MRI-based radiomics to predict radiation-induced effects in mouse salivary glands

Toralf Husevåg

Biological and Medical Physics
60 credits
Department of Physics
Faculty of Mathematics and Natural Sciences

University of Oslo
2022

Acknowledgements

I want to thank all my supervisors whom I would never have completed this thesis without.

Thank you to my main supervisor Prof. Eirik Malinen for countless inspiring conversations and discussions relating to all topics within medical physics, and for enduring my never-ending supply of stupid questions.

Thank you Prof. Nina Frederike Jeppesen Edin for helping me understanding big-picture questions, physiology and biological interpretations, and general thesis structure.

Thank you, PhD student Delmon Arous, for helping me with confusing statistics and segmentation.

Thank you, PhD student Olga Zlygosteva, for answering questions relating to mouse physiology, and spending many hours in collaboration with PhD student Inga Solgård Juvkam keeping the study mice healthy and happy – supplying all raw data used in this work.

Thank you to the people in the Department of Physics at NMBU for spending time discussing methodologies used in, and experiences with, Radiomics, providing great inspiration for this work.

Thanks to all master students at the study room for supplying a great social environment with healthy academic discussions.

Abstract

Radiation therapy (RT) is often included in, or used as a stand-alone, treatment of cancer in the head and neck region (HNC). The salivary glands (SGs) are often in close proximity to the tumour and are therefore not always possible to spare from irradiation when using RT to treat HNC. The response of the SGs to irradiation show variations between patients, indicating that some are more radiosensitive than others. Incorporating more patient-specific biomarkers into treatment planning and evaluations during fractionated radiotherapy are needed to establish a precision oncology framework for mitigation of side-effects such as xerostomia (dry mouth).

Radiomic image features from medical imaging have previously shown potential as biomarkers for risk of developing xerostomia post-RT. Radiomics is a high-throughput method of extracting quantitative information from such images and the features may be broadly categorized as shape-based, first-order, and texture-based. This work evaluated the relation between 828 radiomic features calculated from 2D regions of interests (ROIs) in either T1- or T2-weighted magnetic resonance images (MRI) to known biological changes in the SGs due to damage by ionizing radiation. C57BL/6J mice were used, where 72 individuals were irradiated while 40 belonged to a control group.

The developed radiomic workflow includes creation of ROIs for each image (segmentation), preprocessing, feature extraction, feature selection, and modelling. Radiomic studies are known to have an issue with reproducibility and feature robustness, and therefore all steps in the workflow are evaluated and described in detail. Intensity normalization was performed on a feature-specific level.

The segmented ROIs were evaluated against sublingual gland (SLG) and submandibular gland (SMG) areas from 9 surgical specimens. While the SLG areas had higher correlation to the image-segmented ROIs than the SMG areas ($\rho = 0.71$ and 0.36, respectively), the two types of glands could not be differentiated in the images due to being fused in mice. Saliva production was found to be significantly lower in irradiated individuals relative control comparing data from between day 26 and 105 post-irradiation. Xerostomia was defined into a binary outcome variable by thresholding.

Using only image features from T1 images proved to be significantly better predictors of xerostomia than the T2 features when evaluated on the same data. The relative difference in a T2 first-order feature before and after pilocarpine injections for saliva measurements, delta-p energy, was shown to be a high-performing predictor of xerostomia evaluated on the same day as the MR imaging. Overall, 2D features from the right SG-subunit proved to be higher-performing features than the left subunit, possibly due to some differences in delivered dose.

Both T1 and T2 features obtained from MRI after irradiation were good predictors of late xerostomia, but only T1 features showed a possible predictive ability at baseline. The relative difference in a shape-feature before and after irradiation (delta-feature) showed promise of predicting late xerostomia. Multiple textural features from both T1 and T2 images were good predictors of late xerostomia, possibly related to changes in vascularity or increased fatty tissue in the glands post-irradiation.

The radiomic image features were able to predict saliva production in C57BL/6J mice with varying accuracy. 14 features significantly improved upon models only using time and dose as predictors, indicating that certain features contained information relating to the inter-mouse variations affecting saliva production. However, none of the features were significant under Bonferroni corrected p-threshold, emphasizing the need for validating studies on external data.

