixel. Pixels outside of the LN ROI were 0-padded. For each of the five new sets of texture feature map images, once again, GLCM, GLRLM, GLDM, and GLSZM features were determined. These new features were concatenated with the originally selected 5 features to create a new, S_2, set of features for training ML classifiers.

The decision to select 5 features to study on a deeper layer was made arbitrarily. One could hypothetically create texture feature maps for every feature; however, this is not practical, as it is very computationally costly.

This process was repeated to create the S_3 feature sets, and model performances were evaluated based on previously mentioned metrics. To evaluate feature sets independent of how many features models were built with, at each layer, performance metrics were averaged for models with 1–7 features. To evaluate whether the inclusion of deeper layer features enhanced the overall quality of feature sets for training ML classifiers, a one-tailed t-test with a significance level (α) set at $p = 0.05$ was carried out on the average performance metrics, assessing the null hypothesis that there was no difference in average performance. It should be noted that after training the models on the S_2 dataset (the 5-top 1LTFs + 2LTFs), within the newly selected features, only the 2LTFs could be made into feature maps for 3LTF determination. If, hypothetically, the 5-feature multivariable model trained with the S_2 dataset identifies the most important features as all five of the 1LTFs, this would mean that there is no deeper layer to explore and that there was no added benefit in the inclusion of the 2LTFs. This would be the endpoint of the DTA method.

Focusing solely on these top-5 selected features for the extraction of profound textural features resembles an attention-driven approach that enables superior discrimination. At each layer, only the most informative features from the entire pool of determined features are retained, effectively managing algorithmic complexity by reducing dimensionality. In the realm of deep learning and convolutional neural networks (CNNs), a pooling layer is employed to reduce the dimensions of feature maps, with these selected features aptly representing the entirety of features [37]. Similarly, the utilization of the top-*k* selected features is deemed essential for extracting deeper features, as they possess the capability to encapsulate the most important information.

3. Results

3.1. Patient Characteristics

At the time of diagnosis, enrolled patients had a mean age of 59 years (±10.2), with a majority (n = 66, 93%) being males. Although there is a considerable discrepancy in the ratio of males to females, it should be noted that diagnosis of H&N cancer is far more common in men, as evidenced by a 25-year analysis of cancer prevalence in Canada, which revealed that out of nearly 48,000 total H&N cancer patients, 70% (~35,000) were males [38]. Smoking, drinking, and HPV status were also noted when available. The majority of patients (n = 61, 86%) were treated with chemotherapy (cisplatin, cetuximab, carboplatin, or combination carboplatin + eptoposide), and the remaining patients (n = 10) were treated with radiation alone. Table 1 summarizes the patient, disease, and treatment characteristics for all subjects. Supplementary Table S1 shows an anonymized breakdown of tumour and treatment characteristics for each patient.

Table 1. Summary of patient cohort involved in the study.

Patient Characteristics	***n* (%)**
Patient and Clinical Characteristics *n* = 71 (All Subjects)	
Age (years)	
- Median	- 60.5
- Mean	- 58.9 ± 10.2
Gender	
- Male	- 66 (92.9)
- Female	- 5 (7.0)
Smoking Status:	
- Smoker	- 47 (66.2)
- Non-smoker	- 23 (32.4)
- Unknown	- 1 (1.4)

Table 1. *Cont.*

Patient Characteristics	***n* (%)**
Drinking Status:	
- Drinker	- 50 (70.4)
- Non-drinker	- 15 (21.1)
- Unknown	- 6 (8.4)
Tumour Status	
Primary Tumour (T):	
- T1	- 4 (5.6)
- T2	- 23 (32.3)
- T3	- 7 (9.8)
- T4	- 14 (19.7)
- Unknown	- 23 (32.4)
Histological Type:	
- Squamous cell carcinoma	- 66 (92.9)
- Other	- 5 (7.1)
HPV status:	
- p16 positive	- 40 (56.3)
- p16 negative	- 2 (2.8)
- Unknown	- 29 (40.8)
Treatment Characteristics	
Chemotherapy	- 61 (85.9)
- Cisplatin	- 54 (76.0)
- Carboplatin	- 5 (7.0)
- Carboplatin + etoposide	- 1 (1.4)
- Cetuximab	- 1 (1.4)
No Chemotherapy	- 10 (14.1)
Post Treatment (3-month assessment from MRI)	
Complete Responder—locoregional control (CR)	- 25 (35.2)
Partial Responder—locoregional failure (PR)	- 46 (64.8)

3.2. Models Trained with 1LTFs

Using 1LTFs determined from treatment-planning CT images, both *k*-NN and SVM classifier models trained with S_1 feature set demonstrated the ability to predict treatment outcomes of index LNs, with varying effectiveness, summarized in Tables 2 and 3.

Table 2. Results from 1–7-feature SVM models, trained using S_1 feature set.

Number of Features Selected	**%S_n**	**%S_p**	**%Acc**	**%Prec**	**%Bal Acc**	**AUC**
1	78.3	60.0	71.8	78.3	69.1	0.749
2	71.7	60.0	67.6	76.7	65.9	0.698
3	78.3	60.0	71.8	78.3	69.1	0.723
4	78.3	60.0	71.8	78.3	69.1	0.681
5 *	76.1	56.0	69.0	76.1	66.0	0.645
6	78.3	64.0	73.2	80.0	71.1	0.651
7	67.4	60.0	64.8	75.6	63.7	0.630

* Feature maps were made for selected features, from which 2LTFs were determined.

Table 3. Results from 1–7-feature *k*-NN models, trained using S_1 feature set.

Number of Features Selected	$\%S_n$	$\%S_p$	%Acc	%Prec	%Bal Acc	AUC
1	78.3	60.0	71.8	78.3	69.1	0.622
2	73.9	56.0	67.6	75.6	65.0	0.624
3	69.6	68.0	69.0	80.0	68.8	0.652
4	67.4	60.0	64.8	75.6	63.7	0.656
5 *	67.4	68.0	67.6	79.5	67.7	0.670
6	69.5	68.0	69.0	80.0	68.8	0.664
7	69.5	64.0	67.6	78.0	66.8	0.651

* Feature maps were made for selected for which 2LTFs were determined.

