Master of Science Thesis
VT2016

Investigation of the prognostic value of CT and PET-based radiomic image features in oropharyngeal squamous cell carcinoma

Mohammed Mosad Said

Supervision

Crister Ceberg, MSF och Per Nilsson, Johanna Sjövall, Elisabeth Kjellén, SUS

Department of Medical Radiation Physics,
Clinical Sciences, Lund
Lund University
www.msf.lu.se

Abstract

Background: Medical data in the form of radiographic routine scans is steadily accumulating. The analysis of such data through automated quantitative methods is believed to produce new information which would allow for more personalization of therapy. The present thesis investigated the use of such methods in head and neck cancer.

Material and methods: Pretreatment positron emission tomography (PET) and computed tomography (CT) scans from 74 patients present with oropharyngeal squamous cell carcinoma were analyzed quantitatively and a total of 92 image-based features were calculated. These features attempt to describe the shape and size of the tumor, as well as the heterogeneity within. The prognostic value of these features, as well as common clinical variables, was investigated for tumor recurrence and disease-specific mortality, respectively. Additionally, prediction of treatment failure was attempted using an artificial neural network.

All patients received intensity-modulated radiation therapy and there were thus treatment plans for each patient in addition to the PET/CT scans. The non-uniformity of the dose distribution was studied using custom features based on the gray-level size zone matrix. These custom features measured the number of disconnected regions receiving either too low or too high of radiation dose, and differences in the sizes of such regions.

Results: One PET- and two CT-based features were found to significantly differ between responders and non-responders. The PET-based feature was the correlation ($p = 0.0011$), which is a texture feature derived from the gray-level co-occurrence matrix. It described the irregularity in radiotracer uptake on a voxel-to-voxel basis and results suggest that non-responders have more irregular patterns of uptake. The CT-based features were the variance ($p = 0.0012$) and skewness ($p = 0.0027$), where the former was found to be significantly larger among responders and the skewness more negative. However, image-based features performed quite poorly in treatment failure prediction, as compared to clinical variables, which had an area under the receiver operating characteristic curve (AUC-ROC) of 0.87 (95% confidence interval, 0.73–0.96) for primary tumor recurrence and 0.73 (95% confidence interval, 0.52–0.87) for disease-specific mortality. Three image-based features did, however, contribute significantly when included to the model utilizing clinical variables, which suggests that they may contain additional information that is likely to be of value.

Of the five custom features calculated on the dose distribution, the one emphasizing differences in the number of disconnected regions was observed to be significantly higher among non-responders. No statistical differences were found in the sizes of low-dose and high-dose regions, respectively, between the two groups.

Conclusion: Quantitative analysis of routine scans may provide additional information regarding tumor phenotype, which is likely to be of value when used in conjunction with clinical variables. Additionally, texture analysis of the dose distribution reveals differences between treatment plans that are not captured by dose-volume histogram metrics. These methods are, however, relatively new in use on medical data and there are certain details that require further investigation.

Summary in Swedish

En av de stora utmaningarna inom onkologin är den stora heterogeniteten mellan olika tumörer. Ingen tumör är den andra lik och detta skapar ett behov av individualiserad vård där behandlingen skräddarsys efter patienten. I regel grundas behandlingsalternativen på populationsbaserad forskning och den givna behandlingen är därmed optimal för en så kallad "medelpatient", men en sådan existerar inte och i verkligheten är det vissa som överbehandlas medan andra underbehandlas. Stora framsteg har gjorts i denna individualisering genom rutinmässiga vävnadsprover och identifieringen av ett stort antal biomarkörer. Studier har dock visat på stora variationer i genuttryck inom en och samma tumör och nackdelen med vävnadsprover är att de återger endast en liten del av tumören och därmed misslyckas med att fånga denna heterogenitet.

Ytterligare ett framsteg är den rutinmässiga bildtagningen med datortomografi (CT), som ger en god helhetsbild av tumörstorleken samt spridningen i kroppen. På senare tid har även molekylär bildtagning blivit rutin och information gällande tumörfunktion har blivit tillgänglig. Utifrån dessa bilder görs en någorlunda kvalitativ bedömning som bidrar enormt till behandlingsstrategin. En kvantitativ analys av tumören, som sedd på bilder tagna innan behandlingsstart, kan dock mynna i ny information som har visats vara användbar för att prediktera behandlingsrespons. Analysen är automatiserad och syftar på att undersöka storleken så väl som formen på tumören, men också den inre heterogeniteten, vilket beroende på bildmodalitet kan spegla olika saker.

Det finns idag en enorm mängd data i form av medicinska bilder som bara ökar med tiden. Radiomics är ett relativt nytt ämnesområde där dessa bilder analyseras genom kvantitativa metoder i syftet att utvinna värdefull information, som kan användas i individualiseringen av framtida behandlingar. Den stora mängden data vidare möjliggör användandet av avancerade maskininlärningsmetoder för bättre risk-stratifiering.

I detta examensarbete undersöktes det prognostiska värdet av ett relativt stort antal kvantitativa parametrar beräknade på både CT och positron emission tomography (PET). Av de totalt 92 beräknade parametrarna var det en PET-baserad och två CT-baserade parametrar som lyckades differentiera mellan patienter med positiv och negativ behandlingsrespons. Utöver det användes ett artificiellt neuronnät för att prediktera behandlingsrespons utifrån både bildbaserade parametrar och kliniska variabler. Med denna information, som är tillgänglig redan innan behandlingsstart, kunde responsen predikteras med ganska god säkerhet, vilket kan vara värdefullt ur klinisk synpunkt eftersom alternativa behandlingsstrategier kan utforskas i ett tidigt stadium.

Inom strålbehandling finns dessutom information om dosfördelningen, som bör uppfylla angivna doshomogenitetskrav. Uniformiteten av dosfördelningen analyserades med hjälp av fem speciellt framtagna parametrar, som tog hänsyn till det spatiala förhållandet mellan både lågdos- och högdosområden. Resultaten visade att antalet osammanhängande lågdos- och högdosområden kan vara av betydelse för behandlingsrespons. Detta är dock en hittills obeprövad metod för analys av dosfördelningar och det finns utrymme för förbättringar.

Contents

Abbreviations

^{18}F-FDG	^{18}F-fluorodeoxyglucose
ANN	Artifical neural network
AUC-CSH	Area under the cumulative standardized uptake value histogram
AUC-ROC	Area under the receiver operating characteristic curve
CT	Computed tomography
DVH	Dose-volume histogram
FDR	False discovery rate
GLCM	Gray-level co-occurrence matrix
GLRLM	Gray-level run length matrix
GLSZM	Gray-level size zone matrix
GTV-N	Gross tumor volume - nodal tumor
GTV-T	Gross tumor volume - primary tumor
HNC	Head and neck cancer
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
HU	Hounsfield unit
IMRT	Intensity-modulated radiation therapy
MLP	Multi-layer perceptron
MRI	Magnetic resonance imaging
MRMR	Minimum redundancy maximum relevance
NGTDM	Neighborhood gray-tone difference matrix
NSCLC	Non-small cell lung cancer
OBS	Optimal brain surgeon
OPC	Oropharyngeal cancer
PET	Positron emission tomography
ROI	Region of interest
SCC	Squamous cell carcinoma
SUV	Standardized uptake value
TNM	Tumor-node-metastasis
VOI	Volume of interest

1 Introduction

Head and neck cancer (HNC) accounts for a relatively small fraction of the global cancer burden, and although not the most common type of cancer, HNC has a generally unfavorable prognosis. With almost 700,000 new cases and 400,000 deaths annually worldwide, the 5-year overall survival rate is approximately 50%; though, it depends highly on disease site and stage [1].

The poor prognosis is partly due to the advanced stage at which most patients are diagnosed, where lymph node involvement is quite common. Advances in diagnostics are sure to remedy a great deal, but with the advent of new treatment modalities and the evident heterogeneity between tumors of the same type, the importance of risk stratification for the purpose of personalized medicine becomes apparent. This is especially true for HNC due to the structural complexity of the head and neck region and the close proximity of multiple risk organs, which often entail an organ preserving treatment rationale.

1.1 Head and neck cancer

Malignancies of the head and neck are to a large extent squamous cell carcinomas (SCC) arising from the mucosal lining. Head and neck squamous cell carcinoma (HNSCC) is categorized after site of tumor origin and common sites include the oral and nasal cavities, nasopharynx, oropharynx, hypopharynx and the larynx. Oropharyngeal cancer (OPC) specifically is further divided into tonsillar cancer, and cancer originating from the base of the tongue, the soft palate, and the pharyngeal wall. Prognosis differs markedly between different tumor sites and is worsened by the advanced stage at which pathology is usually confirmed.

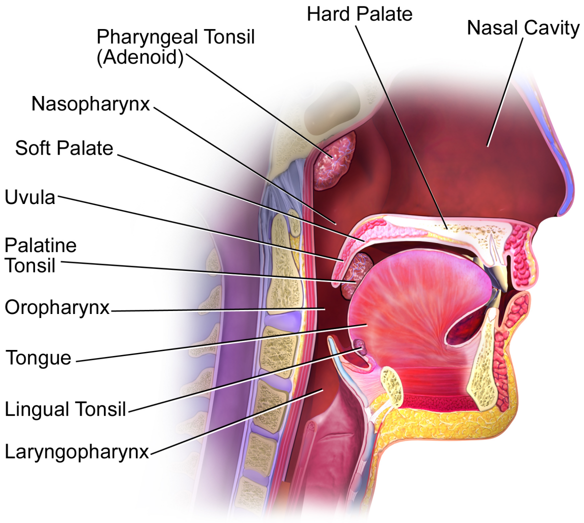


Figure 1: Anatomy of the head and neck [2].

The regional lymphatic system is quite complex and comprise a large number of lymph nodes to which disease may spread. The tumor-node-metastasis (TNM) classification system is the standard for cancer staging and contain information regarding size of the primary tumor, degree and location of regional spread, as well as presence of distant metastases. It is strongly prognostic and thus universally utilized in clinical practice.

Risk factors for HNC include tobacco and alcohol consumption; the combination thereof disproportionately increases risk of carcinogenesis. Despite the steady decline of tobacco- and alcohol-related HNC, the incidence of OPC has increased noticeably during the last decades, which has been linked to a corresponding increase of human papillomavirus (HPV) positive OPC [3]. HPV-infection is now a confirmed risk factor and is more prevalent in a younger and predominantly male population.

Human papillomavirus

There is ample evidence for an association between certain subtypes of HPV-infections and HNC, especially so for oropharyngeal SCC [4]. Additionally, the role of HPV as an etiologic factor in carcinogenesis is well understood and the presence of these viral oncogenes is significantly associated with better prognosis in HNC [4, 5]. At present, HPV-status is not widely utilized for clinical decision making, even though fine-needle biopsies are performed routinely and subsequent analysis yields said status.

Recent studies have demonstrated the ability to differentiate between HPV-positive and HPV-negative HNSCC through tumor texture analysis on contrast-enhanced computed tomography (CT) [6, 7]. Texture analysis on radiographic images yield measures of tumor heterogeneity on a macroscopic scale determined mainly by the imaging modality. The feasibility of HPV-status assessment through texture analysis might be plausible seeing as HPV-positive and HPV-negative HNSCC are in many aspects completely different entities, e.g., etiology, prognosis, and histopathology [4, 7].

1.2 Quantitative imaging

Radiographic imaging is routinely used in diagnostics, staging, and treatment response monitoring. Quantitative methods are, however, largely limited to e.g. measures of tumor cross-sectional diameter on CT and standardized uptake value (SUV) parameters in positron emission tomography (PET). The latter is commonly referred to as a semi-quantitative measure and is highly sensitive to a plethora of variables [8], while the former is often done manually and is therefore relatively time consuming and subject to a degree of interobserver variability.

Radiomics refers to the automatic extraction of a large number of quantitative tumor descriptors and subsequent analysis of clinical importance. Conceptually, it is quite similar to the field of genomics, where tumor phenotype is assessed through gene expression data rather than quantitative image-based features. The radiomics approach is not restricted to a specific imaging modality and includes descriptors of tumor size and shape, as well as intratumoral heterogeneity. These tumor characteristics are general hallmarks of malignancy, where the latter is due to genomic instabilities and microenviromental differences that are often expressed as increased cell proliferation, hypoxia, angiogenesis, and necrosis. Whether texture analysis on radiographic images successfully probes this underlying heterogeneity is currently unclear but there appears to be an association between increased heterogeneity as observed on CT and radiotracer uptake on PET, respectively, and these physiological processes [9–13].

Several studies have recently demonstrated prognostic value of texture features on both CT [13–16] and PET [17–20]. Through analysis of pretreatment CT scans from 1,019 patients, Aerts et al. have found evidence of prognostic potential for image-based biomarkers that are presently not utilized. The patients were from seven independent cohorts and comprised both non-small cell lung cancer (NSCLC) and HNSCC, which suggest that radiomics features might translate between different tumor sites [13]. The four radiomic features declared prognostic constitute a so called radiomic signature, which was externally verified in an independent cohort comprising 542 oropharyngeal SCC patients [21]. Similar results are presented from PET, although on a much smaller scale regarding both the number of patients included and the number of features calculated. Interestingly, however, texture features were shown to be significantly better at differentiating between non-responders, partial responders, and responders in esophageal cancer than SUV parameters [17]. Texture analysis is in fact more commonly used in magnetic resonance imaging (MRI), where high classification accuracies have been achieved in breast cancer through advanced machine learning algorithms that utilize texture features [22, 23]. Despite all compelling results, the reproducibility needs to be assessed and sources of variation need to be examined.

1.3 Thesis objectives

The main objective of the present thesis is to investigate whether quantitative analysis of pretreatment PET/CT scans allows for risk stratification with respect to treatment response in oropharyngeal SCC. Additional information available prior to treatment, such as TNM-classification and planned dose distribution, are also examined. Finally, response prediction is attempted through training of an artificial neural network. The thesis is thus divided in the following parts:

1. Quantitative analysis of tumor size, shape, and heterogeneity as observed on pretreatment PET/CT scans, and how these features associate with different clinical outcomes, as well as HPV-status.
2. Analysis of non-uniformity of dose distribution through a novel method based on texture analysis, where the spatial distribution between underdosed and overdosed tumor subvolumes, respectively, is taken into account.
3. Prediction of treatment failure through artificial neural networks trained with quantitative image-based features and clinical information, respectively.

2 Texture features

There is no universally agreed upon definition of what classifies as texture. Common descriptions often refer to properties such as smoothness, coarseness, granularity, etc. The lack of a strict definition is partly explained by the intricate relationship between texture and tone — where a preponderance of one conceals the other. Nonetheless, there are four different approaches by which texture can be analyzed: statistical, structural, model-based, and transform-based methods. This section contains a short theory of the statistical methods used in this study.

2.1 First-order histogram

First-order histograms are commonly used in descriptive statistics and contain valuable information of the image content. The arithmetic mean and median together with the higher central moments constitute a well-rounded set of features that adequately describe the global distribution of pixel intensities.

2.2 Gray-level co-occurrence matrix

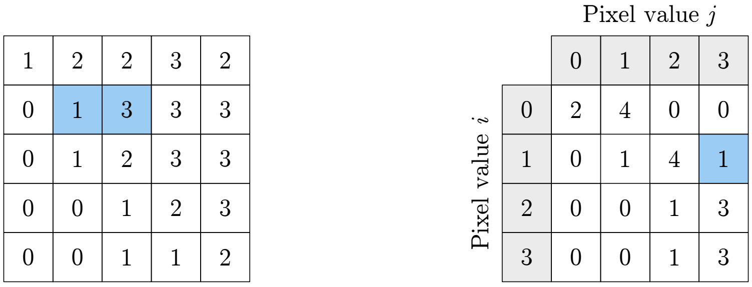
Features derived from the first-order histogram, e.g. the variance, skewness, and kurtosis, are global descriptors of the distribution of pixel values in an image or region of interest (ROI). Their capacity to describe texture is limited because the spatial relationship between pixel values remains unutilized. Texture measures derived from the gray-level co-occurrence matrix (GLCM), however, do capture this relationship and were first proposed by Haralick et al. [24].