Abbreviations

- CFRT (RT): Conformal radiotherapy – dose optimization technique by shaping the irradiation field(s) to the PTV, thus reducing delivered dose to healthy tissue and OARs
- CLAHE: Contrast Limited Adaptive Histogram Equalization – method of increasing the contrast in an image while suppressing noise
- FBC: Fixed bin count - method of image intensity discretization
- FBW: Fixed bin width - method of image intensity discretization
- FID (NMR): Free Induction Decay – basic concept in NMR
- FSE (MRI): Fast Spin Echo – rapid image acquisition technique in MRI
- GTV (RT): TV + subclinical carcinoma
- HN: Head and neck (region)
- HNC: Head and neck cancer / carcinoma in the HN region
- IMRT (RT): Intensity modulated radiotherapy – type of CFRT where the beam intensity is varied *during* the RT to achieve an optimal dose distribution in the PTV
- IQR: Interquartile range – statistical term
- MRI: Magnetic Resonance Imaging
- MRMR: Maximum-relevance minimum-redundancy – method of feature selection
- N4: Non-parametric image processing method to counteract bias field artifacts in MRI
- NMR: Nuclear Magnetic Resonance – the basis for MRI
- OAR (RT): Organ at risk (due to RT of PTV in its proximity)
- PG: parotid gland – type of SG
- PROCCA: Protons Contra Cancer – interdisciplinary research environment considering the biological short- and long-term effects following RT of HNC, which this work is a part of
- PTV (RT): Planning tumour volume (what to irradiate)
- QIBA: The Quantitative Image Biomarker Association – North American association attempting to standardize the process of imaging biomarker identification and clinical implementation
- RARE (MRI): Rapid acquisition with refocused echoes – today referred to as FSE
- RF: radiofrequency – non-ionizing radiation
- ROI: Region of interest – subset of pixels in a 2D image
- RT: Radiotherapy
- SD: Standard deviation – statistical concept
- SE (MRI): Spin echo - MR-imaging pulse sequence where spins are refocused by a 180-degree pulse producing an “echo”
- SG: Salivary gland

- SLG: Sublingual gland - type of SG
- SMG: Submandibular gland - type of SG
- SNR (MRI): Signal to noise ratio
- TE (MRI): Echo time - time from centre of RF-pulse to centre of the signal echo
- TR (MRI): Repetition time - time between centres of consecutive RF-pulses
- TV (RT): Tumour volume (“true” volume of tumour)
- VOI: Volume of interest – subset of voxels in a 3D image

Table of contents

1 Introduction

Cancer is a pathologic term describing groups of cells displaying abnormal growth in some way, characterized by uncontrolled proliferation [1]. In 2020, cancer in the head and neck region (HNC) made up about 800 new cases each year in Norway [2].

Radiotherapy (RT) employs ionizing radiation, typically high-energy X-rays, impinging on the tumour from many angles. The aim is to eradicate the cancer by inactivating all malignant tumour cells. However, irradiation of healthy normal tissue is largely unavoidable and may cause side effects. RT is in many cases central for treating HNC and may be used in combination with other modalities to increase the likelihood of patient survival. RT may be used alone or after surgical resection of the tumour. For the latter, RT aims to eliminate microscopic remains of the tumour, thus reducing the risk of recurrence [3].

However, radiotherapy of cancer in the head and neck region is known to induce side effects such as xerostomia (dry mouth) due to the tumours' proximity to healthy tissue and organs such as the salivary glands (SGs) [4].

Information before start of treatment about the patient currently includes a computed tomography (CT) scan of the head and neck region, where the tumour and organs at risk are delineated. The CT images and the delineated volumes, together with physics-based radiation transport model and optimization techniques, are used to derive e.g. the dose-volume relationships. This is then used to plan the delivery and evaluate the probability of complications in the healthy tissue. However, the SG dose-response have shown high variation between patients and studies have indicated the need for incorporating more patient-specific data into the treatment planning and follow-up to mitigate radiation-induced side effects [5]. Taking such individual patient variations into account is often referred to as *precision oncology*.

Medical imaging data from e.g. CT or magnetic resonance imaging (MRI) may contain patient-specific information with potential applications in precision oncology. Also, medical imaging is inherently non-invasive, which is an advantage over methods requiring tissue sampling. Radiomics attempts to find relationships between medical imaging biomarkers (radiomic features) and some biological phenomenon while also taking inter-patient variations into account. Identified radiomic features may therefore be viable biomarkers for use in precision oncology, as a tool for decision-making and risk assessment both in treatment planning and during the delivery of fractionated RT. This may in turn improve patient care and treatment outcomes [6].