For the SVM classifier models, balanced accuracy scores ranged from 63.7 to 71.1%, with the best model, the six-feature multivariable model, demonstrating the best balanced accuracy of %Bal Acc = 71%. The six selected features for that model were "GLSZM Gray Level Variance", "GLDM Small Dependence Emphasis", "GLSZM Zone Percentage", "GLCM Informational Measure of Correlation 1 (IMC1)", "GLCM Inverse Difference Normalized (IDN)", and "GLCM Inverse Difference Moment (IDM)", the first five of which were also selected in the five-feature multivariable model and computed as feature maps for 2LTF determination. Details and formulas to determine the features can be found in the Pyradiomics documentation and are discussed further in the next section [33].

For *k*-NN classifier models, balanced accuracy scores ranged from 63.7 to 69.1%. The single-feature model had the highest balanced accuracy (% Bal Acc = 69.1%). However, this model had a poor balance between sensitivity (highest of the *k*-NN models, $\%S_n$ = 78.3%) and specificity (second lowest of the *k*-NN models, $\%S_p$ = 60.0%), and had the lowest AUC (0.622) of the seven S_1-trained *k*-NN models. The model with the highest AUC (0.670) was the five-feature multivariable model with a balanced accuracy of % Bal Acc = 68.8%. The associated features were "GLDM Gray Level Non-Uniformity", "GLSZM High Gray Level Zone Emphasis", "GLRLM Gray Level Non-Uniformity", "GLRLM Run Length Non-Uniformity Normalized", and "GLCM Joint Entropy". Feature maps of the aforementioned features were made for 2LTF feature extraction and modeling.

3.3. Models Incorporating 2LTFs

For each of the two classifiers, a set of 1LTF maps were created using Pyradiomics, based on selected features from the five-feature multivariable model. A new feature set, $S_{2,SVM}$ and $S_{2,k\text{-NN}}$, included the selected five 1LTFs concatenated with 350 newly determined 2LTFs, for a total of $S_{2,SVM}/S_{2,k\text{-NN}}$ = 355 features.

When considering models trained using S_1 feature sets with SVM classifier (Table 2), and comparing to model performances after incorporating 2LTFs, as seen in Table 4, findings suggest that regardless of how many selected features models are built upon, the $S_{2,SVM}$ dataset enhanced model sensitivities with average $\%S_n$ = 75 improving significantly to $\%S_n$ = 88 ($p < 0.05$), at the cost of significant decreases in specificity from an average $\%S_p$ = 60% to $\%S_p$ = 50% ($p < 0.05$). Average accuracy increased significantly from %Acc = 70% to %Acc = 75% ($p < 0.05$). Balanced accuracy increased from %Bal Acc = 67.7 to %Bal Acc = 69.3, but not significantly ($p > 0.05$).

Looking at *k*-NN models trained using an $S_{2,k\text{-NN}}$ dataset (Table 5), once again, across all seven models, the incorporation of 2LTFs improved average sensitivity significantly from $\%S_n$ = 71% to $\%S_n$ = 77% ($p < 0.05$). Unlike the SVM models, 2LTFs had no impact on model specificity as the average across all seven models remained unchanged ($\%S_p$ = 63.4%). Average accuracy across all seven models improved significantly from %Acc = 68% for S_1 features to %Acc = 72% after the introduction of 2LTFs ($S_{2,k\text{-NN}}$) ($p < 0.05$). Precision, balanced accuracy, and AUC also improved significantly ($p < 0.05$)

from %Prec = 78.1% to %Prec = 79.5%, %Bal Acc = 67% to %Bal Acc = 70%, and AUC = 0.648 to AUC = 0.701, respectively.

Table 4. Results from 1–7-feature SVM models, trained using $S_{2,SVM}$ feature set.

Number of Features Selected	%S_n	%S_p	%Acc	%Prec	%Bal Acc	AUC
1	91.3	44.0	74.6	75.0	67.7	0.616
2	91.3	48.0	76.1	76.4	69.7	0.615
3	87.0	52.0	74.6	76.9	69.5	0.630
4	87.0	52.0	74.6	76.9	69.5	0.645
5 *	87.0	52.0	74.6	76.9	69.5	0.640
6	87.0	52.0	74.6	76.9	69.5	0.650
7	87.0	52.0	74.6	76.9	69.5	0.640

* Feature maps were made for selected 2LTFs for which 3LTFs were determined.

Table 5. Results from 1–7-feature *k*-NN models, trained using $S_{2,k\text{-NN}}$ feature set.

Number of Features Selected	%S_n	%S_p	%Acc	%Prec	%Bal Acc	AUC
1	82.6	60.0	74.6	79.2	71.3	0.763
2	76.1	64.0	71.8	79.5	70.0	0.690
3	76.1	64.0	71.8	79.5	70.0	0.701
4	76.1	64.0	71.8	79.5	70.0	0.690
5 *	76.1	64.0	71.8	79.5	70.0	0.700
6	78.3	64.0	73.2	80.0	71.1	0.682
7	76.1	64.0	71.8	79.5	70.0	0.682

* Feature maps were made for selected 2LTFs for which 3LTFs were determined.

3.4. Models Incorporating 3LTFs

Selected 2LTFs for the five-feature multivariable models were used to determine 3LTFs. For the SVM model, all five selected features were 2LTFs, namely "GLSZM Small Area Low Gray Level Emphasis" from the "GLCM IMC1" 1LTF map, "GLDM Small Dependence Low Gray Level Emphasis" from the "GLCM IDN" 1LTF map, "GLCM Difference Variance" and "GLCM IDN" from the "GLSZM Zone Percentage" 1LTF map, and finally, "GLCM Correlation" from the "GLSZM Gray Level Variance" 1LTF map. Seventy 3LTFs were determined for each of the five mentioned 2LTFs selected.

Interestingly, for the *k*-NN model, the selected five features were a combination of three 1LTFs ("GLDM, Run Length Non-Uniformity Normalized", "GLCM Joint Entropy", and "GLSZM High Gray Level Zone Emphasis") and two 2LTFs ("GLSZM Size Zone Non-Uniformity" from "GLDM Gray Level Non-Uniformity" 1LTF parametric map and "GLCM Dependence Non-Uniformity" from "GLSZM high Gray Level Zone Emphasis" 1LTF parametric map). Seventy 3LTFs were determined for each of the two selected 2LTFs.