The GLCM $\mathbf{C}(i, j, \mathbf{d})$ of an image contains the frequencies at which gray-level i and j co-occur at distance and direction specified by the vector $\mathbf{d}$. There are four possible directions for a two-dimensional image, i.e., horizontal (0°), vertical (90°), and the two diagonals (45° and 135°). Although the distance is restricted only by the dimensions of the digital image, it is often set to one pixel as to best capture the underlying spatial relationship and the physical limitation is thus determined by the spatial resolution. The normalized GLCM

$$\mathbf{P}(i, j) = \frac{\mathbf{C}(i, j)}{n}, \tag{1}$$

yields the probability, rather than frequency, of finding gray-level i and j at the predetermined distance and direction. The variable n in equation (1) is simply the number of entries in the GLCM and $\mathbf{P}(i, j)$ is essentially a two-dimensional normalized histogram of the image.

The square image in Figure 2a has four possible gray-levels and its GLCM for $\mathbf{d}$ equal to one pixel to the immediate right is shown in Figure 2b. There is for example only one instance where the pixel value 3 is to the right of pixel value 1 and therefore $\mathbf{C}(1, 3) = 1$. This choice of $\mathbf{d}$ is not equivalent to $|\mathbf{d}| = 1$ in the horizontal direction because co-occurrences to the immediate left are omitted. Due to symmetry, however, the GLCM for the horizontal direction is simply the sum of the GLCM in Figure 2b and its transpose.



1	2	2	3	2
0	1	3	3	3
0	1	2	3	3
0	0	1	2	3
0	0	1	1	2

(a)

Pixel value i \ Pixel value j	0	1	2	3
0	2	4	0	0
1	0	1	4	1
2	0	0	1	3
3	0	0	1	3

(b)

Figure 2: **(a)** The pixel values of a 2-bit image of size 5×5 and **(b)** the corresponding GLCM for **d** equal to one pixel to the immediate right. Note that the GLCM is a square matrix with dimensions equal to the number of gray-levels in the original image and that the blue-colored entry in the GLCM corresponds to the blue-colored co-occurrence in the image.

The normalized mean GLCM over all four directions is the basis for further calculations of GLCM texture features. The features are essentially the sum of the GLCM with different weightings applied. There are numerous such weightings and hence just as many GLCM features. For instance, the contrast as defined by Haralick et al. is

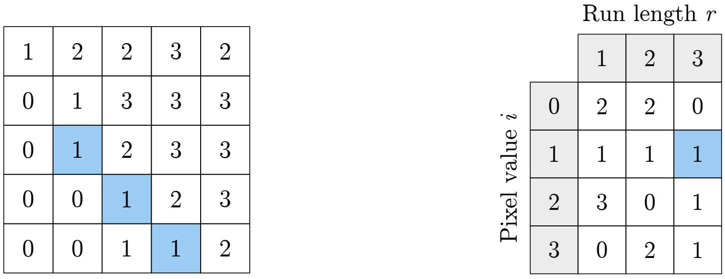
$$\text{Contrast} = \sum_{i=1}^{L}\sum_{j=1}^{L}(i-j)^2 P(i,j), \tag{2}$$

where L is the number of gray-levels in the quantized image [24]. The factor $(i-j)^2$ suppresses all contribution from the diagonal elements in the GLCM while increasing the contribution from elements further away by an amount proportional to the distance squared. An image consisting mainly of the same gray-level will have few off-diagonal elements in the GLCM and thus little to no contrast. Additional features are presented in Appendix A.

2.3 Gray-level run length matrix

Features extracted from the GLCM are statistics of the second-order seeing as co-occurances involve pairs of pixel values. Higher-order statistical features consider more than two pixel values at a time and features computed from the gray-level run length matrix (GLRLM) are examples of such statistics. They were introduced shortly after GLCM-based features by Galloway [25] and have since then been used extensively.

The spatial distribution of pixel values are utilized by counting the lengths of consecutive runs of same gray-level values. In similarity with the GLCM, the elements of the GLRLM $\mathbf{R}(i, r, \mathbf{d})$ contain the number of runs of length r and gray-level i in a direction specified by $\mathbf{d}$. As previously, there are four principal directions in two-dimensions and therefore just as many matrices, which are usually averaged prior to the calculation of the features. The GLRLM in the diagonal direction (135°) of the example image in Figure 2a is illustrated in Figure 3b, where the blue-colored entry corresponds to the run of length 3 by pixels with a gray-level equal to 1, as depicted in Figure 3a.



1	2	2	3	2
0	1	3	3	3
0	1	2	3	3
0	0	1	2	3
0	0	1	1	2

(a)

Run length r

Pixel value i	1	2	3
0	2	2	0
1	1	1	1
2	3	0	1
3	0	2	1

(b)

Figure 3: **(a)** An image consisting of four different gray-levels and **(b)** its corresponding GLRLM for diagonal direction (135°). The size of the GLRLM depends on the number of gray-levels and the length of the longest consecutive run.

In addition to averaging over all directions, the GLRLM is normalized to obtain the probabilities $\mathbf{P}(i, r)$ after which features corresponding to different weightings, or emphasis, are computed. For instance, the short run emphasis (SRE) is calculated as

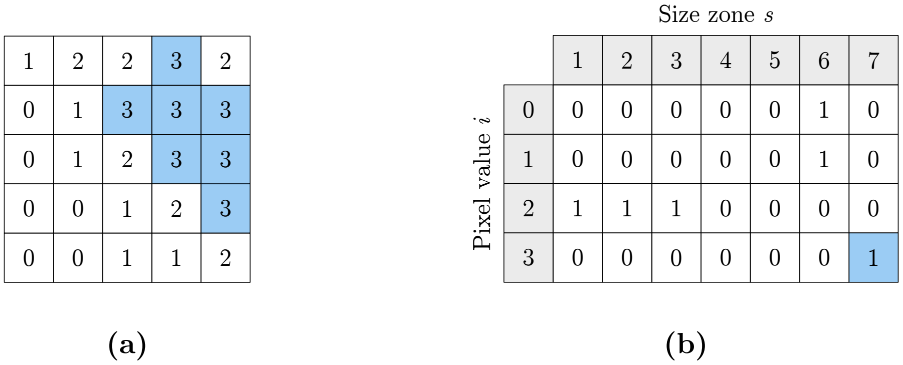
$$\text{SRE} = \sum_{i=1}^{L} \sum_{r=1}^{L_r} \frac{P(i,r)}{r^2}, \tag{3}$$

where L and L_r are the number of gray-levels and the length of the longest run, respectively. The r^2 in the denominator effectively suppresses the contribution of longer runs, hence emphasizing short runs. An image expressing a high degree of homogeneity in a region would likely have a small SRE due to a higher frequency of longer runs. Additional features are designed to capture other aspects of the GLRLM and those used in this work are presented in Appendix A.

2.4 Gray-level size zone matrix

The gray-level size zone matrix (GLSZM) is the basis of another set of higher-order statistical features and is quite similar to the GLRLM in many aspects. The main difference being that sizes of contiguous regions, rather than lengths of consecutive runs, of equal gray-level are counted and assembled in the exact same manner as in the GLRLM [26]. Additionally, it differs from the GLCM and GLRLM in that there is no directionality to consider, i.e., there is only one GLSZM $\mathbf{S}(i, s)$ and thus no direction averaging prior to feature calculation. In the spirit of previous illustrations, the GLSZM is given in Figure 4b, where the blue-colored entry $\mathbf{S}(3, 7) = 1$ corresponds to the rather large region consisting of pixels of value 3 in Figure 4a.

The GLSZM is of size $L \times L_z$, where L_z is the size of the largest contiguous zone. Different weightings of the normalized GLSZM are used to emphasize different characteristics of the matrix and a complete list over all GLSZM-based features used in this study is given in Appendix A.



(a)

1	2	2	3	2
0	1	3	3	3
0	1	2	3	3
0	0	1	2	3
0	0	1	1	2

(b)

Size zone s

Pixel value i	1	2	3	4	5	6	7
0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0
2	1	1	1	0	0	0	0
3	0	0	0	0	0	0	1

Figure 4: **(a)** The image from the previous examples and **(b)** its GLZSM, where the blue-colored entry corresponds to the blue-colored region in the image.

2.5 Neighborhood gray-tone difference matrix

The neighborhood gray-tone difference matrix (NGTDM) differs from other gray-level matrices in that it is a one-dimensional matrix of length equal to the number of gray-levels L. It was developed for the purpose of imitating the human perception of texture and thereby enabling the calculation of quantitative descriptors that closely resemble its visual properties [27].

Each element of the NGTDM $\mathbf{D}(i)$ corresponds to the sum of the absolute difference between pixels of value i and the average of their neighbors

$$D(i) = \begin{cases} \sum_{j \in \{N_i\}} \left| i - \bar{A}_j \right| & : N_i \neq 0 \\ 0 & : N_i = 0, \end{cases} \tag{4}$$

where $\{N_i\}$ is the set of pixels with gray-level i. The neighborhood average $\bar{A}_j$ for a neighborhood of size one, i.e., only the immediate neighbors, is equal to

$$\bar{A}_j = \frac{1}{8} \sum_{m=-1}^{1} \sum_{n=-1}^{1} f(k+m, l+n) \qquad : \text{ for } (m,n) \neq (0,0), \tag{5}$$

where $f(k,l)$ is the pixel of value i. Equation (5) is valid only for non-peripheral regions of size one and in two-dimensions. However, it can easily be extended to three-dimensions — as can the other gray-level matrices.

The texture features are derived in a similar manner as previously, although without the normalization step that is usually a prerequisite. Five different features are defined in [27], namely, the coarseness, contrast, busyness, complexity, and strength. Their definition is given in Appendix A as well.

3 Artificial neural networks

An artifical neural network (ANN) is a nonlinear model capable of constructing decision boundaries of arbitrary complexity. Since the resurgence in interest following the formulation of the back-propagation algorithm, it has been used extensively for pattern recognition and classification tasks in different areas of research and application. The number of clinical trials utilizing ANNs for the purpose of cancer diagnosis has increased almost 20-fold between 1994–2003 [28]. ANNs in image-based cancer diagnosis and classification have been fairly successful in numerous disease sites such as in the detection of pulmonary nodules on chest radiograms [29], and the differentiation between benign and malignant lesions in breast cancer [22, 23, 30].

There are many different types of ANNs and an even larger number of configurations. This section contains the fundamental theory as well as theory of the methods used in this work, as based on the literature by Bishop (1995) and Haykin (2009) [31, 32].

3.1 The artificial neuron

The development of ANNs began with the mathematical model of biological neural networks published in 1943. It was from thereon heavily inspired by the field of neurobiology, with the motivation of imitating the most complex of information processing systems — the brain.

A neural cell, which is the constitutive unit of a biological neural network, can in an overly naive manner be described as consisting of a cell body containing the nucleus etc.; dendrites branching out of the cell body to form a dendritic tree; and an axon extending far out from the cell body, as illustrated in Figure 5a. Synaptic inputs are received at the dendrites and integrated in the cell body before being conveyed to the axon hillock, where an action potential is generated if the sum of the inputs exceeds a certain firing threshold. The action potential propagates through the axon and onwards to the synaptic terminals, where the signal is transmitted further. A neuron is thus, in simple terms, responsible for the transmission of electrochemical impulses.

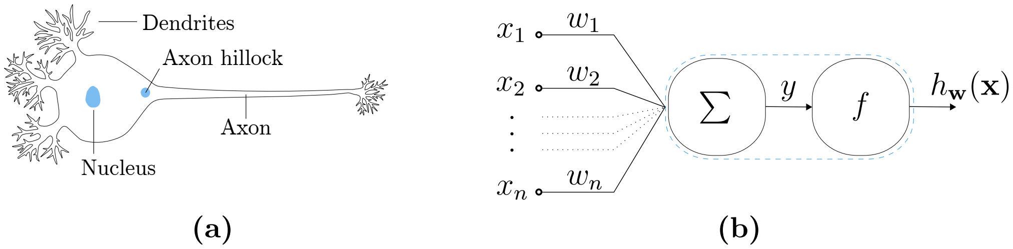


Figure 5: **(a)** Simple illustration of a biological neuron and **(b)** artificial neuron.

The principle of an artificial neuron is similar in that the response to an arbitrary number of input arguments x_n depends on an activation function f applied to the weighted sum of the inputs. The synaptic weights w_n in Figure 5b are introduced as to model the connectivity strengths of different synapses. Mathematically, the response h for an input vector $\mathbf{x}$ using a set of weights $\mathbf{w}$

$$h_{\mathbf{w}}(\mathbf{x}) = f(y), \tag{6}$$

where the activation function with threshold T

$$f(y) = \begin{cases} 1 & : y \geq T \\ 0 & : y < T, \end{cases} \tag{7}$$

and the weighted sum of the inputs is

$$y = \sum_{n=1}^{N} w_n x_n = \mathbf{w}^\intercal \mathbf{x}, \tag{8}$$

where N is the number of input arguments.

This is essentially the construction of a perceptron, which is a linear model capable of binary classification through supervised learning. The weights are obtained through the implementation of a learning algorithm on a training set and supervised learning refers to the requirement of labeled training samples, i.e., that every training sample in the set contains the input arguments $\mathbf{x}$ and the corresponding desired output d.

Activation functions

The perceptron differs from logistic regression mainly in the choice of the Heaviside step function as an activation function, which is rather undesirable in its property of being non-continuous. Differentiability of the activation function is a prerequisite for training a network of perceptrons using the back-propagation algorithm. There are several viable choices of activation functions, e.g., the sigmoid function and the hyperbolic tangent function, where the former

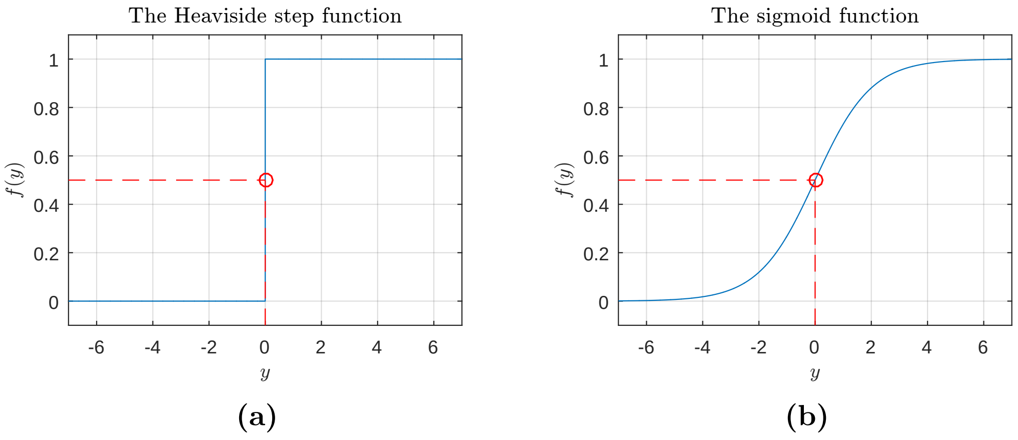
$$f(y) = \frac{1}{1 + \mathrm{e}^{-(y+T)}}, \tag{9}$$

is plotted in Figure 6 together with the Heaviside step function in equation (7).

The desired dichotomous behavior of the step function is preserved using the sigmoid function, with the added property of $f(y)$ corresponding to the posterior probability $P(1 \mid \mathbf{w}, \mathbf{x})$. Moreover, the derivative of the sigmoid function

$$\frac{\mathrm{d}f(y)}{\mathrm{d}y} = f(y)(1 - f(y)), \tag{10}$$

is particularly convenient in its simplicity.



(a) **(b)**

Figure 6: **(a)** The Heaviside step function with threshold $T = 0$ and **(b)** the sigmoid function with the same threshold.

The threshold

The issue of finding the optimal threshold value is easily solved by incorporating it into the training procedure. This is done by augmenting the input vector $\mathbf{x}$ to include a bias $x_0 = 1$, and the weighted sum is thus

$$I = \sum_{n=0}^{N} w_n x_n = w_0 + \sum_{n=1}^{N} w_n x_n = w_0 + y, \tag{11}$$

where w_0 corresponds to the threshold T in equation (9). As a result, the training algorithm yields the optimal threshold in addition to the synaptic weights. These parameters are optimal in the sense that the misclassification error is minimized on the training set. Whether a high accuracy is preserved on unseen samples, however, is a matter of generalization.

3.2 Multi-layer perceptron

A single perceptron is a linear classifier; however, when arranged into several layers, it is capable of constructing nonlinear decision boundaries whose complexity increases with the number of perceptrons, and more so with the number of layers. The multi-layer perceptron (MLP) in Figure 7 is a feed-forward network containing an input layer, several hidden layers, and an output layer consisting of a single node in the case of binary classification.