The work described in this thesis extracts radiomic features from MR images that may be used to predict xerostomia, following irradiation of the salivary glands in C57BL/6J mice. The radiomic features are hypothesized to contain information about the inter- and inter-mice variations in saliva production, which this thesis attempts to validate. The radiomics workflow includes segmentation, feature extraction, feature selection, and modelling [7]–[9].

The segmentation procedure used for creating 2-dimensional regions of interests (ROIs) of the salivary glands was a semi-automatic method using a watershed approach. Various steps in image preprocessing, before extraction of radiomic features, are evaluated and implemented into

the developed radiomics pipeline. Preprocessing includes MRI bias-field correction, intensity normalization, re-segmentation, and discretization by fixed bin widths. The type of normalization is chosen on a feature-specific basis.

As the MR-images were taken over time the radiomic features are evaluated on their ability to predict saliva amounts measured on the same day as the images were acquired, or forward in time. Saliva production was measured after injecting each mouse with pilocarpine. Taking the relative difference between each image feature from before and after irradiation, known as delta-radiomics, have shown previous success with predicting late xerostomia [10] and is also evaluated in this work.

Different types of radiomic features were evaluated in this work. Standard radiomic features were extracted from either T1 or T2-weighted MR-images and two variations of relative differences between features were also derived. The previously established delta-features being the relative difference in time, and the proposed delta-p features being the relative difference before and after pilocarpine injections (i.e. saliva extraction). The two latter types of relative differences in radiomic features are only based on T2-images due to a higher amount of data than the T1-images.

Multiple linear and random forest regression were used to predict the continuous saliva amounts, using the four feature-types. Similarly, a logistic regression and a random forest classifier were used to predict binary outcomes representing xerostomia, found by thresholding based on the expectation values from the control individuals over time. Models were evaluated by either splitting the data into a train and test set, or by resampling methods such as k-fold or leave one out cross-validation (LOOCV).

2 Theory and background

2.1 Radiotherapy of cancer in the head and neck region

As mentioned in section 1, radiotherapy (RT) of head and neck cancer (HNC) is known to induce side effects. Such effects may be either acute or long term, dependent on the RT delivery (radiation dose and technique) and the tissue or organs present in the irradiation fields. In many cases the tumour is located in near vicinity to the salivary glands (SGs), making them difficult to spare from the ionizing radiation. A high dose delivered to the SGs may cause injuries which may be irreparable. Such irradiation-induced injuries to the SGs remains an area of active research, as it is of clinical interest to mitigate such effects [4].

This section describes the basics in radiation physics and radiobiology with implications for radiotherapy, before describing the biology of the healthy and irradiated salivary glands.

2.1.1 An overview of radiation physics and dosimetry

The following chapter is written using Introduction to Radiological Physics and Radiation Dosimetry (Attix, 1986 [11]) as source.

Ionizing radiation is defined by the radiation's ability to excite and ionize atoms of the matter with which it interacts. The radiation may be made up of electromagnetic waves (photons), charged particles (electrons, protons, heavier nuclei), or neutrons.

Two main types of interactions may occur between the photon and an atom of molecule in the irradiated medium (target): *scattering* and *absorption*. In a coherent (Rayleigh) scattering event the photon does not lose energy, while in an incoherent (*Compton*) scattering event some energy is transferred from the photon to the target particle. The *photoelectric effect* is another type of interaction where an absorption of a photon induces either an excited state (causing the emission of a new *characteristic photon* or the release of an *auger electron* during de-excitation) or a direct liberation of an electron (ionization). The final interaction possibility for transferring energy directly from photons to electrons is though *pair production* – the absorption of a photon in the target's coulomb field with a following creation of an electron and a positron.

Whether any single interaction between the radiation particle and a particle in the irradiated material (target) occurs is of a stochastic nature (due to these being quantum mechanical processes) and may be described probabilistically using the *cross-section*. The *total mass attenuation*, describing the loss of *fluence* (number of particles per unit area) in a photon beam going through some material, becomes a linear combination of each interaction cross-section. The dominant interaction is dependent on the atomic number of the target, as well as the energy of the incoming photon (see Figure 2-1).