Newly created 3LTFs were concatenated with the five features selected for both classifiers' five-feature multivariable model (trained using $S_{2,SVM}$ and $S_{2,k\text{-NN}}$ datasets), and models were built based on these new features. Therefore, for the SVM classifier, the final dataset, $S_{3,SVM}$, included (5 × 2LTFs) + (70 3LTFs × five 2LTF maps) for a total of $S_{3,SVM}$ = 355 features. For the *k*-NN classifier, the final dataset, $S_{3,k\text{-NN}}$, included (3 × 1LTFs) + (2 × 2LTFs) + (70 3LTFs × two 2LTF maps) for a total of $S_{3,k\text{-NN}}$ = 145 features.

The performance of the models trained using these features can be found in Tables 6 and 7.

Table 6. Results from 1–7-feature SVM models, trained using $S_{3,SVM}$ feature set.

Number of Features Selected	%S_n	%S_p	%Acc	%Prec	%Bal Acc	AUC
1	71.7	64.0	69.0	78.6	67.9	0.653
2	76.1	64.0	71.8	79.5	70.0	0.730
3	76.1	60.0	70.4	77.8	68.0	0.718
4	73.9	60.0	69.0	77.3	67.0	0.715
5	73.9	60.0	69.0	77.3	67.0	0.723
6	76.1	64.0	71.8	79.5	70.0	0.710
7	76.1	64.0	71.8	79.5	70.0	0.717

Table 7. Results from 1–7-feature *k*-NN models, trained using $S_{3,k\text{-NN}}$ feature set.

Number of Features Selected	%S_n	%S_p	%Acc	%Prec	%Bal Acc	AUC
1	82.6	64.0	76.1	80.9	73.3	0.743
2	76.1	64.0	71.8	79.5	70.0	0.678
3	73.9	64.0	70.4	79.1	69.0	0.688
4	76.1	64.0	71.8	79.5	70.0	0.706
5	76.1	64.0	71.8	79.5	70.0	0.694
6	73.9	68.0	71.8	81.0	71.0	0.695
7	73.9	68.0	71.8	81.0	71.0	0.700

To determine whether DTA improved the quality of feature sets when training classifiers, we compared models trained with the S_1 feature set to models trained with S_n feature sets by comparing the average performances across all seven models for each feature set. This was motivated by the desire to evaluate performance and compare feature sets in a manner independent of the number of features given models are built with. Tables 8 and 9 show the average performance of SVM and *k*-NN models with corresponding *p*-values from a one-tailed *t*-test to evaluate significant change between models trained using S_1 versus S_2 features, and S_1 versus S_3 features. Comparing SVM models trained using the S_1 feature sets to SVM models trained using S_2 feature sets, sensitivity and accuracy improved significantly. Specificity and AUC decreased significantly. Precision decreased, albeit insignificantly. Balanced accuracy also had an insignificant increase. Compared with S_3-trained models, sensitivity decreased insignificantly. Specificity improved significantly, and accuracy, precision, balanced accuracy, and AUC increased, albeit not significantly.

Evaluating the inclusion of deeper layer features in *k*-NN models (Table 9), sensitivity, accuracy, precision, balanced accuracy, and AUC all improved significantly when comparing S_1-trained versus S_3-trained models. Specificity remained unchanged. Similar trends were seen when comparing S_1-trained versus $S_{3,k\text{-NN}}$-trained models, since sensitivity, accuracy, precision, balanced accuracy, and AUC improved significantly, and specificity also improved, albeit not significantly.

For *k*-NN classifier multi-variable models trained using S_1, $S_{2,k\text{-NN}}$, and $S_{3,k\text{-NN}}$ feature sets, the best model (based on %Bal Acc and AUC) was the seven-feature multivariable model trained using the $S_{3,k\text{-NN}}$ feature set (%S_n = 74%, %S_p = 68%, %Acc = 72%, %Prec = 80%, %Bal Acc = 71%, and AUC = 0.700). The seven selected features included two 1LTFs ("GLRLM Run Length Non Uniformity Normalized" and "GLCM Joint Entropy"), two 2LTFs ("GLSZM Size Zone Non Uniformity" from the 1LTF "GLDM Gray Level Non-Uniformity" feature map and "GLDM Dependence Non-Uniformity" from the 1LTF "GLSZM High Gray Level Zone Emphasis" feature map), and three 3LTFs ("GLCM

Autocorrelation" from the 2LTF "GLDM Gray Level Non-Uniformity_GLSZM Size Zone Non-Uniformity" feature map, "GLCM Cluster Shade" from the 2LTF "GLSZM High Gray Level Zone Emphasis_GLDM Dependence Non-Uniformity" feature map, and "GLCM Cluster Tendency" from the 2LTF "GLDM Gray Level Non-Uniformity_GLSZM Size Zone Non-Uniformity" feature map).

Table 8. Comparing SVM models trained using S_1 versus S_2 as well as S_1 versus S_3 feature sets. In bold are metrics that improved significantly ($p < 0.05$).

Average Performance (%)	S_1 Models	$S_{2,SVM}$ Models	One-Tailed *p*-Value	S_1 Models	$S_{3,SVM}$ Models	One-Tailed *p*-Value
Sensitivity	**75.5 ± 4.0**	**88.2 ± 1.9**	**0.0002**	75.5 ± 4.0	74.8 ± 1.6	0.3800
Specificity	60.0 ± 2.1	50.3 ± 2.9	0.0003	**60.0 ± 2.1**	**62.3 ± 1.9**	**0.0150**
Accuracy	**70.0 ± 2.8**	**74.8 ± 0.5**	**0.0038**	70.0 ± 2.8	70.4 ± 1.3	0.3941
Precision	77.6 ± 1.4	76.6 ± 0.6	0.0864	77.6 ± 1.4	78.5 ± 0.9	0.1264
Balanced Accuracy	67.7 ± 2.4	69.3 ± 0.6	0.0981	67.7 ± 2.4	68.6 ± 1.3	0.2549
AUC	68.2 ± 4.0	63.4 ± 1.3	0.0285	68.2 ± 4.0	70.9 ± 2.4	0.1484

Table 9. Comparing *k*-NN models trained using S_1 versus S_2 as well as S_1 versus S_3 feature sets. In bold are metrics that improved significantly ($p < 0.05$).