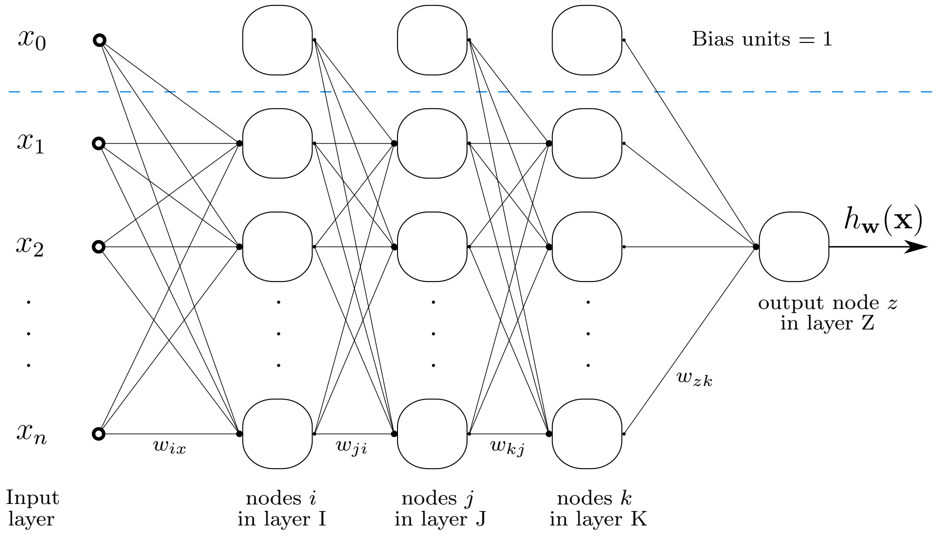


Figure 7: A feed-forward multi-layer neural network with three hidden layers I, J, and K, respectively, and an output layer Z containing a single output node.

Each node in the network functions as a perceptron with input I as in equation (11) and output O given by the activation function, i.e.,

$$I_k = \sum_{j=0}^{J} w_{kj} O_j, \tag{12}$$

and

$$O_k = f(I_k) = \frac{1}{1 + \mathrm{e}^{-I_k}}, \tag{13}$$

correspond to the input and output of a node in layer K, respectively, where O_j is the output of a node in the preceding layer J. The nodes are thus immensely linked and the input vector $\mathbf{x}$ propagates through the entire MLP network before a response $h_{\mathbf{w}}(\mathbf{x}) = O_z$ is produced in the output layer.

The connectivity of the network in Figure 7 is quite common due to its simplicity, but it is not necessarily optimal. The optimal number of nodes in each layer together with the number of layers and connectivity depend on both the objective and available data. These parameters decide the architecture, or topology, which is commonly determined heuristically, although several algorithms exist that incorporate the process of finding the optimal topology into the learning algorithm.

Decision boundaries

The role of each node and layer in an MLP is best illustrated with the well-known exclusive-or problem, which is modified to a concrete example rather than the implementation of the exclusive-or logical operator. Given the task of predicting treatment failure in patients who have received either

– no treatment,
– chemotherapy (CTx),
– radiotherapy (RTx) or
– chemoradiotherapy (CRTx),

and assuming adverse outcomes for the alternative of no treatment and CRTx, a single perceptron would at best produce a decision boundary similar to that in Figure 8a or 8b, where the shaded region corresponds to prediction of treatment failure. In either case, a misclassification error occurs due to the perceptron's incapability of discerning linearly inseparable classes.

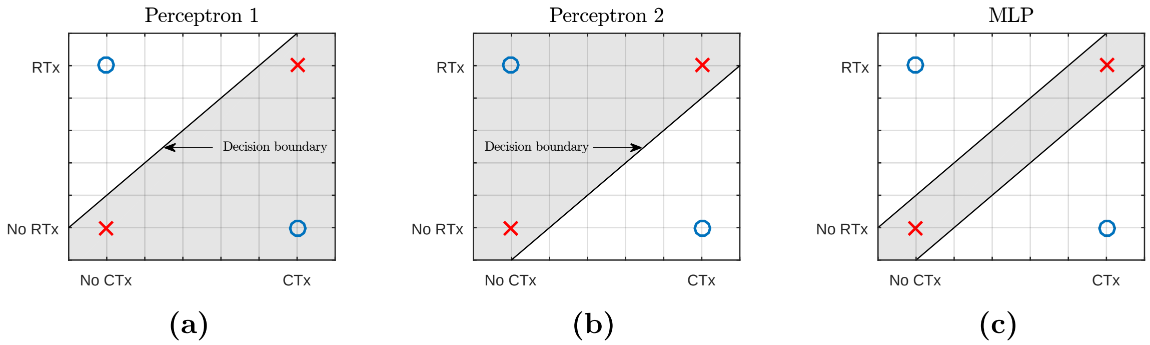


Figure 8: Example of three different decision boundaries in the prediction of treatment failure (crosses), where **(a)** and **(b)** are boundaries resulting from single perceptrons and **(c)** is that of an MLP consisting of the previous two perceptrons and an additional output node.

The intersection of the shaded regions in Figure 8a and 8b would adequately separate the two classes and therefore successfully predict treatment failure. The decision boundary in Figure 8c is obtained by the simplest MLP, namely, an MLP with a single hidden layer consisting of two nodes and a single output node. The two hidden nodes correspond to the perceptrons constructing the boundaries in Figure 8a and 8b, whereas the output node's function is to activate as a response to the simultaneous activation of both nodes in the hidden layer, i.e., $O_z = 1$ if and only if a given observation lies in both shaded regions.

Generally, every node in the first hidden layer corresponds to a hyperplane of dimensions $N - 1$, where N is the number of inputs excluding the bias. The next layer constructs regions contained by the hyperplanes, while additional hidden layers give rise to decision boundaries of arbitrary complexity. An MLP is thus capable of forming nonlinear hypotheses whose complexity is mainly determined by the network topology.

3.3 The back-propagation algorithm

The back-propagation algorithm is a powerful technique for learning the weights of a feed-forward network of arbitrary topology. Initially, the weights are assigned random values within a specified interval and then iteratively updated by an amount $\Delta\mathbf{w}$. The optimal weights are eventually obtained when the performance of the ANN remains unchanged, which necessitates the definition of a performance measure. The cross-entropy error of an ANN with a single output unit is

$$E_{\text{tot}} = -\frac{1}{M}\sum_{m=1}^{M} d_m \log(h_m) + (1 - d_m)\log(1 - h_m), \tag{14}$$

where d is the desired response and M is the number of training samples. There are, of course, other measures of performance, e.g., the mean squared error, although the cross-entropy is more suitable for binary classification while still being mathematically convenient. To reduce the notational clutter, only the error attributed from a single training sample is considered.

The output h of the network depends on all its weights, and consequently so does the cross-entropy error. It is thus simply a matter of minimizing the error with respect to the weights in an efficient manner. The method of gradient descent is suitable for this task, where the optimal set of weights are obtained through an iterative process. The updated value of a specific weight w_{zk}, mapping between the output node z and a node in the preceding layer K is

$$w_{zk}^{(t+1)} = w_{zk}^{(t)} - \eta \left.\frac{\partial E}{\partial w_{zk}}\right|_{w_{zk}^{(t)}}, \tag{15}$$

where t and η are the iteration step and learning rate, respectively. The correction applied after a single iteration is thus

$$\Delta w_{zk} = -\eta \left.\frac{\partial E}{\partial w_{zk}}\right|_{w_{zk}^{(t)}}. \tag{16}$$

The computation of the gradient in equation (16) requires the repeated implementation of the chain rule as follows

$$\frac{\partial E}{\partial w_{zk}} = \frac{\partial E}{\partial O_z}\frac{\mathrm{d}O_z}{\mathrm{d}I_z}\frac{\partial I_z}{\partial w_{zk}}. \tag{17}$$

The middle derivative on the right-hand side of equation (17) is simply $f'(I_z)$ previously evaluated in equation (10). With $O_z = h$, the derivative of the cross-entropy error function with respect to O_z is

$$\frac{\partial E}{\partial O_z} = \left(\frac{d-1}{h-1} - \frac{d}{h}\right), \tag{18}$$

and

$$\frac{\partial I_z}{\partial w_{zk}} = \frac{\partial}{\partial w_{zk}} \sum_{k=1}^{K} w_{zk} O_k = O_k. \tag{19}$$

Substituting these results into equation (16) and (17) yields

$$\Delta w_{zk} = -\eta \left(\frac{d-1}{h-1} - \frac{d}{h}\right) h(1-h) O_k = -\eta(h-d) O_k, \tag{20}$$

which is remarkable in that the updates to weights mapping between the output layer and the preceding layer is calculated solely from the desired output and the outputs of the nodes being connected. This is valid only for weights mapping to the output layer; however, the derivation of the corresponding equation for weights of the hidden layers is essentially identical with the difference of involving additional partial derivatives.

The correction to a weight w_{kj} mapping between the final hidden layer K and the preceding layer J is analogous to equation (16) with the corresponding gradient

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial O_z} \frac{\mathrm{d} O_z}{\mathrm{d} I_z} \frac{\partial I_z}{\partial O_k} \frac{\mathrm{d} O_k}{\mathrm{d} I_k} \frac{\partial I_k}{\partial w_{kj}}. \tag{21}$$

Equation (21) implies a successive decrease in computational efficiency as updates of weights further from the output node are processed. This appears at first as a legitimate concern, although, upon closer inspection, the first two derivatives in equation (21) have been evaluated in the calculation of Δw_{zk}. These two derivatives are thus strategically defined as the error δ_z attributed to nodes in the output layer

$$\delta_z = \frac{\partial E}{\partial I_z} = \frac{\partial E}{\partial O_z} \frac{\mathrm{d} O_z}{\mathrm{d} I_z} = h - d. \tag{22}$$

Similarly, the error attributed to nodes of the preceding layer K is

$$\delta_k = \frac{\partial E}{\partial I_k} = \delta_z \frac{\partial I_z}{\partial O_k} \frac{\mathrm{d} O_k}{\mathrm{d} I_k}, \tag{23}$$

and the gradient in equation (21) becomes

$$\frac{\partial E}{\partial w_{kj}} = \delta_z \frac{\partial I_z}{\partial O_k} \frac{\mathrm{d} O_k}{\mathrm{d} I_k} \frac{\partial I_k}{\partial w_{kj}} = \delta_k \frac{\partial I_k}{\partial w_{kj}}. \tag{24}$$

Generally, the correction to a weight in a hidden layer depends on the updates in all subsequent layers. It is therefore computationally efficient to calculate the updates to weights in the output layer first (thereby obtaining δ_z) and then progress backwards — hence the name back-propagation.

What remains now is the evaluation of equation (24), where the last partial derivative

$$\frac{\partial I_k}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}} \sum_{j=1}^{J} w_{kj} O_j = O_j, \tag{25}$$

and similarly

$$\frac{\partial I_z}{\partial O_k} = \frac{\partial}{\partial O_k} \sum_{k=1}^{K} w_{zk} O_k = w_{zk}. \tag{26}$$

The error attributed to nodes in the final hidden layer is thus

$$\delta_k = \delta_z f'(I_k) w_{zk} = (h-d) O_k (1-O_k) w_{zk}, \tag{27}$$

and the update to a weight w_{kj} connecting the final hidden layer K and the preceding layer J is

$$\Delta w_{kj} = -\eta(\delta_k O_j) = -\eta(h-d) O_k (1-O_k) w_{zk} O_j, \tag{28}$$

whereas for weights immediate to the output layer Z

$$\Delta w_{zk} = -\eta(\delta_z O_k) = -\eta(h-d) O_k. \tag{29}$$

To generalize beyond the two outermost layers, the update to an arbitrary weight w_{ji} connecting node j in layer J, which may or may not be the output layer, to node i in the preceding layer I, is

$$\Delta w_{ji} = -\eta(\delta_j O_i), \tag{30}$$

$$\delta_j = \begin{cases} \sum_{k=1}^{K} w_{kj} O_j (1-O_j) \delta_k & : j \neq z \\ h-d & : j = z, \end{cases} \tag{31}$$

where K is the layer following J. The optimal weights can thus be obtained through an iterative process now that the gradient can be computed.

3.4 Optimization

The procedure of finding the optimal weights is in reality not straightforward and there are several pitfalls innate in the search. One being the difficulty of finding the global minimum rather than one of the many local minima and there are numerous techniques aimed at solving this optimization problem.

Training scheme

There are essentially three training schemes in which the training set is utilized differently. An update of the weight vector **w** may be based on either a single training sample or the entire set, which correspond to the methods of stochastic and batch gradient descent, respectively. Stochastic gradient descent has the advantage of adding an element of randomization inherent to the noise in the training samples, whereby shallow local minima may be overcome. This is, however, only a minor effect and the main advantage appears with extremely large data sets where the issue of insufficient computer memory materializes.

A disadvantage with stochastic gradient descent is its incompatibility with a few advanced training algorithms, which either rely on information derived from the entire data set or are computationally costly for use with separate training cases. An appropriate choice in the latter is the method of mini-batch gradient descent, where a subsample rather than the entire set is utilized in the application of each correction $\Delta\mathbf{w}$.

Line search

The learning rate in equation (15) determines the step size taken along the direction of the gradient in weight space. The size of η is decisive in assuring fast convergence of the learning algorithm and an ill-chosen learning rate can result in an oscillatory behavior, which reduces the effective rate of learning and can at worst cause the learning algorithm to diverge and never reach a minimum. A global value of the learning rate is often set through a method of trial and error, which is far from ideal since the optimal value varies with each iteration and is usually different for each weight in the network.

There are several methods for incorporating an adaptive learning rate into the learning algorithm. Line search computes an optimal learning rate for each iteration by minimizing the error with respect to η so that

$$\eta^{(t)} = \arg\min_{\eta} E(\mathbf{w}^{(t)} - \eta \nabla E|_{\mathbf{w}^{(t)}}). \tag{32}$$

This yields the value of η for which the error along a search direction equal to the gradient is minimized, as illustrated in Figure 9a. As a result, line search guarantees fast convergence whilst negating the oscillations that stem from a fixed learning rate that is too high. This is, however, achieved at the expense of a diminished capacity to escape local minima due to the error being strictly non-increasing.

Conjugate gradient method

Gradient descent in its simplest form is highly inefficient for realistic problems, where the error surface is usually non-quadratic and often quite complex. Even though line search mitigates difficulties related to the learning rate, the inefficiency due to suboptimal search directions remains. Consecutive search directions are orthogonal as a consequence of the line search optimization, i.e., the minimization of the error along the search direction $\mathbf{s}^{(t)}$ yields the updated weight vector $\mathbf{w}^{(t+1)}$, at which the gradient component parallel to $\mathbf{s}^{(t)}$ is zero and so the inner product

$$\mathbf{g}^{(t+1)\mathsf{T}}\mathbf{s}^{(t)} = 0, \tag{33}$$

where $\mathbf{g}^{(t+1)}$ is the gradient evaluated at $\mathbf{w}^{(t+1)}$. This is illustrated in the two-dimensional example in Figure 9a, where a fraction of the initial progress made in one direction is undone in the second iteration due to the orthogonality of search directions.

Equating the search direction with the negative of the gradient as done in gradient descent is thus suboptimal. A way of finding non-interfering, or conjugate, search directions is by imposing the conjugacy condition

$$\mathbf{s}^{(t+1)\mathsf{T}}\mathbf{A}\mathbf{s}^{(t)} = 0, \tag{34}$$

where $\mathbf{A}$ is a positive-definitive square matrix. This is a more general case of orthogonality, which reduces to the conventional notion when $\mathbf{A}$ equals the identity matrix. Under this condition, the search directions $\mathbf{s}$ are mutually conjugate with respect to $\mathbf{A}$ and thus linearly independent. Consequently, a set of W such vectors form the basis for the weight space, where W is the number of weights in the ANN. The difference between the optimal weight vector $\mathbf{w}^*$ and the initialization $\mathbf{w}^{(1)}$ can therefore be expressed as

$$\mathbf{w}^* - \mathbf{w}^{(1)} = \sum_{t=1}^{W} \alpha^{(t)} \mathbf{s}^{(t)}, \tag{35}$$

where $\alpha^{(t)}$ corresponds to the step size in each search direction and is equal to the learning rate $\eta^{(t)}$ in the case of conjugate gradients. The implication of equation (35) is that the optimal weights can be obtained in W iterations, which is valid for quadratic error surfaces. This is illustrated in Figure 9b, where $\mathbf{w}^*$ is obtained in two iterations.