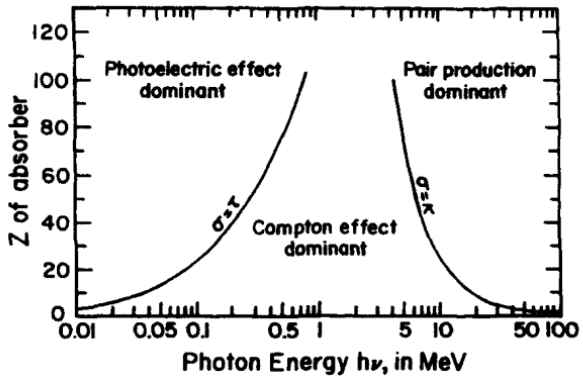


Figure 2-1: Dominant electron-releasing interaction for varying photon energy and atomic number Z in target. σ, τ, κ are the cross-sections for Compton scattering, photoelectric effect, and pair production, respectively. Image is from [11].

The production of x-rays, either by a x-ray tube or linear accelerator, involves accelerating electrons before they collide with a heavy element (target). The rapid retardation of electrons in collision with the target, *brehmsstrahlung* (braking-radiation), causes an induction of high-frequency EM waves: x-rays. The *quality* of the produced x-rays, being a metric describing some aspect of the x-ray beam's energy distribution, may be manipulated by beam *filtering* through some metal.

Knowing how much imparted energy, of which the expectation value per unit mass becomes *dose*, a specific region of a material has received due to irradiation is a question of *dosimetry*. *Ionometry* correlates the dose to a measurement of the number of ionizations within some enclosed area (e.g. measured using an electrometer), being an example of *absolute dosimetry* as the method does not require external information for calculation of the dose. However, in practice the method is often used for *relative dosimetry* requiring a calibration of the *ion chamber* at some conditions where the dose is known (e.g. the dose to a known mass of water at ambient conditions).

2.1.2 A brief overview of radiobiology and radiotherapy

This majority of this section is based on the 2019 book Radiobiology for the radiologist by Eric J. Hall [12].

Contained in a double layered lipid membrane, the cell nucleus, the DNA is considered the radiosensitive biomolecule of the cell. The DNA contains all genetic information for production of necessary proteins in the human body.

Corresponding to the molecular structure of a specific protein, the information in a gene is *transcribed* onto a messenger RNA (mRNA) before the information is *translated* into the protein outside the cell nucleus.

The *self-replication* of cells in done by the doubling of all necessary proteins for sustaining both the old and new cell, in addition to copying the whole DNA (*DNA replication*), before division into two separate cells. The part of the *cell cycle* where the cell divides is known as *mitosis*, or M-phase. The cell grows and prepares for division during *interphase,* during which the DNA is accessible for transcription, which is further subdivided into the three phases G1, S, and G2. Cellular growth occurs during the G-phases, while the DNA replication occurs in S-phase. Before mitosis, the DNA is efficiently packed into a coil known as the chromosome. After DNA replication a *sister chromatid* is connected to the original chromosome at the *centromere*, containing all genetic information to sustain further cell life following the cell division during mitosis.

While charged particles have a higher probability of directly ionizing a biomolecule in an interaction event, x-rays deposit their energy in proximity creating *free radicals*. The free radicals, or the ionizing irradiation directly, damages or inactivates the cell by altering the DNA.

Ionizing irradiation may alter the structure of the DNA in a damaging way by either depositing energy directly onto the DNA, or by induction of reactive *free-radicals* in the nearby environment. The damage is characterized by discontinuities in the DNA stand (strand breaks). The severity of the damage, and the cell's potential for repair, differs for a *single strand break* (SSB) and a *double strand break* (DSB). The damage may be lethal for the cell but may also be possible to be repair. Following a DSB, the possible reparation mechanisms vary based on where in the cell cycle the cell is. If the cell is in late G2 or S-phase, the DNA may be completely restored by using information available in the sister chromatid. This is known as homologous recombination (HR). If the cell is in G1 an attempt to repair the cell is by non-homologous end-joining (NHEJ). The separated DNA-strands are re-attached at the site of the strand breaks, leading to a potential loss of genetic information.

Following an incomplete reparation of irradiation-induced DNA damage, a mutation may occur as a change in the genomic structure. If the cell is able to divide and survive with said mutation the risk of cancer may increase, or the function of physiological systems may become affected. Cellular mechanisms exists to mitigate mutations following irradiation damage, such as controlled cell death (apoptosis) or an irreversible arrest in the G1 or S-phase [13]. If the cell is lacking the apoptotic response pathways, the sister chromatids may become inseparable during mitosis leading to incomplete division and uncontrolled cell death. The latter is known as the mitotic catastrophe, or mitotic cell death.