Average Performance (%)	S_1 Models	$S_{2,k\text{-NN}}$ Models	One-Tailed *p*-Value	S_1 Models	$S_{3,k\text{-NN}}$ Models	One-Tailed *p*-Value
Sensitivity	**70.8 ± 3.6**	**77.3 ± 2.3**	**0.0002**	**70.8 ± 3.6**	**76.1 ± 2.8**	**0.0006**
Specificity	63.4 ± 4.5	63.4 ± 1.4	0.5000	63.4 ± 4.5	65.1 ± 2.3	0.1779
Accuracy	**68.2 ± 1.9**	**72.4 ± 1.0**	**0.0001**	**68.2 ± 1.9**	**72.2 ± 1.7**	**0.0004**
Precision	**78.1 ± 1.8**	**79.5 ± 0.2**	**0.0466**	**78.1 ± 1.8**	**80.0 ± 0.7**	**0.0186**
Balanced Accuracy	**67.1 ± 1.9**	**70.3 ± 2.3**	**0.0016**	**67.1 ± 1.9**	**70.6 ± 1.3**	**0.0020**
AUC	**64.8 ± 1.7**	**70.1 ± 2.6**	**0.0079**	**64.8 ± 1.7**	**70.1 ± 1.9**	**0.0026**

Investigating the spatial distribution of features, the proposed DTA method was explored for two more layers, as seen in Figure 3.

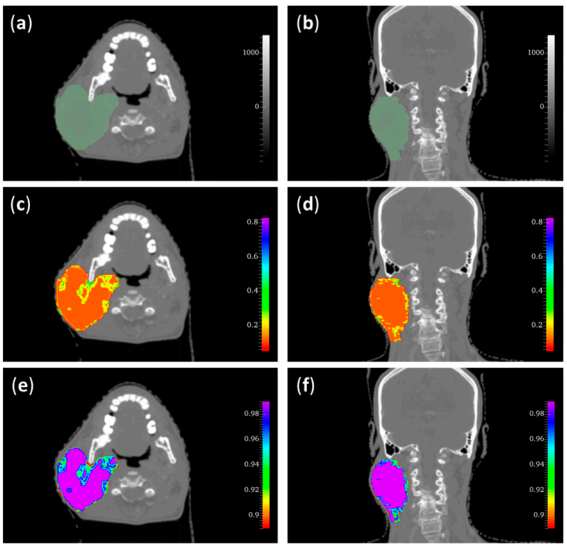

Figure 3. (**a**,**c**,**e**) Axial and (**b**,**d**,**f**) coronal treatment-planning CT and LN segmentation (**top**), 1LTF (GLSZM Zone Percentage) map (**middle**), and 2LTF (GLSZM Zone Entropy—GLCM IDN) map (**bottom**).

In summary, the results suggest that feature sets were enhanced by incorporating 2LTFs and 3LTFs for classifier training. The best performance for each classifier was the seven-feature multivariable model trained using S_3 feature sets. The seven-feature SVM classifier multivariable model had $\%S_n$ = 76%, $\%S_p$ = 64%, %Acc = 72%, %Prec = 80%, %Bal Acc = 70%, and AUC = 0.717, and the *k*-NN classifier model, $\%S_n$ = 74%, $\%S_p$ = 68%, %Acc = 72%, %Prec = 81, %Bal Acc = 71, and AUC = 0.700. Figures 4 and 5 represent the average performance of all seven models trained using S_1, S_2, and S_3 feature sets for SVM and *k*-NN classifiers, respectively. To our knowledge, this is the first time DTA methodology has been investigated for CT scans.

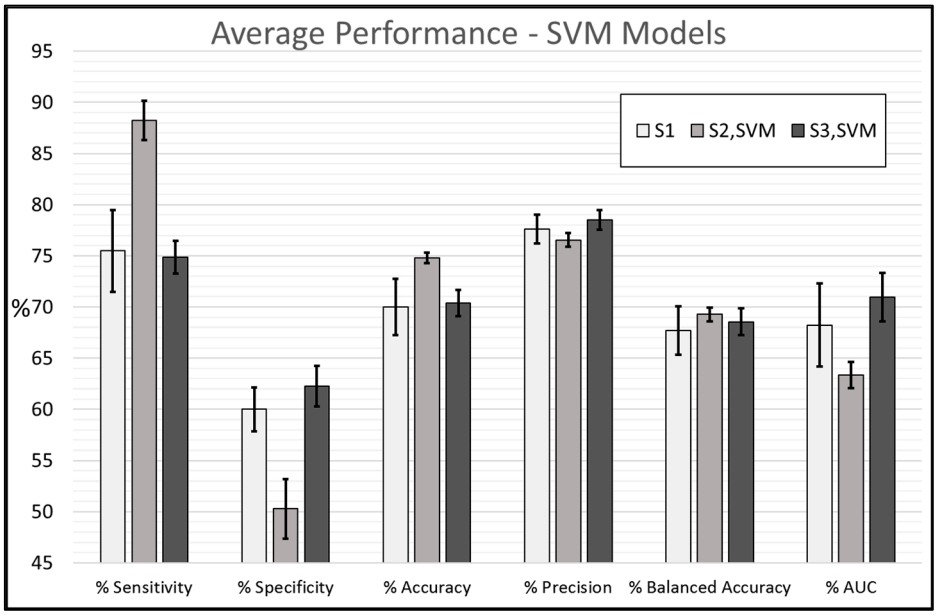

Figure 4. Average performance for SVM models with 1–7-feature models trained using S_1, $S_{2,SVM}$, $S_{3,SVM}$ feature sets.