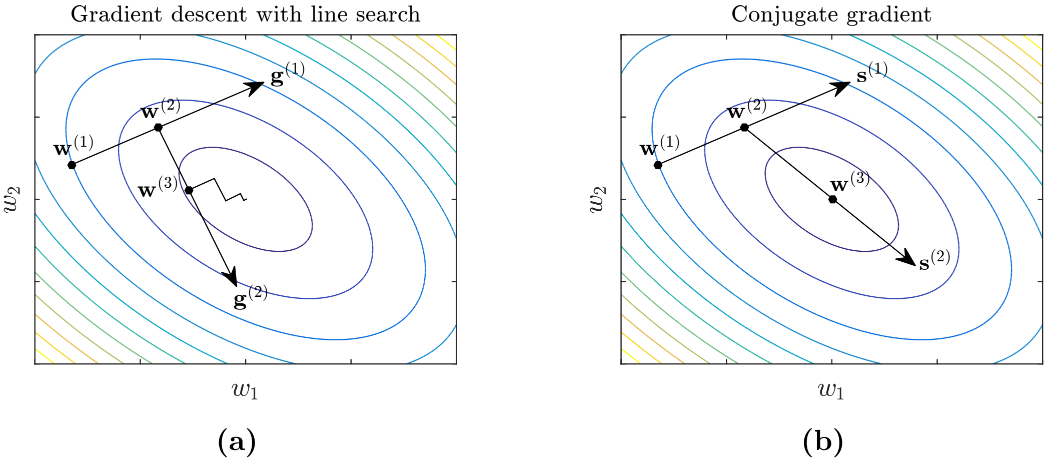


Figure 9: Contour plot demonstrating the method of line search and the difference between **(a)** gradient descent and **(b)** conjugate gradient on a two-dimensional problem.

Although the error surface is generally non-quadratic, it can be approximated locally with a quadratic function of the form

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\intercal \mathbf{H} \mathbf{w} + \mathbf{b}^\intercal \mathbf{w} + c, \tag{36}$$

where $\mathbf{H}$ is the Hessian matrix, $\mathbf{b}$ is a constant vector and c is a scalar. The search directions must be $\mathbf{H}$-conjugate, i.e., conjugate with respect to the matrix $\mathbf{H}$, and each search direction is a linear combination of the local gradient and the previous search direction as

$$\mathbf{s}^{(t)} = -\mathbf{g}^{(t)} + \beta^{(t)} \mathbf{s}^{(t-1)}, \tag{37}$$

with the exception of the first search direction which is set equal to the negative of the gradient. The scaling factor $\beta^{(t)}$ in equation (37) is determined explicitly by multiplying both sides with $\mathbf{s}^{(t-1)\intercal}\mathbf{H}$ and imposing the conjugacy condition in equation (34), which yields the expression

$$\beta^{(t)} = \frac{\mathbf{s}^{(t-1)\intercal} \mathbf{H} \mathbf{g}^{(t)}}{\mathbf{s}^{(t-1)\intercal} \mathbf{H} \mathbf{s}^{(t-1)}}. \tag{38}$$

Similarly, the explicit expression for the step size $\alpha^{(t)}$ is

$$\alpha^{(t)} = \frac{\mathbf{s}^{(t)\mathsf{T}}\mathbf{g}^{(t)}}{\mathbf{s}^{(t)\mathsf{T}}\mathbf{H}\mathbf{s}^{(t)}}. \tag{39}$$

The calculation of the scaling factor and step size using equation (38) and (39) requires knowledge of the Hessian matrix $\mathbf{H}$, which is computationally too expensive to reevaluate at each iteration. These parameters are therefore not explicitly determined but rather obtained by other means. There are for example several approximations for the scaling factor, of which the Polak-Ribière formula

$$\beta^{(t)} = \frac{\mathbf{g}^{(t)\mathsf{T}}(\mathbf{g}^{(t)} - \mathbf{g}^{(t-1)})}{\mathbf{g}^{(t-1)\mathsf{T}}\mathbf{g}^{(t-1)}}, \tag{40}$$

has proven to be most suitable for non-quadratic functions. The step size, however, is usually determined through a line search and the evaluation of both parameters depend therefore solely on gradient information obtained through the back-propagation algorithm.

The conjugate gradient algorithm can now be summarized as follows:

1. Initialize the weight vector $\mathbf{w}^{(1)}$.
2. Calculate the gradient $\mathbf{g}^{(t)}$ through the back-propagation algorithm.
3. If $t = 1$ or $t \bmod W = 0$, set the search direction $\mathbf{s}^{(t)} = -\mathbf{g}^{(t)}$,
   else calculate $\beta^{(t)}$ using the Polak-Ribière formula and set $\mathbf{s}^{(t)} = -\mathbf{g}^{(t)} + \beta^{(t)}\mathbf{s}^{(t-1)}$.
4. Determine the learning rate $\eta^{(t)}$ through a line search.
5. Update the weight vector to obtain $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta^{(t)}\mathbf{s}^{(t)}$.
6. Increment t by one and repeat from step 2 until the stopping criterion is met.

The modulus operator in step 3 functions as to restart the algorithm every Wth iteration by setting the search direction equal to the negative of the gradient. This is necessary because the implication of convergence in W iterations is no longer valid since the error surface is non-quadratic. Restarting the algorithm in this manner prevents the deterioration in conjugacy of search directions that would otherwise occur.

3.5 Generalization

During supervised learning, the model is presented a set of observations that is assumed to be representative of a larger population. The model is then trained to optimal performance on these observations, when in reality it is the ability to perform on unseen samples that is sought, i.e., to generalize beyond the training set. This is essentially an overfitting problem, which arises due to the complexity of the model. There are numerous methods that partly prevent this effect, e.g., regularization through weight decay, and network pruning methods that reduce the number of parameters and thereby also model complexity. The latter method is used in this study and it is therefore described next.

Optimal brain surgeon

The optimal brain surgeon (OBS) method is a network pruning algorithm, where weights that contribute the least to the overall performance of the model are removed in a sequential manner until a stopping criterion is met. Retraining after each iteration is not necessary, as is the case with its predecessor — the optimal brain damage algorithm, but instead an appropriate correction to the weight vector $\Delta\mathbf{w}$ is computed.

The calculation of both the optimal weight for deletion and subsequent correction vector require the evaluation of the inverse of the Hessian $\mathbf{H^{-1}}$, which, as mentioned previously, is computationally costly. However, the Hessian may be estimated using the Levenberg-Marquardt approximation, which for a cross-entropy error function and binary classification yields

$$\mathbf{H}_M \approx \sum_{m=1}^{M} h_m(1-h_m)\mathbf{g}_m\mathbf{g}_m^\intercal = \mathbf{H}_{M-1} + h_M(1-h_M)\mathbf{g}_M\mathbf{g}_M^\intercal, \tag{41}$$

where M is the number of training samples, and h_m and $\mathbf{g}_m$ are the model output and gradient of the mth sample, respectively. The inverse of the Hessian is then computed through a similar recursion

$$\mathbf{H}_M^{-1} = \mathbf{H}_{M-1}^{-1} - \frac{h_m(1-h_m)\mathbf{H}_{M-1}^{-1}\mathbf{g}_M\mathbf{g}_M^\intercal\mathbf{H}_{M-1}^{-1}}{1 + h_m(1-h_m)\mathbf{g}_M\mathbf{H}_{M-1}^{-1}\mathbf{g}_M^\intercal}, \tag{42}$$

with the initialization $\mathbf{H}_0^{-1} = \epsilon^{-1}\mathbf{I}$, where ϵ is a small number and $\mathbf{I}$ is the identity matrix.

The increase in error following the deletion of a specific weight is defined as the weight's saliency

$$S_n = \frac{w_n^2}{2\mathbf{H}_{n,n}^{-1}}, \tag{43}$$

where $\mathbf{H}_{n,n}^{-1}$ is the diagonal element of $\mathbf{H}^{-1}$ at position (n, n). With information regarding the relative importance of each weight, the least contributing weight can be removed and the subsequent correction to the remaining weights is

$$\Delta\mathbf{w} = -\frac{w_n}{\mathbf{H}_{n,n}^{-1}}\mathbf{H}^{-1}\mathbf{1}_n, \tag{44}$$

where $\mathbf{1}_n$ is a vector of size equal to the number of weights with value one in the nth element and zeros elsewhere. Equation (43) and (44) are valid under the assumptions that the network is properly trained to a global or local minimum and that the error function is approximately quadratic, i.e., the effect of a small perturbation can be approximated solely through the second derivative since the first derivative is zero at the minimum and higher-order terms of the Taylor expansion may be neglected. These two equations are essentially the result of an optimization of the quadratic term of the Taylor expansion with respect to perturbations to the weight vector, under the constraint that a perturbation corresponds to the complete removal of a single weight. The OBS algorithm is thus summarized in the following steps:

1. Train the ANN to minimum error.
2. Compute $\mathbf{H}_M^{-1}$ through the recursion in equation (42).
3. Find the weight n whose saliency S_n is smallest.
4. Calculate the correction $\Delta\mathbf{w}$ for weight n and update the entire weight vector $\mathbf{w}$.
5. Repeat from step 2 until the stopping criterion is met.

4 Material and Methods

4.1 Patient cohort

With approval from the Regional Ethics Board of Lund, Sweden (EPN Lund, Dnr 2013/742), patients were selected for retrospective analysis from a cohort used in the doctoral dissertation of Johanna Sjövall [33]. The primary focus of Sjövall's dissertation was on the management of regional lymph node metastases and consequently, all patients had lymph node involvement while distant metastases were present in none. Similarly, the primary aim of this study was on quantitative imaging of the nodal tumor, with little attention given to the primary tumor volume (GTV-T). This focus on nodal tumors was also motivated by the degree of dental artifacts within the GTV-T on computed tomography (CT) for the majority of all patients, and the fact that approximately half of all patients had their tonsils surgically removed.

The two main inclusion criteria were the retrievability of imaging and dose distribution data, and the completeness of clinical and follow-up information. Additionally, inclusion was narrowed further by only considering patients diagnosed with oropharyngeal cancer (OPC). In total, 82 cases were retrieved from the archive, of which 74 patients with OPC were retained for the positron emission tomography (PET) and CT part of the study, and 65 for the dose distribution part. All patients had histologically proven squamous cell carcinomas (SCC) and additional characteristics of all eligible patients are presented in Table 1.

Follow-up data was of at least 2-year maturity and contained information regarding site of recurrence, i.e., primary, nodal, or distant, and survival data. Locoregional control is in this study defined as the absence of recurrence in both the primary site and regional lymph nodes.

Table 1: Characteristics of the 74 patients eligible for analysis.

Patient characteristics	No.	(%)
Median age (range)	60	(44–87)
Sex		
Male	57	(77)
Female	17	(23)
Primary site		
Tonsil	60	(81)
Base of tongue	10	(14)
Other	4	(5)
HPV		
HPV-positive	62	(84)
HPV-negative	12	(16)
Tumor stage		
T1	13	(18)
T2	41	(55)
T3	12	(16)
T4	8	(11)
Nodal stage		
N1	11	(15)
N2	62	(84)
N3	1	(1)
Histologic grade		
Well differentiated	3	(8)
Moderately differentiated	18	(64)
Poorly differentiated	47	(24)
Not specified	6	(4)
Treatment		
Radiotherapy alone	69	(93)
Concurrent chemotherapy	3	(4)
Neoadjuvant chemotherapy	2	(3)
Tumor recurrence		
Primary site	13	(18)
Regional lymph node	9	(12)
Distant metastasis	4	(5)
Disease-specific mortality	10	(14)

HPV: Human papillomavirus

Image acquisition and treatment delivery

All patients underwent pre-treatment ^{18}F-fluorodeoxyglucose (^{18}F-FDG) PET/CT on a Philips Gemini TF PET/CT scanner and adhered to the same imaging protocol. Intravenous injection of 4 MBq/kg body weight ^{18}F-FDG was followed by an incubation time of 1 hour (range, 51–99 minutes) and images were acquired for 2 minutes per bed position from the upper abdomen to the base of the skull. Images were reconstructed to $4 \times 4 \times 4$ mm voxel dimensions and attenuation and scatter corrections were applied.

Non-contrast enhanced full-dose CT scans were taken for the purpose of treatment planing using the same scanner. The CT images had an in-plane resolution of 1.1719×1.1719 mm and a reconstructed slice thickness of 3 mm.

The prescribed radiation dose was 68 Gy delivered in fractions of 2 Gy through intensity-modulated radiation therapy (IMRT) at Skåne University Hospital in Lund, Sweden. The large majority (n = 69) were treated with definitive radiation therapy alone and the remaining five patients received either concurrent or neoadjuvant chemotherapy.

4.2 Feature extraction

Three different groups of features were calculated on the retained data set, namely: shape features, texture features, and standardized uptake value (SUV) parameters. Whereas the last is defined only for PET, the other two feature groups are computable for both modalities. Texture features were calculated in an attempt to probe heterogeneity in ^{18}F-FDG uptake and nodal tumor density, respectively. Additionally, the gray-level size zone matrix (GLSZM), described in an earlier chapter, was utilized to calculate five custom feature on the dose distribution for the purpose of investigating the association between coldspots and hotspots, and locoregional tumor control, i.e. whether the size of underdosed and overdosed regions is critical for locoregional tumor control.

PET/CT features were extracted for the gross tumor volume of the nodal tumor (GTV-N) and the custom GLSZM-based features from the GTV-T, which were delineated by radiation oncologists as is routine prior to external radiation therapy. All features were computed in three-dimensions, i.e., the gray-level co-occurrence matrix (GLCM) and gray-level run length matrix (GLRLM) were averaged over 13 directions, and a distance of 1 voxel was used for the GLCM and the neighborhood gray-tone difference matrix (NGTDM).

The imaging and dose data were transferred into Matlab R2015b, where all features were computed through modules contained in the RADIOMICS package available at [34]. A few adjustments were, however, introduced to tailor for use in PET/CT rather than PET and magnetic resonance imaging (MRI), for which the package was originally developed and used by Vallières et al. [35]. An overview of the entire procedure is given in Figure 11.

Shape

A total of four shape features were extracted for the description of both size and shape of the volume of interest (VOI). These geometrical descriptors were computed only on CT scans because of the superior spatial resolution of CT when compared to PET, which may allow for a more accurate estimation. Examples for two of the features are given in Figure 10 and the four shape features were:

1. Volume — obtained from the number of voxels contained in the VOI and the voxel dimensions.
2. Size — defined as the longest diameter of the VOI, as computed in three-dimensions..
3. Solidity — defined in two-dimensions as the ROI area divided by the area of the convex hull and is an overall measure of concavity. This is extended into three-dimensions by taking the ratio of volumes, rather than areas.
4. Eccentricity — is in two-dimensions a measure of circularity defined as the ratio between the longest chord l_c and the longest perpendicular chord l_{p1}. As with the solidity, this feature extends easily into three-dimensions where an additional perpendicular chord l_{p2} is introduced and the definition changes accordingly to

$$\text{Eccentricity} = \sqrt{1 - \frac{l_{p1} \cdot l_{p2}}{l_c^2}}. \tag{45}$$

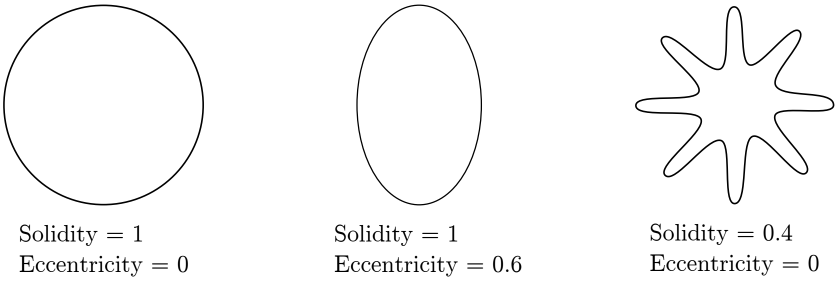


Figure 10: The solidity and eccentricity for three different shapes.

Texture

To allow for a meaningful three-dimensional texture analysis, the image volume was isotropically resampled to the in-plane resolution prior to feature extraction. First-order histogram features were, however, computed on the original image volume since the spatial relation between voxels was irrelevant.