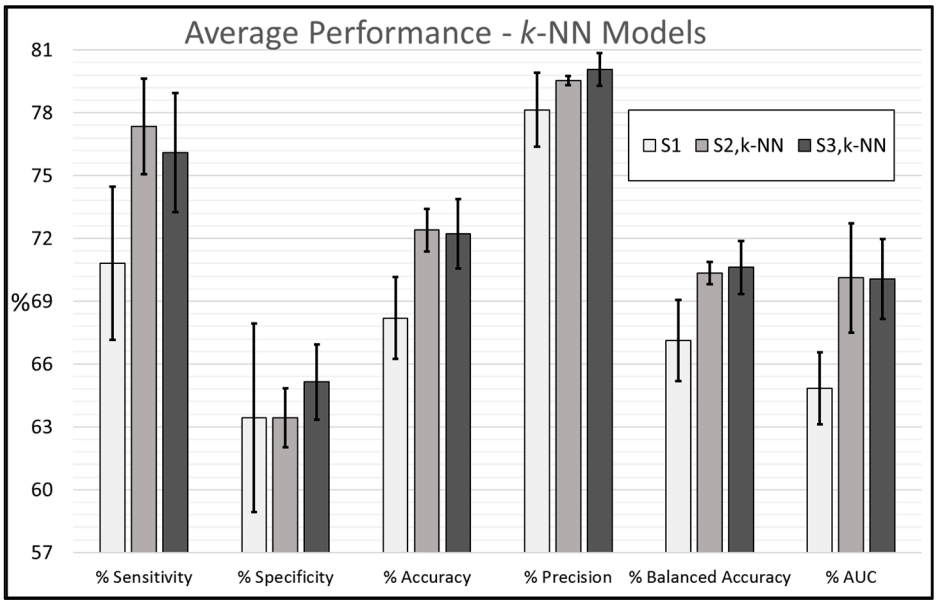

Figure 5. Average performance for *k*-NN models with 1–7-feature models trained using S_1, $S_{2,k\text{-NN}}$, $S_{3,\,k\text{-NN}}$ feature sets.

4. Discussion

In this study, texture features determined from treatment-planning CT scans of H&N cancer patients yielded phenotypic insights regarding treatment endpoints. Pre-treatment GLCM, GLDM, GLRLM, and GLSZM texture features (S_1) from treatment-targeted LNs proved useful in training SVM and *k*-NN ML classifiers for binary treatment outcome prediction. Feature selection was performed using SFS for models with 1–7 features. For SVM classifier models, the best balanced accuracy was found with a six-feature multivariable model with $\%S_n$ = 78.3%, $\%S_p$ = 64, %Acc = 73.2, %Prec = 80.0, %Bal Acc = 71.1, and AUC = 0.651. The six selected features were "GLSZM Gray Level Variance", "GLDM Small Dependence Emphasis", "GLSZM Zone Percentage", "GLCM Informational Measure of Correlation 1 (IMC1)", "GLCM Inverse Difference Normalized (IDN)", and "GLCM Inverse Difference Moment (IDM)", the first five of which were computed as feature maps for 2LTF determination. "GLSZM Gray Level Variance" is a Pyradiomics feature that measures the variance in gray level intensities for the 'zones' within the GLSZM [33]. A gray level 'zone' is defined as the number of connected voxels (or pixels) that share the same gray level intensity [33]. Therefore, identification of GLSZM Gray Level Variance supports the notion that ROI heterogeneity impacts treatment efficacy. The next feature was "GLDM Small Dependence Emphasis", which is a measure of the distribution of small dependencies, with a greater value indicative of smaller dependence and less homogeneous textures [33]. A "dependency" in regard to GLDM is the number of connected voxels (or pixels) within a specific distance and magnitude that are dependent on a central voxel [33]. The next feature was "GLSZM Zone Percentage", which measures the coarseness of the texture by taking the ratio of the number of zones and the number of voxels in the ROI. The remaining three features, "GLCM IMC1", "GLCM IDN" and "GLCM IDM", are various methods of quantifying texture heterogeneity.

Similarly, for the *k*-NN classifier models, the highest multivariable model balanced accuracy came from the six-feature model with $\%S_n$ = 69.5%, $\%S_p$ = 68%, %Acc = 69%, %Prec = 80%, %Bal Acc = 68.8%, and AUC = 0.664. The six selected features were "GLDM Gray Level Non-Uniformity", "GLSZM High Gray Level Zone Emphasis", "GLRLM Gray Level Non-Uniformity", "GLRLM Run Length Non-Uniformity Normalized", "GLCM Joint Entropy", and "GLSZM Gray Level Non-Uniformity", the first five of which were made into feature maps for 2LTF determination. "GLDM Gray Level Non-Uniformity" measures the similarity of gray-level intensity values in the image, with lower values correlating with greater similarity in intensity values [33]. Quantifying "similarity" in pixel intensities within the ROI can be thought of as analogous to measuring homogeneity. The next three features, "GLSZM High Gray Level Zone Emphasis", "GLRLM Gray Level Non-Uniformity", and "GLRLM Run Length Non-Uniformity", are all measures of heterogeneity within the ROI [33]. The final selected feature that was used to determine 2LTFs was "GLCM Joint Entropy", which measures the randomness or variability in the neighbourhood intensity values within the GLCM.

Predicting biological endpoints with radiomics features is a growing area of research. For example, Tang et al. reported contrast-enhanced CT radiomics features acquired pre-treatment to be useful in predicting recurrence within two years of locally advanced esophageal SCC with radiomics features alone, clinical features alone, and combined clinical (7 features) and radiomics features (10 features), with $\%S_n$ = 87%, 79%, and 89% respectively (n = 220) [27]. Another study by Huang et al. reports success in predicting metastasis and extranodal extension in H&N patients with preoperative CT-scan radiomics features, even when compared to experienced radiologists (n = 464) [39]. For predicting metastasis, Huang et al. found %Acc = 73.8% for radiologist performances and %Acc = 77.5% for model performance [39]. For predicting extranodal extension, radiologists performed with %Acc = 70.4%, whereas the model performed with %Acc = 80% [39].

Radiomics studies are not limited to features determined from CT images. For example, for the prediction of preoperative cavernous sinus invasion from pituitary adenomas, a condition of interest for determining optimal treatment planning, radiomics features

evaluated from contrast-enhanced T_1 MRI scans were used to train a linear support vector machine model with %Acc = 80.4%, $\%S_n$ = 80.0%, $\%S_p$ = 80.7%, and AUC = 0.826 [40]. In another study, MRI radiomics were investigated to differentiate between low-grade glioma and glioblastoma peritumoral regions [41], and yet another investigated prediction of response to neoadjuvant chemotherapy in patients with locally advanced breast cancer.

Previously, the DTA method was investigated for the first two layers for QUS texture features determined from LN quantitative ultrasound parametric maps, for the same cohort of patients [24]. DTA methodology and the inclusion of 2LTFs improved model performance in the QUS study as well (seven-feature SVM model $\%S_n$ = 81% improved to $\%S_n$ = 85%, $\%S_p$ = 76% improved to $\%S_p$ = 80%, %Acc = 79% improved to %Acc = 83%, %Prec = 86% improved to %Prec = 89%, and AUC increased from 0.82 to 0.85) [24]. Overall, models trained using QUS features outperformed the models in that study, suggesting that they reveal more phenotypic insight regarding treatment efficacy. This could be due to the fact that in the QUS study, features were determined from QUS parametric maps, ultrasound parameters known to be associated with tissue microstructures [42], whereas, in this study, features were determined from the CT image itself. Most importantly, however, both studies confirm that the inclusion of deeper layer texture features through the DTA methodology can improve model training.

However, it should be noted that radiomics studies do not always yield effective predictive capabilities, as was reported by Keek et al. who investigated radiomics features (GLCM, GLRLM, and GLSZM features in addition to first-order and shape features) from H&N SCC for prediction of overall survival, locoregional recurrence, and distant metastasis after concurrent chemo-radiotherapy (n = 444), using Cox proportional hazards regression and random survival forest models, and found that radiomics features from the peritumoral regions are not useful for the prediction of time to overall survival, locoregional recurrence, and/or distant metastasis, which the authors posit may be related to high variability between training and validation datasets [43]. Another study (with a large cohort (n = 726)) by Ger et al. found that prediction of overall survival did not improve after incorporating radiomics features, which was concluded when comparing a model trained using HPV status, tumour volume, and two radiomics features to a model using just tumour volume alone, and found that the AUC of the radiomics-included model was lower than the AUC of the model with tumour volume alone; however, the authors did comment on the potential advantages of using LN radiomics features instead of primary tumour features [44].

Furthermore, deeper layer texture features were determined only for feature maps made from top-five selected features from previously trained models. One could, hypothetically, create feature maps for every feature and subsequently acquire deeper layer textures for all available features before any model building; however, the proposed method of initial model building and feature selection makes sense not only intuitively, but also practically. For example, consider that in this work, the S_1 dataset included 70 (24 GLCM, 16 GLRLM, 16 GLDM, and 14 GLSZM) 1LTFs. If feature maps were made from all 70 features and 2LTFs were determined for all 70 feature maps, the subsequent S_2 dataset would include 4970 features (70 1LTFs + 4900 2LTFs). If the same process was repeated to incorporate 3LTFs, S_3 would include >300,000 features. However, through model building and feature selection at each layer, only important features are highlighted and "zoomed in" on or "focused on" through deep texture analysis. The calculation of textural features from the LN ROIs ranged in the order of minutes, whereas the computational time to create the texture feature maps ranged from a few hours to a few days depending on the complexity of the calculation, size of ROI, and quantization of pixel intensities. In this study, the number of features was greater than the number of samples, which means we had an underdetermined equation system. In this situation, the probability of overfitting is considerably high. To circumvent this challenge, we applied feature selection to reduce the dimension of data.

Although the results were promising, it should be noted that due to a small sample size, these models are not yet generalizable for clinical applications. Moreover, patients

in this study were recruited with the presence of bulky neck disease, which represents a subset of all H&N cancer patients. However, the utility of this work may be clinically useful since it is exactly such patients with bulky disease who typically respond poorly to treatment and can benefit from adaptive radiotherapy in the future with response-predictive input on the basis of imaging. Furthermore, predictive models could incorporate clinical features, such as smoking and drinking history, along with HPV status, in model training. In this work, the feasibility of radiomics features was evaluated, and in particular, the influence of 2LTFs and 3LTFs was investigated. Further, features could be determined from the primary tumour as well as the LNs; however, in this study, some patients had unknown primary tumours. Additionally, with the understanding that cancerous cells extend beyond the visible GTV margins, some radiomics studies also evaluate features from the tumour margins [45]. Despite the limitations, the results in this study were promising, suggesting that treatment response can be predicted from treatment-planning CT scans and associated LN segmentations. Additionally, it seems that DTA methodology enhanced the quality of feature sets, and these results were consistent with previous work on QUS features for the same patient cohort. In the future, texture features and, in particular, the proposed DTA methodology will be investigated for the same patient cohort, using features determined from diagnostic contrast-enhanced T_1 MRI scans. Finally, an investigation will be performed to compare models trained using features from each of the three modalities, as well as training models on a combination of QUS + CT + MRI features.

Lastly, it is worth bringing attention to the fact that as an alternative to determining radiomics features and model building, another approach could be to utilize deep learning methods to build models directly from the images. Huynh et al. investigated the effectiveness of predictive models using conventional radiomics features with deep learning models and found that CNNs trained on images achieved the highest performance and that adding radiomics and clinical features to these models could enhance the performance further [46]. When testing models with radiomics and clinical features, it was found that they were susceptible to overfitting and, in particular, poor cross-institutional generalizability perhaps due to small sample sizes and variability in data procuring [46]. However, although deep learning approaches yield attractive results, they can be considered as "black boxes" with limited transparency and interpretability. The potential for radiomics models is in the fact that the features are well defined, and identification of important radiomics features with respect to any biological endpoint allows for further study and physiological investigation, in an effort to better understand the nature of the condition in question.

5. Conclusions

This study was designed based on the hypothesis that the index LNs of H&N cancer patients contain radiological phenotypes that can be correlated to and prove insightful in predicting the treatment response of the primary tumour and nodal disease. With modern healthcare approaches centered on personalized medical care, reliable and generalizable predictive models would provide clinicians with yet another tool in the treatment-planning protocol. Predictive models would benefit both CR and PR patients in better understanding their potential treatment outcomes, which can assist them in their decision making. Finally, it was found that DTA methodology and deep layer texture features can enhance predictive model performance and are worth consideration in future radiomics studies going forward.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/radiation4010005/s1. Table S1: Anonymized clinical characteristics of interest as reported in patient notes from institutional database.