Another prerequisite for the calculation of the second- and higher-order statistics is the quantization into a manageable number of gray-levels. Each VOI was uniformly quantized into 32 gray-levels prior to feature extraction and the reason for this was twofold. The number of quantization-levels affects the sparsity of the texture matrices, where denser matrices are more desirable from a statistical point of view, and are generally the result of fewer quantization-levels. Furthermore, variations in image intensity are reduced as a consequence of this quantization procedure and there exists a fine line where the gain in information, as a result of noise dampening, is offset by a loss of information concerning actual tumor heterogeneity. The latter effect was not examined and the appropriate number of quantization-levels was determined mainly with the first point in mind and with regard to previous findings [36]. The type of quantization algorithm, however, was chosen as to cause the least possible distortion of raw imaging data [37]; although the Lloyd-Max quantization algorithm was more preferable to counteract the large range in Hounsfield units (HU) introduced by the presence of air and/or bone in the VOI, which negatively impact the effective range for soft tissue analysis. The effect of air and bone in the VOI was partly eliminated through two fixed thresholds.

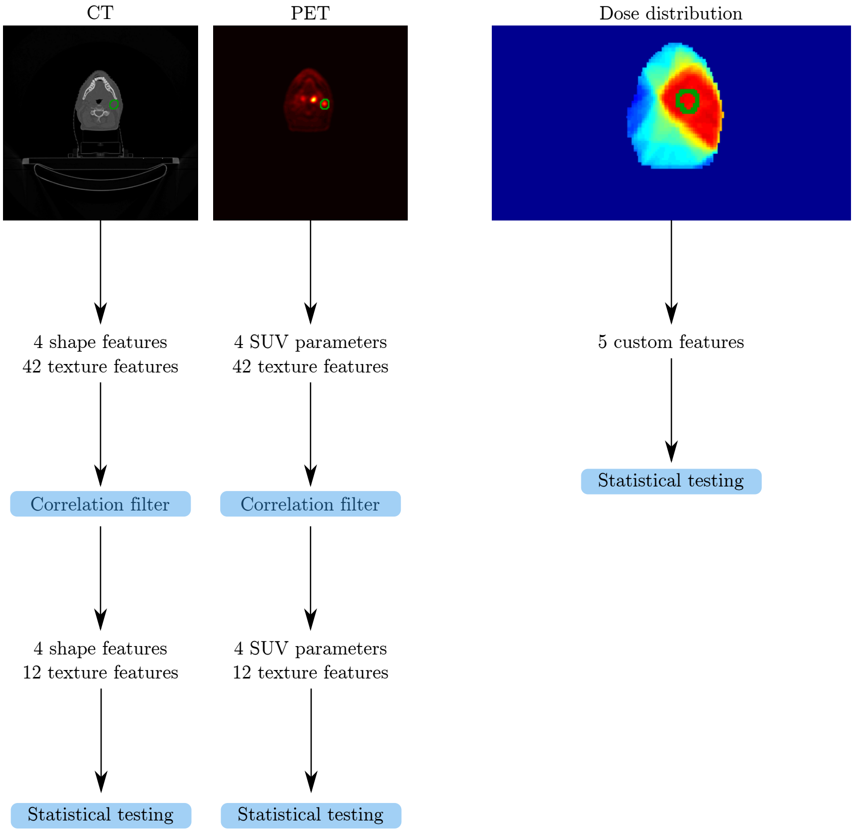


Figure 11: Overview of the extraction and subsequent management of radiomic features from image and dose distribution data. In PET and CT analysis, features are extracted from the GTV-N, whereas for the dose distribution the GTV-T is analyzed.

Following these preliminary steps, the following texture features were calculated: the variance, skewness, and kurtosis obtained from the first-order histogram; 8 GLCM-based features; 13 GLRLM-based features; 13 GLSZM-based features, and 5 NGTDM-based features, resulting in a total of 42 texture features per image modality. The definitions of second- and higher-order statistical features are given in Appendix A.

The five bespoke features extracted from the normalized dose distribution were defined as

$$\text{Dose deviation emphasis (DDE)} = \frac{1}{S}\sum_{i=1}^{L}\sum_{z=1}^{S} P(i,z)|i-d|, \tag{46}$$

$$\text{Small coldspot emphasis (SCE)} = \frac{1}{S}\sum_{i=1}^{d}\sum_{z=1}^{S} P(i,z)|i-d|(S-z), \tag{47}$$

$$\text{Small hotspot emphasis (SHE)} = \frac{1}{S}\sum_{i=d}^{L}\sum_{z=1}^{S} P(i,z)|i-d|(S-z), \tag{48}$$

$$\text{Large coldspot emphasis (LCE)} = \frac{1}{S}\sum_{i=1}^{d}\sum_{z=1}^{S} P(i,z)|i-d|z, \tag{49}$$

$$\text{Large hotspot emphasis (LHE)} = \frac{1}{S}\sum_{i=d}^{L}\sum_{z=1}^{S} P(i,z)|i-d|z, \tag{50}$$

where $P(i,z)$ is the normalized GLSZM containing the probabilities of observing regions of size z and intensity i, and L is the number of quantization-levels. The GLSZM is further normalized with respect to S, which is the total number of voxels enclosed in the VOI. In contrast to PET/CT features, the normalized dose distribution was quantized into 19 levels with predefined decision boundaries as to assure a valid comparison between distributions. Relative doses between 99.5% and 100.5% were rebinned to the midmost level $d = 10$ and the remaining dose values were handled accordingly with the same dose increment of 1%. The resulting GLSZM and the different emphases are illustrated in Figure 12.

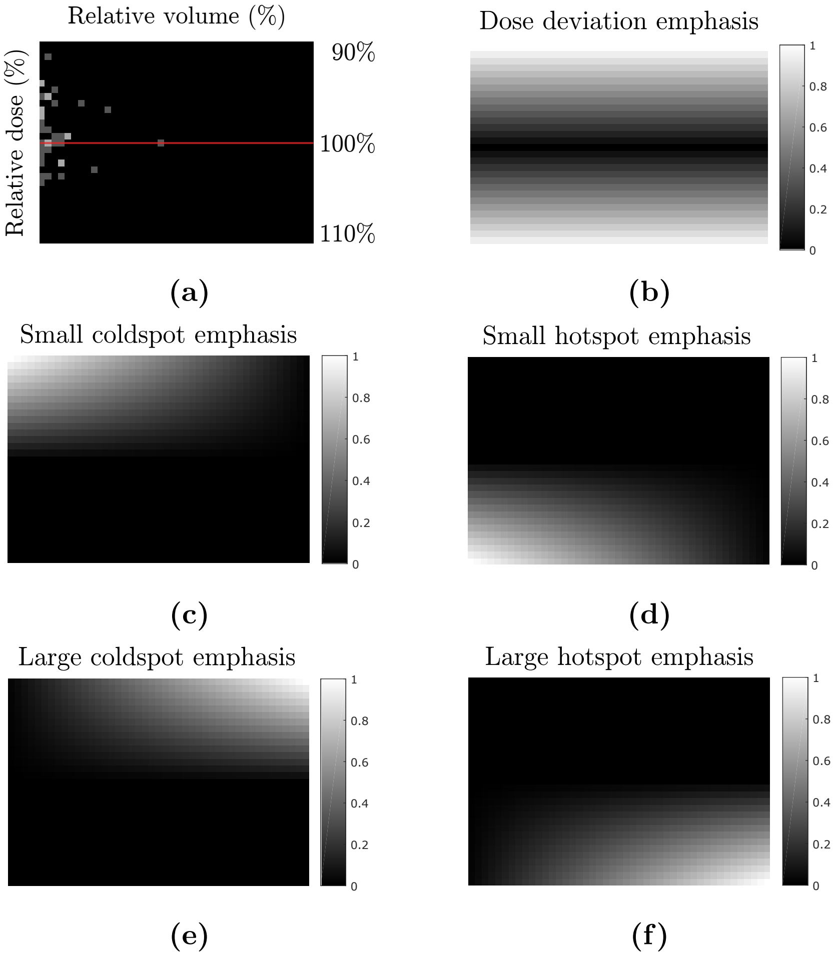


Figure 12: (a) Example of a customized GLSZM and its properties, and **(b)** – **(f)** the different emphases visualized, where each focuses on a different aspect of the GLSZM.

Standardized uptake value

Attenuation and scatter corrected PET images were converted to SUV maps through information contained in the DICOM header, thus enabling the calculation of three common SUV parameters. These three parameters, as normalized to body weight, were the $\mathrm{SUV}_{\mathrm{mean}}$ and $\mathrm{SUV}_{\mathrm{max}}$ defined as the average and highest SUV within the VOI, respectively, and the $\mathrm{SUV}_{\mathrm{peak}}$ defined as the average between the highest SUV and its 26 immediately adjacent voxel values. Additionally, the area under the cumulative SUV histogram (AUC-CSH) was calculated since it has been shown to be a measure of the degree of heterogeneity [38]. The cumulative SUV histogram is, in construction, quite similar to the dose-volume histogram (DVH) commonly used in radiation therapy.

4.3 Model training

The model used for prediction purposes was a multi-layer perceptron (MLP) consisting of a single hidden layer containing three nodes, and a single output unit for binary classification. Sigmoidal activation functions were used for all nodes and the cross-entropy error was adopted as a performance measure. The neural network was set up and trained through code written locally in Matlab R2015b.

The relative simplicity of the neural network was due to the limited number of observations; using leave-one-out cross validation, the number of cases available for model training was 73, which, according to general guidelines, allows for approximately 10 parameters [39]. The relatively small network in Figure 13 contains in total 16 weights. To counter overfitting, the optimal brain surgeon (OBS) algorithm was implemented and a reduction in the number of parameters was thereby achieved. The stopping criterion was heuristically set to a 30% increase of the post-training cross-entropy error, i.e., the removal of parameters was discontinued once the cross-entropy error post-deletion was to exceed 130% of the initial training error.

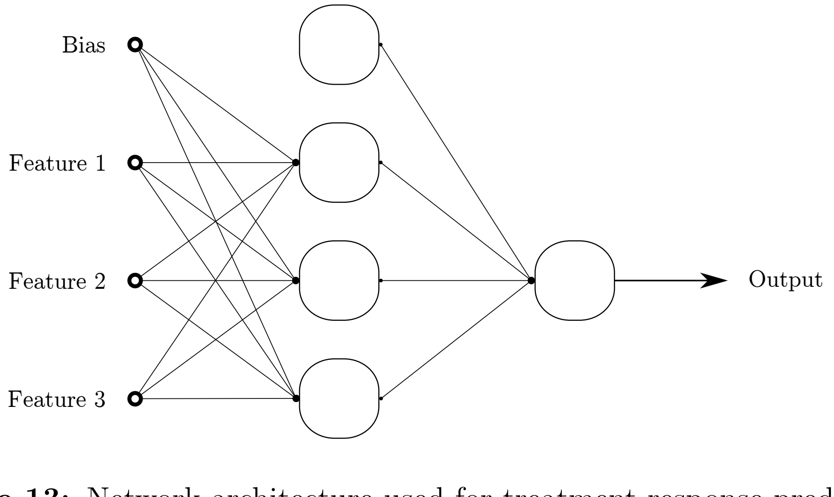


Figure 13: Network architecture used for treatment response prediction.

Training was performed using the entire training set with the back-propagation algorithm alongside the conjugate gradient method and line searches. All features were normalized and rescaled to zero mean and unit variance, and the network parameters were initialized from a random uniform distribution with zero mean and a small variance, which is recommended for rapid convergence [32]. The stochastic aspect of such initialization was handled through multiple repetitions of the entire training procedure, except for the OBS algorithm, which was implemented only on the final model. As a result, the probability of convergence on a shallow local minimum was reduced at the expense of a manageable increase in computation time.

4.4 Statistical analysis

Correlation-based feature subset reduction

The extracted features are highly correlated by definition, especially those derived from the same statistical texture matrix. This redundancy was first reduced by selecting the three least correlated texture features from each matrix, e.g., the 8 GLCM-based features were reduced to a set of 3, among which the correlation was minimized. Pearson's r was used as a measure of correlation. It was thus hypothesized that the information contained within each texture matrix is limited and can adequately be expressed by the three least correlated features. The NGTDM-based features underwent a similar reduction, and the GLRLM- and GLSZM-based features were combined such that 3 features were selected from the combined set of 26 features. These two groups were merged prior to reduction because they consistently exhibited high intergroup correlation, which was also observed in previous studies [17, 36]. The remaining groups, namely, the shape and SUV parameters, were not affected by this procedure, although they too exhibit high intragroup correlation.

Multiple hypotheses testing

Following this initial feature subset reduction, multiple hypotheses were tested using the non-parametric Mann-Whitney U test since most image features were not normally distributed. The hypotheses tested were whether image features differed between HPV-positive and HPV-negative patients, and between responders and non-responders, where the latter was tested for both locoregional failure and disease-specific mortality independently. Correction for multiple testing was performed by control of the false discovery rate (FDR), which was restricted to 5% using the Benjamini-Hochberg method. Additionally, the hypothesis that clinical information, e.g., TNM-classification, HPV- and smoking-status, age, weight, and sex, could differentiate between responders and non-responders was tested through multiple comparisons using either the Mann-Whitney U test, Fisher's exact test, or the χ^2 test for trends, depending on the type of variable. The custom GLSZM-based features were tested in similar manner against treatment response.

Feature selection for prediction

For response prediction using the model architecture presented in Figure 13, a set of three features were selected from the entire set using the correlation based method described earlier followed by the minimum redundancy maximum relevance (MRMR) algorithm and multivariable analysis. The MRMR algorithm is an entropy-based feature selection method commonly used for gene expression data, where a feature set of predefined size is selected such that the mutual information between predictors is minimized and information regarding the dependent variable is maximized [40]. In total, the initial PET/CT feature set contained 92 features, which was reduced to 32 by removal of highly correlated texture features. Of the remaining 32 features, a set of 10 was obtained through the MRMR algorithm and the final three features were determined through sequential forward selection.

Model evaluation

The entire feature subset selection methodology was incorporated into the leave-one-out cross validation procedure and the holdout sample was thus withheld from feature selection and subsequent model training. This was done to counter the large bias that would otherwise occur by inclusion of the holdout sample into the feature selection algorithm [41]. The predictive performance of the model was evaluated using the area under the receiver operating characteristic curve (AUC-ROC) and 95% confidence intervals were generated with a bootstrap method and threshold averaging. Statistical analysis was performed in both Matlab R2015b and the R software version 3.2.3.

5 Results

The prognostic value of several clinical variables is recounted in Table 2, where the clinical endpoints are primary tumor recurrence and disease-specific mortality. Locoregional control is not presented in Table 2 since none of the variables are significant after false discover rate (FDR) correction, and only T-stage ($p = 0.043$) and weight ($p = 0.033$) are significant prior to correction for multiple testing.

Table 2: Association between clinical information, and primary tumor recurrence and disease-related mortality. Number of patients and percentages reported for all variables except for age and weight, where the median is given in years and kilogram, respectively.

	Primary tumor recurrence			Disease-specific mortality		
Characteristic	Non-resp.	Resp.	p	Non-resp.	Resp.	p
Age	61	60	0.72	60	61	0.89
Weight	76	79	0.54	63	81	0.00089**
Sex			0.99			0.044*
Male	10 (77%)	47 (77%)		5 (50%)	52 (81%)	
Female	3 (23%)	14 (23%)		5 (50%)	12 (19%)	
Smoker			0.0067**			0.71
Yes	13 (100%)	38 (62%)		8 (80%)	43 (67%)	
Never	0 (0%)	23 (38%)		2 (20%)	21 (33%)	
HPV			0.43			0.00076**
Positive	10 (77%)	52 (85%)		4 (40%)	58 (91%)	
Negative	3 (23%)	9 (15%)		6 (60%)	6 (9%)	
T-stage			0.0084**			0.0055**
T1	1 (8%)	12 (20%)		1 (10%)	12 (19%)	
T2	4 (31%)	37 (61%)		3 (30%)	38 (59%)	
T3	5 (38%)	7 (11%)		2 (20%)	10 (16%)	
T4	3 (23%)	5 (8%)		4 (40%)	4 (6%)	
N-stage			0.85			0.75
N1	2 (15%)	9 (15%)		1 (10%)	10 (16%)	
N2	11 (85%)	51 (84%)		9 (90%)	53 (83%)	
N3	0 (0%)	1 (1%)		0 (0%)	1 (1%)	
Histologic grade			0.75			0.15
G3	0 (0%)	3 (6%)		1 (13%)	2 (3%)	
G2	4 (31%)	14 (25%)		3 (37%)	15 (25%)	
G1	9 (69%)	38 (69%)		4 (50%)	43 (72%)	
Chemotherapy			0.21			0.016**
Yes	2 (15%)	3 (5%)		3 (30%)	2 (3%)	
No	11 (85%)	58 (95%)		7 (70%)	62 (97%)	

*Significant at $p < 0.05$. **Significant at a false discovery rate of 5%.

Histologic grade, G3: well differentiated; G2: moderately differentiated; G1: poorly differentiated.

Table 3: Median and p-values of PET/CT features for the three grouping variables HPV-status, locoregional control, and disease-related mortality, in nodal oropharyngeal cancer. The volume and size are given in mm^3 and mm, respectively.

	HPV			Locoregional control			Disease-specific mortality		
Feature	Pos.	Neg.	p	Non-resp.	Resp.	p	Non-resp.	Resp.	p
Shape									
Volume	16000	9400	0.029*	16000	14000	0.89	16000	14000	0.86
Size	48	45	0.24	45	48	0.39	45	48	0.41
Solidity	0.90	0.91	0.94	0.92	0.89	0.42	0.92	0.89	0.56
Eccentricity	0.79	0.80	0.41	0.78	0.80	0.16	0.72	0.82	0.039*
SUV									
SUV$_{max}$	8.3	6.5	0.041*	7.3	8.4	0.15	7.4	8.2	0.41
SUV$_{peak}$	6.3	4.5	0.019*	5.4	6.3	0.12	5.1	6.2	0.27
SUV$_{mean}$	3.7	2.6	0.015*	3.2	3.6	0.37	2.8	3.6	0.27
AUC-CSH	0.38	0.38	0.46	0.40	0.38	0.16	0.38	0.38	0.53
CT texture									
Global									
Variance	38	25	0.59	19	40	0.0033*	15	39	0.0012**
Skewness	-1.7	-0.92	0.0060*	-1.0	-1.7	0.014*	-0.46	-1.7	0.0027**
Kurtosis	7.6	4.2	0.075	7.5	6.6	0.23	17	6.8	0.083
GLCM									
Correlation	0.64	0.69	0.40	0.61	0.65	0.36	0.60	0.65	0.28
Energy	0.038	0.018	0.077	0.046	0.033	0.49	0.070	0.033	0.16
Contrast	5.6	11	0.082	5.4	6.3	0.45	3.0	6.3	0.086
GLRLM and GLSZM									
SRLGE†	0.0033	0.0049	0.093	0.0044	0.0033	0.22	0.0057	0.0033	0.024*
GLN†	13000	4700	0.023*	14000	12000	0.88	16000	12000	0.56
ZSN‡	270	220	0.48	210	270	0.16	170	290	0.015*
NGTDM									
Coarseness	0.00095	0.0016	0.033*	0.00088	0.0011	0.97	0.00088	0.0011	0.79
Complexity	510	610	0.19	390	540	0.31	370	540	0.041*
Busyness	1.4	0.61	0.12	1.5	1.1	0.74	2.2	0.98	0.29
PET texture									
Global									
Variance	12	7.4	0.056	12	10.1	0.67	11	11	0.93
Skewness	0.55	0.60	0.38	0.34	0.60	0.11	0.42	0.58	0.65
Kurtosis	-0.57	-0.36	0.42	-0.67	-0.50	0.34	-0.43	-0.57	0.54
GLCM									
Correlation	0.59	0.56	0.035*	0.53	0.59	0.0011**	0.53	0.59	0.095
Entropy	8.7	8.4	0.12	8.8	8.7	0.40	8.6	8.7	0.83
Sum average	0.014	0.013	0.63	0.014	0.013	0.24	0.014	0.014	0.86
GLRLM and GLSZM									
LZLGE‡	0.24	0.18	0.87	0.14	0.27	0.10	0.16	0.24	0.48
RLN†	2700	1600	0.033*	2800	2500	0.99	2800	2500	0.77
LRHGE†	240	230	0.40	280	230	0.28	250	240	0.79
NGTDM									
Busyness	0.11	0.11	0.42	0.10	0.11	0.78	0.10	0.11	0.76
Contrast	0.30	0.45	0.11	0.30	0.320	0.58	0.26	0.33	0.32
Strength	14	24	0.023*	12	15	0.44	14	15	0.69

*Significant at $p < 0.05$. **Significant at a false discovery rate of 5%.

†GLRLM- and ‡GLSZM-based features.

Full definition of acronyms given in Appendix A.

Tumor stage is as anticipated strongly prognostic for all endpoints, where higher stages are associated with adverse outcomes. Similarly, human papillomavirus (HPV) positivity is associated with a favorable outcome with respect to survival, although this is not reflected by primary tumor control. In contrast, tobacco consumption appears to be strongly detrimental with regard to primary tumor control, where all 13 non-responders were either current or former smokers. This is, however, not apparent in the disease-specific mortality, although a non-significant trend is demonstrated. Additionally, patient sex and the prescription of chemotherapy were both significant at $p < 0.05$. Possible interactions between variables are, however, not examined even though they are almost certain to exist.

The association between radiomics features and HPV-status, locoregional control, and disease-specific mortality, respectively, is presented in Table 3. Several features are significant for HPV-status at $p < 0.05$ but none remain after FDR-correction. Interestingly however, the three standardized uptake value (SUV) parameters, namely, $\mathrm{SUV}_{\mathrm{max}}$, $\mathrm{SUV}_{\mathrm{peak}}$, and $\mathrm{SUV}_{\mathrm{mean}}$, are all found to significantly differ between HPV-positive and HPV-negative oropharyngeal cancer (OPC), where a higher SUV is observed in HPV-positive lymph nodes.

For locoregional control, both the variance and the skewness of the first-order histogram are significant at $p < 0.05$. Moreover, they remain significant after FDR-correction for disease-specific mortality as endpoint. This suggests that the distribution of Hounsfield units (HU) is generally more skewed towards lower densities for responders than for non-responders.

Of the second- and higher-order texture features, only the correlation as calculated from the gray-level co-occurrence matrix (GLCM) on positron emission tomography (PET) is significant post-correction. Correlation is a measure of dependence between pairs of pixels at a specified distance and a lower correlation on PET signifies a more irregular spatial distribution of radiotracer uptake, which is observed for non-responders.

Table 4: Median of the custom GLSZM features extracted from the dose distribution within the GTV-T, and their association with primary tumor recurrence and locoregional control.

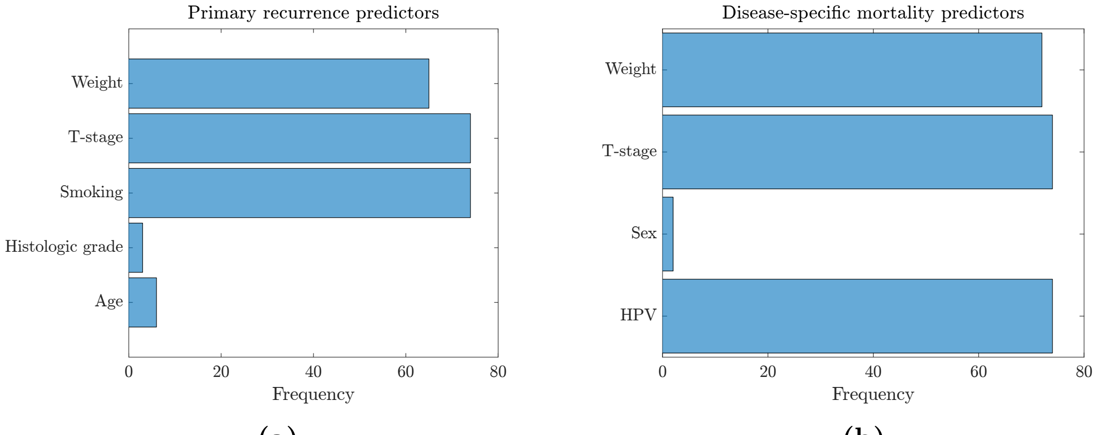
	Primary tumor recurrence			Locoregional control		
Parameter	Non-resp.	Resp.	p	Non-resp.	Resp.	p
DDE	0.22	0.19	0.032*	0.22	0.19	0.022*
LCE	0.0019	0.0022	0.35	0.0026	0.0022	0.80
SCE	0.11	0.094	0.70	0.11	0.094	0.56
LHE	0.0019	0.0022	0.70	0.0026	0.0022	0.89
SHE	0.12	0.10	0.11	0.11	0.093	0.19

*Significant at $p < 0.05$.
DDE: Dose deviation emphasis.
LCE: Large coldspot emphasis, SCE: Small coldspot emphasis.
LHE: Large hotspot emphasis, SHE: Small hotspot emphasis.

Results for texture analysis of the dose distribution within the gross tumor volume of the primary tumor (GTV-T) are presented in Table 4. Responders and non-responders did not differ with respect to mean absorbed dose for either primary recurrence ($p = 0.83$) or locoregional control ($p = 0.94$); neither did they with regard to two common dose-volume histogram (DVH) parameters: $V_{95\%}$ ($p = 0.96$ and $p = 0.51$) and $V_{105\%}$ ($p = 0.17$ and $p = 0.36$). Similarly, no significant difference is observed in the sizes of coldspots and hotspots between responders and non-responders. However, the overall deviation from the mean is found to be significant for both primary and locoregional recurrence. Note that this deviation considers connectivity of regions and is thus not comparable to the variance extracted from first-order histograms.

Predictive performance of PET and computed tomography (CT) features is subpar, with best performance in prediction of disease-specific mortality, where the area under the receiver operating characteristic curve (AUC-ROC) is 0.66 (95% confidence interval, 0.43–0.85). For primary recurrence as endpoint, the performance is essentially no better than chance with an AUC-ROC of 0.50 (95% confidence interval, 0.28–0.71).

In contrast, clinical information is highly prognostic for primary tumor recurrence and disease-specific mortality with an AUC-ROC of 0.87 (95% confidence interval, 0.73–0.96) and 0.73 (95% confidence interval, 0.52–0.87), respectively. Due to the leave-one-out cross-validation procedure, the final three features selected as predictors varied between different iterations. In Figure 14, however, it is apparent that the features selected are predominantly those previously declared significant through univariable analysis. The three most selected variables for each endpoint are thus likely not correlative and of value for general model performance. The additional contribution of image-based features to the already well-performing clinical variables is found non-significant for prediction of primary recurrence, whereas three features contributed significantly for disease-specific mortality, namely: the contrast based on the neighborhood gray-tone difference matrix ($p = 0.00060$), the eccentricity ($p = 0.0077$), and the variance based on the gray-level co-occurrence matrix ($p = 0.019$).



(a) **(b)**

Figure 14: The number of times each variable is used as a predictor for **(a)** primary tumor recurrence and **(b)** disease-specific mortality. The total number of iterations is 74, which is equal to the number of observations, and three features are selected in each iteration.

6 Discussion

Routine radiographic images contain more information than is readily utilized in clinical practice. Radiomics refers to the automatic extraction of a large number of image-based features from routine scans and for the purpose of generating additional information regarding the tumor phenotype. Successful implementation thereof would be valuable for patient health since non-responders could be identified before treatment is even begun and alternative treatment strategies could be explored.

The feasibility of the radiomics approach in oropharyngeal squamous cell carcinoma (SCC) was the main focus of the present thesis, where a relatively large number of features were extracted from positron emission tomography (PET) and computed tomography (CT) scans. The prognostic value of these features was assessed and results appear encouraging. Several features are found to significantly differ between responders and non-responders, which in a controlled setting may allow for accurate risk estimation.

The majority of the promising features were in fact derived from CT scans and thus represent tumor morphology. This is quite contradictory seeing as tumor metabolism is generally considered more indicative of disease progression. Possible explanations might relate to the instability of quantitative PET imaging due to multiple confounding factors [8]. Additionally, the poor resolution of PET, as compared to CT, might further diminish the usefulness of texture features, which comprise the majority of image-based features. These shortcomings notwithstanding, a second-order statistics based on the pair-wise spatial distribution of voxels in PET was found highly significant for locoregional control. This feature — the correlation — suggests that the spatial distribution of ^{18}F-fluorodeoxyglucose (^{18}F-FDG) in non-responders is more irregular on a voxel-to-voxel basis. However, the reliability of voxel-scale measures in PET is questionable due to the partial volume effect and additional factors relating to the image reconstruction. Texture features are generally quite sensitive with respect to several factors, which need to be further examined.

In contrast, features extracted from the first-order histogram do not depend on as many factors and are possibly more robust as a consequence. Two such features, namely, the global variance and skewness as computed on CT, differed significantly between responders and non-responders for both clinical endpoints. The implication is that responders exhibit a larger variance and more negative skewness in Hounsfield units (HU) than non-responders. What this actually signifies is, however, unclear; although common first-order histogram metrics have previously been associated with hypoxia and angiogenesis [9].

Of the shape features, the eccentricity was significantly larger among responders, corresponding to more elongated lymph nodes. This is consistent with previous observations, where lymph nodes of the neck were evaluated using a roundness index and results identify roundness as a fair measure of the degree of malignancy [42].

Interestingly, none of the standardized uptake value (SUV) parameters demonstrated prognostic value in univariable analysis; all three were, however, significant with respect to human papillomavirus (HPV) status, indicating a higher ^{18}F-FDG uptake in HPV-positive tumors. Furthermore, of the three texture features reported to significantly differ between HPV-positive and HPV-negative oropharyngeal cancer (OPC) by Fujita et al. [6], two were examined in the present work, and although non-significant, the contrast as extracted from the gray-level co-occurrence matrix (GLCM) demonstrated a weak trend

($p = 0.082$). Methodological differences are, however, likely to exist seeing as the range of values differ markedly and clear definitions of the texture features are not provided. This is in fact a common trend among publications examining the feasibility of texture analysis on medical images, which further prevent meaningful comparisons across studies and delay much needed evaluation of reproducibility [36]. We are unfortunately guilty of contributing to this trend seeing as all features were extracted from the gross tumor volume of the nodal tumor (GTV-N), whereas all previous work focused on the primary tumor; hence, direct comparisons might be invalid.

Another limitation lies in the dichotomization of outcomes and the choice of statistical tests. Disease outcome should ideally be analyzed taking the time factor into account. We opted against doing so in the present work partially because of the difficulties that it would entail in the implementation of the artificial neural network.

Texture analysis of the dose distribution

Second- and higher-order statistics were employed as measures of intratumoral heterogeneity in PET/CT since the first-order histogram was inadequate at describing the spatial relationship between voxels. The dose-volume histogram (DVH) commonly used in the evaluation of radiation treatment plans suffers from the same shortcoming, i.e., the spatial relationship between voxels of low and high dose, respectively, is unaccounted for.

Differences in treatment plans between responders and non-responders were investigated through five custom features derived from the gray-level size zone matrix (GLSZM). There was no noticeable difference in the relative sizes of coldspots and hotspots, respectively, for examined patients, likely due to the fulfillment of general dose criteria, and possibly due to ill-defined features. What constitutes a large coldspot and hotspot, respectively, was chosen as the entire tumor volume, which is by definition impossible. More appropriate features definitions would perhaps emphasize the 5–10% relative volume as large. Moreover, the degree of emphasis is also subject to questioning; the linear emphasis employed in the present thesis is perhaps not sufficient in revealing small disparities between responders and non-responders, whereas a quadratic relation might be. The large majority of texture features do in fact scale quadratically.

The dose deviation emphasis (DDE), which was defined as to put no emphasis on differences in sizes and merely measure the deviation from the mean dose, was found to significantly differ between responders and non-responders. A larger deviation is observed among non-responders, indicating a higher number of disconnected regions receiving either a lower or higher radiation dose. This deviation is not captured by the variance extracted from the first-order histogram, which is a measure of something slightly different. Whereas the first-order histogram contains frequencies of different voxel values, each row of the GLSZM contains the number of times a contiguous region of a specific intensity is observed. The DDE is thus a measure of deviation in the number of connected regions rather than the number of voxels. This might be capable of capturing differences in the number of small low-dose regions at the margins of the volume of interest (VOI), where they usually occur.