Author Contributions: Conceptualization, A.S. and G.J.C.; methodology, A.S.; software, L.S.; data curation, A.S.; validation, A.S. and L.S.; formal analysis, A.S.; investigation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S., A.M. and G.J.C.; supervision, A.P.-M. and G.J.C.; funding acquisition, G.J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC – RGPIN:2019-06846) as well as Terry Fox Research Institute (TFRI)/Lotte & John Hecht Foundation (project #1115) and the Canadian Institute of Health Research (CIHR) (project #162136) Program Project. The funding agencies had no role in the study design, study methodology, study results, or preparation of the manuscript.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Sunnybrook Health Sciences Centre (SUN-3047).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data can be made available upon request (contact Czarnota Lab at Sunnybrook Health Sciences Centre).

Acknowledgments: We would like to thank all the patients for their participation in the study. Our sincere gratitude to the physicians and other healthcare staff for their support in patient care. Thank you to Ian Poon, Andrew Bayley, and Irene Karam for their efforts in treating these patients and contouring patient scans. We would also like to thank Toronto Metropolitan University, University of Toronto, the Terry Fox Foundation, and the Lotte and John Hecht Foundation for their willingness to collaborate on this project.

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

References

1. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
2. Chu, A.; Sehgal, C.; Greenleaf, J. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognit. Lett.* **1990**, *11*, 415–419. [CrossRef]
3. Thibault, G.; Fertil, B.; Navarro, C.; Pereira, S.; Cau, P.; Levy, N.; Sequeira, J.; Mari, J.-L. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification. *Int. J. Pattern Recognit. Artif. Intell.* **2013**, *27*, 1357002. [CrossRef]
4. Sun, C.; Wee, W.G. Neighboring gray level dependence matrix for texture classification. *Comput. Vis. Graph. Image Process.* **1983**, *23*, 341–352. [CrossRef]
5. Argiris, A.; Karamouzis, M.V.; Raben, D.; Ferris, R.L. Head and neck cancer. *Lancet* **2008**, *371*, 1695–1709. [CrossRef] [PubMed]
6. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
7. Skarsgard, D.P.; Groome, P.A.; Mackillop, W.J.; Zhou, S.; Rothwell, D.; Dixon, P.F.; O'Sullivan, B.; Hall, S.F.; Holowaty, E.J. Cancers of the upper aerodigestive tract in Ontario, Canada, and the United States. *Cancer* **2000**, *88*, 1728–1738. [CrossRef]
8. Vineis, P.; Alavanja, M.; Buffler, P.; Fontham, E.; Franceschi, S.; Gao, Y.T.; Gupta, P.C.; Hackshaw, A.; Matos, E.; Samet, J.; et al. Tobacco and cancer: Recent epidemiological evidence. *J. Natl. Cancer Inst.* **2004**, *96*, 99–106. [CrossRef]
9. Blot, W.J.; McLaughlin, J.K.; Winn, D.M.; Austin, D.F.; Greenberg, R.S.; Preston-Martin, S.; Bernstein, L.; Schoenberg, J.B.; Stemhagen, A.; Fraumeni, J.F. Smoking and drinking in relation to oral and pharyngeal cancer. *Cancer Res.* **1988**, *48*, 3282–3287.
10. Gasco, M.; Crook, T. The p53 network in head and neck cancer. *Oral Oncol.* **2003**, *39*, 222–231. [CrossRef]
11. Bryant, A.K.; Sojourner, E.J.; Vitzthum, L.K.; Zakeri, K.; Shen, H.; Nguyen, C.; Murphy, J.D.; Califano, J.A.; Cohen, E.E.W.; Mell, L.K. Prognostic role of p16 in nonoropharyngeal head and neck cancer. *J. Natl. Cancer Inst.* **2018**, *110*, 1393–1399. [CrossRef]
12. McKaig, R.G.; Baric, R.S.; Olshan, A.F. Human papillomavirus and head and neck cancer: Epidemiology and molecular biology. *Head Neck* **1998**, *20*, 250–265. [CrossRef]
13. Chau, L.; McNiven, A.; Arjune, B.; Bracken, G.; Drever, L.; Fleck, A.; Grimard, L.; Poon, I.; Provost, D. Dose Objectives for Head and Neck IMRT Treatment Planning Recommendation Report. 2014. Available online: https://www.cancercareontario.ca/sites/ccocancercare/files/guidelines/full/DoseObj_HN_IMRT_TrtmtPlngRec_0.pdf (accessed on 15 March 2023).
14. Gormley, M.; Creaney, G.; Schache, A.; Ingarfield, K.; Conway, D.I. Reviewing the epidemiology of head and neck cancer: Definitions, trends and risk factors. *Br. Dent. J.* **2022**, *233*, 780–786. [CrossRef]
15. Cooper, J.S.; Porter, K.; Mallin, K.; Hoffman, H.T.; Weber, R.S.; Ang, K.K.; Gay, E.G.; Langer, C.J. National Cancer Database Report on Cancer of The Head and Neck: 10-Year Update. *Head Neck* **2009**, *31*, 748–758. [CrossRef]
16. Taylor, A.; Powell, M.E.B. Intensity-modulated radiotherapy—What is it? *Cancer Imaging* **2004**, *4*, 68–73. [CrossRef] [PubMed]
17. Teoh, M.; Clark, C.H.; Wood, K.; Whitaker, S.; Nisbet, A. Volumetric modulated arc therapy: A review of current literature and clinical use in practice. *Br. J. Radiol.* **2011**, *84*, 967–996. [CrossRef] [PubMed]
18. Dagogo-Jack, I.; Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 81–94. [CrossRef] [PubMed]