Another consideration regards the selection of VOI used for analysis. We have chosen to analyze the gross tumor volume of the primary tumor (GTV-T) as it is usually the main focus of therapy. This is, however, inappropriate in the present study seeing as half of all patients had their primary sites surgically removed and are thus presented with a lower tumor burden. Consequently, a higher frequency of primary recurrences is observed among patients with primary tumors intact, although the difference appears non-significant ($p = 0.20$).

In future studies, one would preferably examine either the clinical or planning target volume, where the variability in absorbed dose is usually higher. Texture analysis of dose distributions has to our knowledge not been attempted before and although limited in several ways, this study demonstrates a proof of concept which could potentially be of value once all shortcomings are addressed.

Treatment response prediction

The radiomics framework enables the use of advanced machine learning methods, which are generally unfeasible in medical applications due to the limited number of observations available. These methods are often developed for use with big data and their high complexity is thus upheld by an enormous amount of observations. Quantitative analysis of routine scans would allow for structuring of sufficiently large data sets for machine learning applications. The patient material used in the present thesis is not nearly sufficient for such methods and the employed artificial neural network was therefore severely restricted due to apparent limitations. Model performance is essentially comparable to logistic regression due to the relatively simple network architecture, and whether added difficulties relating to network training are worthwhile is questionable. Realization of the radiomics framework would, however, likely involve advanced methods and the implementation thereof in this work was mainly for an educational purpose. Nevertheless, the performance of a neural network is rarely worse than logistic regression and it is therefore still pertinent for the current task [28].

Model choice disregarded, clinical information appeared most prognostic and yielded excellent performance as indicated by a large area under the receiver operating characteristic curve (AUC-ROC), especially in predicting primary tumor recurrence where an AUC-ROC of 0.87 (95% confidence interval, 0.73–0.96) was achieved. Similar results are reported by Bryce et al., where clinical data was utilized in the prediction of survival in squamous cell carcinoma of the head and neck through artificial neural networks [43]. Information regarding HPV- and smoking-status, and weight was unavailable for their patients and different variables were utilized instead, e.g., nodal stage and hemoglobin. This suggests that there is additional information to be gained from other clinical variables and it is possible that clinical information alone might suffice for response prediction.

Image-based features performed poorly for both endpoints, especially for recurrence prediction, which might be due to a disadvantageous model architecture. There was in general fewer image-based features of prognostic value, as concluded by the number of features that were significant at a false discovery rate of 5%. The requirement of including three features, when there is perhaps only two non-correlated predictors, was therefore likely detrimental for model performance seeing as the third feature added no real value but instead increased the general uncertainty of the model. Nevertheless, three image-based features significantly contributed to a model already containing three clinical variables, which implies that radiomics features might possess additional information regarding tumor phenotype. Two of these features were, however, disregarded by the feature selection procedure and only the eccentricity was retained, which would also explain the generally poor performance of image-based models. Additionally, these features might be valuable only in conjunction with clinical variables. Due to restrictions on network size, the maximum number of features that could be attempted for a given model was three, and the combination of both clinical and image features could not be explored fully.

Feature selection

The features utilized in the present work were predominately of a statistical kind and the large majority attempted to capture the intratumoral heterogeneity. Immediate comparisons to specific endpoints without preliminary feature subset reduction would result in an inflation of type-1 errors, and although corrections for multiple testing exist, e.g., the Bonferroni correction and false discovery rate correction, they are often too conservative. Large feature sets are thus undesirable unless carefully defined or properly managed, where the latter inevitably results in a loss of information and the aim of feature reduction methods is to essentially minimize this loss.

The inherently high redundancy between texture features was at first reduced through a correlation-based method, which disregards the specific endpoint entirely and is therefore not susceptible to false discoveries. It is because of this advantage that the method was implemented prior to the minimum redundancy maximum relevance (MRMR) algorithm, which does not share the same property. However, its advantage is also one of its drawbacks, namely that removed features might contain valuable information regarding the specific endpoint, and the hypothesis that three least-correlating features adequately describe a given texture matrix does not necessarily hold. With this in mind, the reduction achieved through the correlation-based method was substantial only for gray-level run length features and GLSZM-based features. These two groups both describe regional variations, and the directional averaging of the gray-level run length matrix (GLRLM) prior to feature extraction makes the GLRLM and the GLSZM very similar. Features derived from these two matrices are thus highly correlated, which is also observed in previous studies where the GLRLM-based features were omitted altogether in favor of GLSZM-based features [17, 36].

Highly informative features may have been lost in the feature selection process and it is therefore more worthwhile to define a feature set that is non-redundant. Approximately half of all features calculated in this study were essentially the reciprocal of the other half, e.g., the short run emphasis and long run emphasis both emphasize length and either one is likely to suffice for the given task. For prediction purposes, a selection procedure is necessary regardless seeing as the number of features included in a model is usually restricted by the number of observations. However, a well-defined feature set might allow for direct use of the MRMR algorithm or perhaps even multivariable analysis without any preliminary steps. Information regarding the specific endpoint is thus considered in all steps and the type-1 error inflation is reduced by the careful definition of features.

Alternative features

Quantitative features are not limited to those used in this work and there are plenty more that can be applied. For instance, Coroller et al. extracted 635 radiomic features of which the majority were derived through filter-based approaches [16]. In such methods, the VOI is processed prior to feature extraction to highlight different aspects of the tumor. Spatial filtering by the Laplacian of the Gaussian can for example reveal rapid changes in pixel values, which in PET/CT corresponds to changes in radiotracer uptake and tissue density, respectively.

Another method is based on the wavelet transform, which is unique in its space-frequency decomposition property, and is the basis for a large number of additional features. The wavelet transform is similar to the Fourier transform in that basis functions are utilized in the decomposition of the original image into frequencies. In contrast, however, the wavelet transform retains information regarding both position and frequency, although at worse resolutions due to Heisenberg's uncertainty principle, i.e., both properties cannot be known precisely and a loss in resolution is inevitable for the simultaneous presentation of both position and frequency. This is a very appealing property which is commonly used in analysis of non-stationary signals. Features based on wavelet coefficients have previously been used to produce excellent classification performance through advanced machine learning algorithms in magnetic resonance imaging (MRI) [44]. Statistical texture features were, however, shown to be significantly better in terms of classification performance for a variety of classifiers in MRI [45]. In radiomic analysis, statistical texture features are usually computed on both the original VOI and the wavelet transformation. The wavelet transform is thus utilized slightly differently and these two methods are consequently complementary.

As many as 384 of the 634 used by Coroller et al. and 288 of the 440 by Aerts et al. were obtained through wavelet analysis [13, 16]. Proper management of large feature sets such as these is no easy task and require well-devised strategies. Both Coroller et al. and Aerts et al. used independent validation cohorts, although the latter group employed independent cohorts in the selection of radiomic features as well. Features were thereby selected with regard to their stability for test-retest and multiple delineation, i.e., the selected features were relatively insensitive with respect to variations in tumor delineation and stochastic aspects of imaging, respectively.

A disadvantage with radiomic features is that they need to be defined beforehand and knowledge regarding what is of importance must be available. This knowledge is apparently not sufficiently specific, as evident by the extremely large feature sets. Convolutional neural networks do not share the prerequisite of manual feature definition but rather take the entire image as input and establish translational, rotational, and scale-invariance through specialization of the initial layers. They are commonly used in image-based object recognition and were in a recent publication compared with the radiomics approach in PET. Convolutional neural networks demonstrated better performance in treatment response prediction for esophageal cancer, as compared to advanced machine learning methods that utilize predefined texture features similar to those employed in the present thesis [46]. Differences in FDG uptake patterns between responders and non-responders are assessed immediately, and feature definitions and all accompanying difficulties related to standardization are thus avoided altogether.

Feature variability

As mentioned previously, statistical texture features, or rather quantitative image-based features in general depend on several factors, some relate to the image acquisition itself whereas other factors arise from postprocessing and subsequent feature extraction. The dependence of global and NGTDM-based features on scanner model was for example investigated with use of a CT phantom containing regions of different textures resembling those observed in non-small cell lung cancer, which were created by three-dimensional printing. Interestingly, texture features were found to differ between manufacturers more so than between different scanner models from the same manufacturer [47]. The realization of the radiomics framework would presumably involve images from different scanners and normalization of the apparent variability is therefore likely necessary.

Additional variability is introduced by reconstruction parameters, where GLCM-based features, followed by GLRLM-based features, were found most robust in a study exploring the influence of five different reconstruction parameters in PET [48]. Local features, such as those derived from the GLCM and NGTDM, were also deemed more stable with respect to partial volume effects [38, 49]. GLCM-based features, specifically, demonstrate higher reproducibility as reported in two recently published test-retest studies where features were extracted from two pretreatment scans taken less than a week apart on the same patients, and GLCM-based features were generally observed to vary less as compared to other feature groups [50, 51]. Features describing local variations appear thus more favorable with regard to image acquisition and preprocessing.

The effects of multiple delineations on different feature groups further attest the higher robustness of local features. Interestingly, GLSZM-based features were shown to be most sensitive to discrepancies in target definition, more so than GLRLM-based features [51]. Due to the highly correlative nature of the two matrices, future research might omit the GLSZM altogether in favor of GLRLM-based features, seeing as they appear more robust.

Parameters related to the quantization procedure have also been shown to affect feature stability. Leijenaar et al. investigated the effect of two different quantization procedures, namely, uniform quantization into a predetermined number of levels as was done for the PET/CT part of the present study, and quantization with predefined decision boundaries, which was actually done in the dose distribution part [52]. These were, however, both studied in PET and for different intensity resolutions, i.e. different quantization levels and bin sizes, respectively. For predefined decision boundaries, an inconsistency across different resolutions was observed and feature values were thus found to depend on the intensity resolution. In contrast, a predetermined quantization level resulted in widely different resolutions for different patients, which is thought to adversely affect inter-patient comparison seeing as FDG distribution is analyzed on extremely different scales. Nevertheless, the general performance of GLCM- and GLRLM-based features were found to be largely unaffected by these variations in the feature values. Furthermore, the performance of the correlation, as derived from the GLCM, was considered least sensitive to variations in intensity resolution and quantization methods. This is quite noteworthy seeing as this feature was the only one found to significantly differ between responders and non-responders for locoregional control in the present thesis. The general robustness of local features, in particular the correlation, is highly encouraging for future works where reproducibility is likely to be assessed.

7 Conclusion

Quantitative analysis of pretreatment positron emission tomography and computed tomography revealed significant differences in three features between responders and non-responders. They could, however, not differentiate between human papillomavirus positive and negative oropharyngeal cancer in this work, as was suggested in previous studies. The use of quantitative image-based features could prove valuable in risk stratification once sources of variability are properly addressed. The radiomic features employed in the present thesis performed poorly in treatment response prediction using an artificial neural network, as compared to clinical variables such as tumor stage. However, they appear to contain additional information that is not expressed by clinical variables and the combination of both radiomic features and clinical variables is likely to be better than either alone.

Custom features based on the gray-level size zone matrix calculated on the dose distribution uncovered differences between responders and non-responders that are not captured by the dose-volume histogram. These differences relate to the number of disconnected regions receiving either too low or too high of a radiation dose. There are several details regarding the extraction of these custom features that need to be examined, which warrants for further investigation.

Acknowledgments

I would like to thank my supervisors Crister Ceberg and Per Nilsson for their continuous support and enthusiasm throughout the entire process. If not for their welcoming offices and the lively discussions held therein, none of this would have been possible, and so I thank you once again. Special thanks go to Johanna Sjövall for her valuable input and for thoroughly going through the patient material with me, and to Mahnaz Haghanegi for her help with archived patients. Lastly, I would like to thank my family for always believing in me even when I struggled to do so myself.

References

[1] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[2] Blausen.com staff. Blausen gallery 2014. *Wikiversity Journal of Medicine*. `http://dx.doi.org/10.15347/wjm/2014.010` Accessed: 2016-06-01.

[3] Lalle Hammarstedt, David Lindquist, Hanna Dahlstrand, Mircea Romanitan, Jeanna Joneberg, Nomi Creson, Johan Lindholm, Weimin Ye, Tina Dalianis, Eva Munck-Wikland, et al. Human papillomavirus as a risk factor for the increase in incidence of tonsillar cancer. *International journal of cancer*, 119(11):2620–2623, 2006.

[4] E.C. Ward and C.J. van As-Brooks. *Head and Neck Cancer: Treatment, Rehabilitation, and Outcomes*. Plural Publishing, Incorporated, 2014.

[5] Carole Fakhry, William H Westra, Sigui Li, Anthony Cmelak, John A Ridge, Harlan Pinto, Arlene Forastiere, and Maura L Gillison. Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial. *Journal of the National Cancer Institute*, 100(4):261–269, 2008.

[6] Akifumi Fujita, Karen Buch, Baojun Li, Yusuke Kawashima, Muhammad M Qureshi, and Osamu Sakai. Difference between hpv-positive and hpv-negative non-oropharyngeal head and neck cancer: Texture analysis features on ct. *Journal of computer assisted tomography*, 40(1):43–47, 2016.

[7] K Buch, A Fujita, B Li, Y Kawashima, MM Qureshi, and O Sakai. Using texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinomas on ct. *American Journal of Neuroradiology*, 36(7):1343–1348, 2015.

[8] Ronald Boellaard. Standards for pet image acquisition and quantitative data analysis. *Journal of nuclear medicine*, 50(Suppl 1):11S–20S, 2009.

[9] Balaji Ganeshan, Vicky Goh, Henry C Mandeville, Quan Sing Ng, Peter J Hoskin, and Kenneth A Miles. Non–small cell lung cancer: histopathologic correlates for texture parameters at ct. *Radiology*, 266(1):326–336, 2013.

[10] Hubert Vesselle, Rodney A Schmidt, Jeffrey M Pugsley, Melissa Li, Steve G Kohlmyer, Eric Vallières, and Douglas E Wood. Lung cancer proliferation correlates with [f-18] fluorodeoxyglucose uptake by positron emission tomography. *Clinical Cancer Research*, 6(10):3837–3844, 2000.

[11] Joseph G Rajendran, David L Schwartz, Janet O'Sullivan, Lanell M Peterson, Patrick Ng, Jeffrey Scharnhorst, John R Grierson, and Kenneth A Krohn. Tumor hypoxia imaging with [f-18] fluoromisonidazole positron emission tomography in head and neck cancer. *Clinical Cancer Research*, 12(18):5435–5441, 2006.

[12] Reinhard Bos, Jacobus JM van der Hoeven, Elsken van der Wall, Petra van der Groep, Paul J van Diest, Emile FI Comans, Urvi Joshi, Gregg L Semenza, Otto S Hoekstra, Adriaan A Lammertsma, et al. Biologic correlates of 18fluorodeoxyglucose uptake in human breast cancer measured by positron emission tomography. *Journal of Clinical Oncology*, 20(2):379–387, 2002.

[13] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5, 2014.

[14] Balaji Ganeshan, Elleny Panayiotou, Kate Burnand, Sabina Dizdarevic, and Ken Miles. Tumour heterogeneity in non-small cell lung carcinoma assessed by ct texture analysis: a potential marker of survival. *European radiology*, 22(4):796–802, 2012.

[15] David V Fried, Susan L Tucker, Shouhao Zhou, Zhongxing Liao, Osama Mawlawi, Geoffrey Ibbott, and Laurence E Court. Prognostic value and reproducibility of pretreatment ct texture features in stage iii non-small cell lung cancer. *International Journal of Radiation Oncology* Biology* Physics*, 90(4):834–842, 2014.

[16] Thibaud P Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H Mak, and Hugo JWL Aerts. Ct-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, 2015.

[17] Florent Tixier, Catherine Cheze Le Rest, Mathieu Hatt, Nidal Albarghach, Olivier Pradier, Jean-Philippe Metges, Laurent Corcos, and Dimitris Visvikis. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine*, 52(3):369–378, 2011.

[18] Gary JR Cook, Connie Yip, Muhammad Siddique, Vicky Goh, Sugama Chicklore, Arunabha Roy, Paul Marsden, Shahreen Ahmad, and David Landau. Are pretreatment 18f-fdg pet tumor textural features in non–small cell lung cancer associated with response and survival after chemoradiotherapy? *Journal of Nuclear Medicine*, 54(1):19–26, 2013.