19. Denison, T.A.; Bae, Y.H. Tumor heterogeneity and its implication for drug delivery. *J. Control. Release* **2012**, *164*, 187–191. [CrossRef] [PubMed]
20. Ganeshan, B.; Miles, K.A. Quantifying tumour heterogeneity with CT. *Cancer Imaging* **2013**, *13*, 140–149. [CrossRef] [PubMed]
21. Lin, G.; Keshari, K.R.; Park, J.M. Cancer metabolism and tumor heterogeneity: Imaging perspectives using MR imaging and spectroscopy. *Contrast Media Mol. Imaging* **2017**, *2017*, 6053879. [CrossRef] [PubMed]
22. Tufail, S.; Riggs, H.; Tariq, M.; Sarwat, A.I. Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. *Electronics* **2023**, *12*, 1789. [CrossRef]
23. Kawashima, Y.; Miyakoshi, M.; Kawabata, Y.; Indo, H. Efficacy of texture analysis of ultrasonographic images in the differentiation of metastatic and non-metastatic cervical lymph nodes in patients with squamous cell carcinoma of the tongue. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2023**, *136*, 247–254. [CrossRef] [PubMed]
24. Safakish, A.; Sannachi, L.; DiCenzo, D.; Kolios, C.; Pejović-Milić, A.; Czarnota, G.J. Predicting Head & Neck Cancer Treatment Outcomes with Pre-Treatment Quantitative Ultrasound Texture Features & Optimizing Machine Learning Classifiers with Texture-of-Texture Features. *Front. Oncol.* **2023**, *13*, 1258970. [CrossRef] [PubMed]
25. Chen, S.-W.; Hsieh, T.-C.; Yen, K.-Y.; Yang, S.-N.; Wang, Y.-C.; Chien, C.-R.; Liang, J.-A.; Kao, C.-H. Interim FDG PET/CT for predicting the outcome in patients with head and neck cancer. *Laryngoscope* **2014**, *124*, 2732–2738. [CrossRef] [PubMed]
26. Tran, W.T.; Suraweera, H.; Quaioit, K.; Cardenas, D.; Leong, K.X.; Karam, I.; Poon, I.; Jang, D.; Sannachi, L.; Gangeh, M.; et al. Predictive quantitative ultrasound radiomic markers associated with treatment response in head and neck cancer. *Future Sci. OA* **2019**, *6*, FSO433. [CrossRef] [PubMed]
27. Tang, S.; Ou, J.; Wu, Y.P.; Li, R.; Chen, T.W.; Zhang, X.M. Contrast-enhanced CT radiomics features to predict recurrence of locally advanced oesophageal squamous cell cancer within 2 years after trimodal therapy A case-control study. *Medicine* **2021**, *100*, e26557. [CrossRef]
28. Yip, S.S.F.; Aerts, H.J.W.L. Applications and limitations of radiomics. *Phys. Med. Biol.* **2016**, *61*, R150–R166. [CrossRef] [PubMed]
29. Shur, J.D.; Doran, S.J.; Kumar, S.; ap Dafydd, D.; Downey, K.; O'connor, J.P.B.; Papanikolaou, N.; Messiou, C.; Koh, D.-M.; Orton, M.R. Radiomics in oncology: A practical guide. *Radiographics* **2021**, *41*, 1717–1732. [CrossRef]
30. Bera, K.; Braman, N.; Gupta, A.; Velcheti, V.; Madabhushi, A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat. Rev. Clin. Oncol.* **2022**, *19*, 132–146. [CrossRef]
31. Eisenhauer, E.A.; Therasse, P.; Bogaerts, J.; Schwartz, L.H.; Sargent, D.; Ford, R.; Dancey, J.; Arbuck, S.; Gwyther, S.; Mooney, M.; et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **2009**, *45*, 228–247. [CrossRef]
32. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341. [CrossRef]
33. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef] [PubMed]
34. Elreedy, D.; Atiya, A.F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Inf. Sci.* **2019**, *505*, 32–64. [CrossRef]
35. Moslemi, A. Engineering Applications of Artificial Intelligence A tutorial-based survey on feature selection: Recent advancements on feature selection. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107136. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 5999–6009.
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
38. Statistics Canada and the Public Health Agency of Canada. Canadian Cancer Statistics: A 2022 Special Report on Cancer Prevalence. Toronto, Ontario, Canada. 2022. Available online: https://www.cancer.ca/Canadian-Cancer-Statistics-2022-EN (accessed on 10 August 2023).
39. Huang, T.-T.; Lin, Y.-C.; Yen, C.-H.; Lan, J.; Yu, C.-C.; Lin, W.-C.; Chen, Y.-S.; Wang, C.-K.; Huang, E.-Y.; Ho, S.-Y. Prediction of extranodal extension in head and neck squamous cell carcinoma by CT images using an evolutionary learning model. *Cancer Imaging* **2023**, *23*, 84. [CrossRef]
40. Niu, J.; Zhang, S.; Ma, S.; Diao, J.; Zhou, W.; Tian, J.; Zang, Y.; Jia, W. Preoperative prediction of cavernous sinus invasion by pituitary adenomas using a radiomics method based on magnetic resonance images. *Eur. Radiol.* **2019**, *29*, 1625–1634. [CrossRef] [PubMed]
41. Malik, N.; Geraghty, B.; Dasgupta, A.; Maralani, P.J.; Sandhu, M.; Detsky, J.; Tseng, C.-L.; Soliman, H.; Myrehaug, S.; Husain, Z.; et al. MRI radiomics to differentiate between low grade glioma and glioblastoma peritumoral region. *J. Neurooncol.* **2021**, *155*, 181–191. [CrossRef] [PubMed]
42. Lizzi, F.L.; Ostromogilsky, M.; Feleppa, E.J.; Rorke, M.C.; Yaremko, M.M. Relationship of Ultrasonic Spectral Parameters to Features of Tissue Microstructure. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **1987**, *34*, 319–329. [CrossRef]

43. Keek, S.; Sanduleanu, S.; Wesseling, F.; de Roest, R.; Brekel, M.v.D.; van der Heijden, M.; Vens, C.; Giuseppina, C.; Licitra, L.; Scheckenbach, K.; et al. Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemoradiotherapy. *PLoS ONE* **2020**, *15*, e0232639. [CrossRef]
44. Ger, R.B.; Zhou, S.; Elgohari, B.; Elhalawani, H.; Mackin, D.M.; Meier, J.G.; Nguyen, C.M.; Anderson, B.M.; Gay, C.; Ning, J.; et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- And PET-imaged head and neck cancer patients. *PLoS ONE* **2019**, *14*, e0222509. [CrossRef]
45. Osapoetra, L.O.; Sannachi, L.; DiCenzo, D.; Quiaoit, K.; Fatima, K.; Czarnota, G.J. Breast lesion characterization using Quantitative Ultrasound (QUS) and derivative texture methods. *Transl. Oncol.* **2020**, *13*, 100827. [CrossRef]
46. 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]