[19] Nai-Ming Cheng, Yu-Hua Dean Fang, Joseph Tung-Chieh Chang, Chung-Guei Huang, Din-Li Tsan, Shu-Hang Ng, Hung-Ming Wang, Chien-Yu Lin, Chun-Ta Liao, and Tzu-Chen Yen. Textural features of pretreatment 18f-fdg pet/ct images: prognostic significance in patients with advanced t-stage oropharyngeal squamous cell carcinoma. *Journal of Nuclear Medicine*, 54(10):1703–1709, 2013.

[20] Issam El Naqa, PW Grigsby, A Apte, E Kidd, E Donnelly, D Khullar, S Chaudhari, Deshan Yang,

M Schmitt, Richard Laforest, et al. Exploring feature-based approaches in pet images for predicting cancer treatment outcomes. *Pattern recognition*, 42(6):1162–1171, 2009.

[21] Ralph TH Leijenaar, Sara Carvalho, Frank JP Hoebers, Hugo JWL Aerts, Wouter JC van Elmpt, Shao Hui Huang, Biu Chan, John N Waldron, Brian O'sullivan, and Philippe Lambin. External validation of a prognostic ct-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncologica*, 54(9):1423–1429, 2015.

[22] Ke Nie, Jeon-Hor Chen, J Yu Hon, Yong Chu, Orhan Nalcioglu, and Min-Ying Su. Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast mri. *Academic radiology*, 15(12):1513–1525, 2008.

[23] Kirsi Holli, Anna-Leena Lääperi, Lara Harrison, Tiina Luukkaala, Terttu Toivonen, Pertti Ryymin, Prasun Dastidar, Seppo Soimakallio, and Hannu Eskola. Characterization of breast cancer types by texture analysis of magnetic resonance images. *Academic radiology*, 17(2):135–141, 2010.

[24] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 6:610–621, 1973.

[25] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.

[26] Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira, and JJ Mari. Texture indexes and gray level size zone matrix: application to cell nuclei classification. *Pattern Recognition and Information Processing*, pages 140–145, 2009.

[27] Moses Amadasun and Robert King. Textural features corresponding to textural properties. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(5):1264–1274, 1989.

[28] Paulo J Lisboa and Azzam FG Taktak. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415, 2006.

[29] Giuseppe Coppini, Stefano Diciotti, Massimo Falchini, Natale Villari, and Guido Valli. Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. *Information Technology in Biomedicine, IEEE Transactions on*, 7(4):344–357, 2003.

[30] Yuzheng Wu, Maryellen L Giger, Kunio Doi, Carl J Vyborny, Robert A Schmidt, and Charles E Metz. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology*, 187(1):81–87, 1993.

[31] Christopher M. Bishop. *Neural networks for pattern recognition.* Oxford university press, 1995.

[32] Simon S. Haykin. *Neural networks and learning machines.* Prentice Hall, 2009.

[33] Johanna Sjövall. *PET in the evaluation of head and neck cancer treatment.* PhD thesis, Department of Clinical Sciences, Lund University, Sweden, 2015.

[34] MATLAB programming tools for radiomics analysis. `https://github.com/mvallieres/radiomics`. Accessed: 2015-12-17.

[35] M Vallières, CR Freeman, SR Skamene, and I El Naqa. A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in medicine and biology*, 60(14):5471, 2015.

[36] Fanny Orlhac, Michaël Soussan, Jacques-Antoine Maisonobe, Camilo A Garcia, Bruno Vanderlinden, and Irène Buvat. Tumor texture analysis in 18f-fdg pet: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *Journal of Nuclear Medicine*, 55(3):414–422, 2014.

[37] David A Clausi. Comparison and fusion of co-occurrence, gabor and mrf texture features for classification of sar sea-ice imagery. *Atmosphere-Ocean*, 39(3):183–194, 2001.

[38] Floris HP van Velden, Patsuree Cheebsumon, Maqsood Yaqub, Egbert F Smit, Otto S Hoekstra, Adriaan A Lammertsma, and Ronald Boellaard. Evaluation of a cumulative suv-volume histogram method for parameterizing heterogeneous intratumoural fdg uptake in non-small cell lung cancer pet studies. *European journal of nuclear medicine and molecular imaging*, 38(9):1636–1647, 2011.

[39] David J Livingstone and David T Manallack. Statistics using neural networks: chance effects. *Journal of medicinal chemistry*, 36(9):1295–1297, 1993.

[40] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[41] Richard Simon, Michael D Radmacher, Kevin Dobbin, and Lisa M McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18, 2003.

[42] Branko Krišto and Marko Buljan. The lymph node roundness index in the evaluation of lymph nodes of the neck. *Collegium antropologicum*, 39(1):165–169, 2015.

[43] Thomas J Bryce, Mark W Dewhirst, Carey E Floyd, Vera Hars, and David M Brizel. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *International Journal of Radiation Oncology* Biology* Physics*, 41(2):339–345, 1998.

[44] Sandeep Chaplot, LM Patnaik, and NR Jagannathan. Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network. *Biomedical Signal Processing and Control*, 1(1):86–92, 2006.

[45] Namita Aggarwal and RK Agrawal. First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing*, 3:146–153, 2012.

[46] Petros-Pavlos Ypsilantis, Musib Siddique, Hyon-Mok Sohn, Andrew Davies, Gary Cook, Vicky Goh, and Giovanni Montana. Predicting response to neoadjuvant chemotherapy with pet imaging using convolutional neural networks. *PloS one*, 10(9):e0137036, 2015.

[47] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, et al. Measuring computed tomography scanner variability of radiomics features. *Investigative radiology*, 50(11):757–765, 2015.

[48] Paulina E Galavis, Christian Hollensen, Ngoneh Jallow, Bhudatt Paliwal, and Robert Jeraj. Variability of textural features in fdg pet images due to different acquisition modes and reconstruction parameters. *Acta Oncologica*, 49(7):1012–1016, 2010.

[49] Mathieu Hatt, Florent Tixier, Catherine Cheze Le Rest, Olivier Pradier, and Dimitris Visvikis. Robustness of intratumour 18f-fdg pet uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *European journal of nuclear medicine and molecular imaging*, 40(11):1662–1671, 2013.

[50] Florent Tixier, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18f-fdg pet. *Journal of Nuclear Medicine*, 53(5):693–700, 2012.

[51] Ralph TH Leijenaar, Sara Carvalho, Emmanuel Rios Velazquez, Wouter JC Van Elmpt, Chintan Parmar, Otto S Hoekstra, Corneline J Hoekstra, Ronald Boellaard, André LAJ Dekker, Robert J Gillies, et al. Stability of fdg-pet radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta Oncologica*, 52(7):1391–1397, 2013.

[52] Ralph TH Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter JC van Elmpt, Esther GC Troost, Ronald Boellaard, Hugo JWL Aerts, Robert J Gillies, and Philippe Lambin. The effect of suv discretization in quantitative fdg-pet radiomics: the need for standardized methodology in tumor texture analysis. *Scientific reports*, 5, 2015.

[53] A Gebejes and R Huertas. Texture characterization based on grey-level co-occurrence matrix. *Proceedings of ICTIC for Conference of Informatics And Management Sciences*, pages 375–378, 2013.

[54] A Chu, Chandra M Sehgal, and James F Greenleaf. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11(6):415–419, 1990.

[55] Belur V Dasarathy and Edwin B Holder. Image characterizations based on joint gray level—run length distributions. *Pattern Recognition Letters*, 12(8):497–502, 1991.

Appendix A Statistical texture features

This appendix contains definitions of the statistical texture features used in this work, which were computed in Matlab R2015b using the RADIOMICS package available at [34].

Gray-level co-occurrence features

Texture features derived from the gray-level co-occurrence matrix (GLCM) as defined by Haralick et al. [24], with minor modifications to the homogeneity, correlation, variance, and dissimilarity [53] are presented below. Four common metrics appearing in several texture features are defined as

$$\mu_i = \sum_{i=1}^{L} i \sum_{j=1}^{L} P(i,j), \qquad \sigma_i = \sum_{i=1}^{L} (i-\mu_i)^2 \sum_{j=1}^{L} P(i,j),$$

$$\mu_j = \sum_{j=1}^{L} j \sum_{i=1}^{L} P(i,j), \qquad \sigma_j = \sum_{j=1}^{L} (j-\mu_j)^2 \sum_{i=1}^{L} P(i,j),$$

where $P(i,j)$ is the normalized gray-level co-occurrence matrix, L is the number of gray-levels, and μ and σ are the mean and standard deviation, respectively, calculated for rows and columns separately.

$$\text{Energy} = \sum_{i=1}^{L}\sum_{j=1}^{L} P(i,j)^2 \tag{A.1}$$

$$\text{Contrast} = \sum_{i=1}^{L}\sum_{j=1}^{L} (i-j)^2 P(i,j) \tag{A.2}$$

$$\text{Entropy} = -\sum_{i=1}^{L}\sum_{j=1}^{L} P(i,j)\log_2(P(i,j)) \tag{A.3}$$

$$\text{Homogeneity} = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{P(i,j)}{1+|i-j|} \tag{A.4}$$

$$\text{Correlation} = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i\sigma_j} \tag{A.5}$$

$$\text{Variance} = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\Big((i-\mu_i)^2P(i,j)+(j-\mu_j)^2P(i,j)\Big) \tag{A.6}$$

$$\text{Dissimilarity} = \sum_{i=1}^{L}\sum_{j=1}^{L}|i-j|P(i,j) \tag{A.7}$$

$$\text{Sum average} = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}(iP(i,j)+jP(i,j)) \tag{A.8}$$

Gray-level run length features

Several features based on the gray-level run length matrix (GLRLM) are presented below. Features (A.9) to (A.13) are calculated as defined by Galloway [25], while the remaining are presented in [26, 54, 55]. First, the mean over rows and columns, respectively, are defined as

$$\mu_i = \sum_{i=1}^{L} i \sum_{r=1}^{L_r} P(i,r), \qquad \mu_r = \sum_{r=1}^{L_r} r \sum_{i=1}^{L} P(i,r),$$

where $P(i,r)$ is the normalized GLRLM, and L and L_r are the number of gray-levels and the length of the longest run, respectively.

$$\text{Short run emphasis (SRE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r}\frac{P(i,r)}{r^2} \tag{A.9}$$

$$\text{Long run emphasis (LRE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r} r^2P(i,r) \tag{A.10}$$

$$\text{Gray-level nonuniformity (GLN)} = \sum_{i=1}^{L}\left(\sum_{r=1}^{L_r}P(i,r)\right)^2 \tag{A.11}$$

$$\text{Run-length nonuniformity (RLN)} = \sum_{r=1}^{L_r}\left(\sum_{i=1}^{L}P(i,r)\right)^2 \tag{A.12}$$

$$\text{Run percentage (RP)} = \frac{\sum_{i=1}^{L}\sum_{r=1}^{L_r}P(i,r)}{\sum_{r=1}^{L_r} r\sum_{i=1}^{L}P(i,r)} \tag{A.13}$$

$$\text{Low gray-level run emphasis (LGRE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r}\frac{P(i,r)}{i^2} \tag{A.14}$$

$$\text{High gray-level run emphasis (HGRE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r} i^2 P(i,r) \tag{A.15}$$

$$\text{Short run low gray-level emphasis (SRLGE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r}\frac{P(i,r)}{i^2 r^2} \tag{A.16}$$

$$\text{Short run high gray-level emphasis (SRHGE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r}\frac{i^2 P(i,r)}{r^2} \tag{A.17}$$

$$\text{Long run low gray-level emphasis (LRLGE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r}\frac{r^2 P(i,r)}{i^2} \tag{A.18}$$

$$\text{Long run high gray-level emphasis (LRHGE)} = \sum_{i=1}^{L}\sum_{r=1}^{L_r} i^2 r^2 P(i,r) \tag{A.19}$$

$$\text{Gray-level variance (GLV)} = \frac{1}{L \cdot L_r}\sum_{i=1}^{L}\sum_{r=1}^{L_r}(iP(i,r) - \mu_i)^2 \tag{A.20}$$

$$\text{Run-length variance (RLV)} = \frac{1}{L \cdot L_r}\sum_{i=1}^{L}\sum_{r=1}^{L_r}(rP(i,r) - \mu_r)^2 \tag{A.21}$$

Gray-level size zone features

As previously, the mean over rows and columns are defined as

$$\mu_i = \sum_{i=1}^{L} i \sum_{z=1}^{L_z} P(i,z), \qquad \mu_z = \sum_{z=1}^{L_z} z \sum_{i=1}^{L} P(i,z),$$

where $P(i,z)$ is the normalized gray-level size zone matrix (GLSZM), L is the number of gray-levels, and L_z is the size of the largest contiguous zone. The following features are calculated as described in [25, 26, 54, 55]:

$$\text{Small zone emphasis (SZE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z}\frac{P(i,z)}{z^2} \tag{A.22}$$

$$\text{Large zone emphasis (LZE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z} z^2 P(i,z) \tag{A.23}$$

$$\text{Gray-level nonuniformity (GLN)} = \sum_{i=1}^{L}\left(\sum_{z=1}^{L_z} P(i,z)\right)^2 \tag{A.24}$$

$$\text{Zone size non-uniformity (ZSN)} = \sum_{z=1}^{L_z}\left(\sum_{i=1}^{L} P(i,z)\right)^2 \tag{A.25}$$

$$\text{Zone percentage (ZP)} = \frac{\sum_{i=1}^{L}\sum_{z=1}^{L_z} P(i,z)}{\sum_{z=1}^{L_z} z \sum_{i=1}^{L} P(i,z)} \tag{A.26}$$

$$\text{Low gray-level zone emphasis (LGZE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z}\frac{P(i,z)}{i^2} \tag{A.27}$$

$$\text{High gray-level zone emphasis (HGZE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z} i^2 P(i,z) \tag{A.28}$$

$$\text{Small zone low gray-level emphasis (SZLGE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z}\frac{P(i,z)}{i^2 z^2} \tag{A.29}$$

$$\text{Small zone high gray-level emphasis (SZHGE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z}\frac{i^2 P(i,z)}{z^2} \tag{A.30}$$

$$\text{Large zone low gray-level emphasis (LZLGE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z}\frac{z^2 P(i,z)}{i^2} \tag{A.31}$$

$$\text{Large zone high gray-level emphasis (LZHGE)} = \sum_{i=1}^{L}\sum_{z=1}^{L_z} i^2 z^2 P(i,z) \tag{A.32}$$

$$\text{Gray-level variance (GLV)} = \frac{1}{L \cdot L_z}\sum_{i=1}^{L}\sum_{z=1}^{L_z}(iP(i,z)-\mu_i)^2 \tag{A.33}$$

$$\text{Zone size variance (ZSV)} = \frac{1}{L \cdot L_z}\sum_{i=1}^{L}\sum_{z=1}^{L_z}(zP(i,z)-\mu_z)^2 \tag{A.34}$$

Neighborhood gray-tone difference features

With $D(i)$ as defined in the section concerning the neighborhood gray-tone difference matrix (NGTDM), and n_i as

$$n_i = \frac{N_i}{N},$$

where N_i is the number of pixels wtih gray-level i and N being the total number of pixels, the following NGTDM-based features are calculated according to [27]:

$$\text{Coarseness} = \left(\epsilon + \sum_{i=1}^{L} n_i D(i) \right)^{-1}, \tag{A.35}$$

$$\text{Contrast} = \left(\frac{1}{L_{\text{eff}}(L_{\text{eff}} - 1)} \sum_{i=1}^{L} \sum_{j=1}^{L} n_i n_j (i-j)^2 \right) \left(\frac{1}{L} \sum_{i=1}^{L} D(i) \right), \tag{A.36}$$

$$\text{Busyness} = \frac{\sum_{i=1}^{L} n_i D(i)}{\sum_{i=1}^{L} \sum_{j=1}^{L} (i n_i - j n_j)} \qquad : n_i \neq 0,\ n_j \neq 0, \tag{A.37}$$

$$\text{Complexity} = \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{|i-j|(n_i D(i) + n_j D(j))}{L(n_i + n_j)} \qquad : n_i \neq 0,\ n_j \neq 0, \tag{A.38}$$

$$\text{Strength} = \frac{\sum_{i=1}^{L} \sum_{j=1}^{L} (n_i + n_j)(i-j)^2}{\epsilon + \sum_{i=1}^{L} D(i)} \qquad : n_i \neq 0,\ n_j \neq 0, \tag{A.39}$$

where ϵ in equation (A.35) and (A.39) is a small number to assure a finite result and L_{eff} is the effective number of gray-levels.