The standard for modelling cell culture survival following ionizing irradiation is by using the linear quadratic (LQ) model. The LQ-model assumes an exponential decrease in number of surviving cells with increased dose, based on a second-order polynomial in the exponent.

$$\ln(S) = -(\alpha D + \beta D^2) \qquad 2.1\text{-}1.$$

The two parameters in the LQ-model, seen as alpha and beta in equation 2.1-1, are cell-specific. Their ratio describes the radiosensitivity for a cell type characterized by an early or late response

to irradiation [14]. As tumour cells usually have a higher alpha / beta ratio compared to the healthy surrounding tissue, they are more sensitive to multiple small irradiation doses delivered over a large time-span than early responders having a lower alpha / beta ratio. *Temporally fractionated* dose-delivery regimes are therefore used to maximize the killing of cancer cells, mathematically formulated through the tumour control probability (TCP), while maintaining a low normal tissue complication probability (NTCP).

The maximization of TCP while keeping NTCP low is a key component of radiotherapy. Another important aspect for consideration to increase the probability of a complication-free cure using RT, are organs positioned close to the tumour. Their close proximity to the tumour, and therefore potentially in the irradiation field, makes them organs at risk (OAR). Special techniques for optimizing the dose distribution within a patient with respect to lowering the dose to OARs exists, such as conformal radiotherapy (CFRT) and intensity-modulated radiotherapy (IMRT) [15].

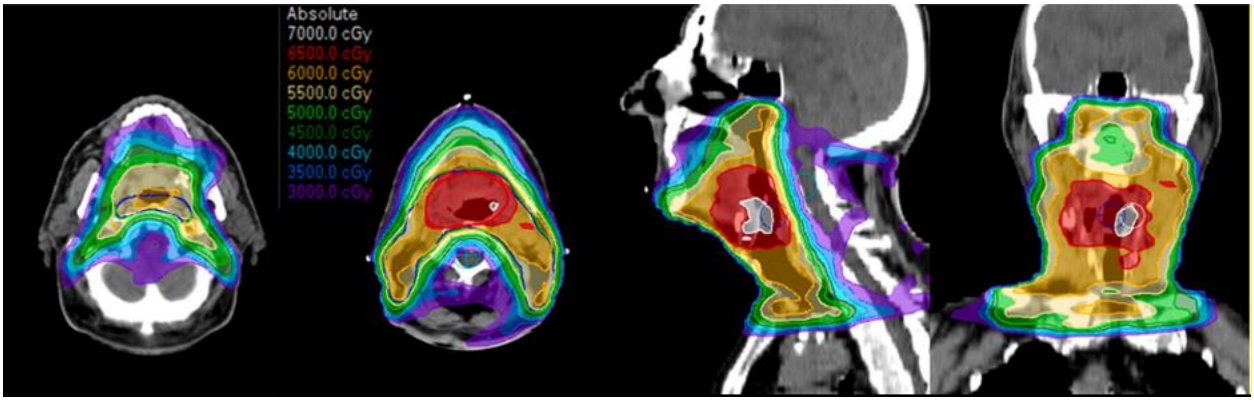


Figure 2-2: Example of a RT plan for treatment of HNC, showing the expected dose distributions in the transversal (left), sagittal (middle), and coronal (right) planes. Image from [16].

When using RT for treatment of HNC, the salivary glands are often considered OARs. However, even when using CFRT or IMRT in the dose planning process, the sparing of SGs is not always possible for all patients as seen in Figure 2-2. The understanding of radiation-induced biologic responses in the SGs is therefore of clinical importance [4].

While the radiosensitivity between cell types is described mathematically by the LQ-model, inter-patient differences are present in their response to similar doses [17]. As the human body is a complex system, macroscopic clinical factors such as overall health and age have great impact on overall survival and the potential for late-effects following RT.

2.1.3 The salivary glands

Saliva is functionally important for various bodily functions such as speech, digestion, control of the body's water balance (absence of saliva causes a thirst sensation), and protecting the teeth

from decaying [18]. Saliva is produced by the salivary glands (SGs) and is composed of water, mucus, proteins, minerals, salts, and the enzyme amylase. Humans have three paired major salivary glands: the parotid gland (PG), sublingual gland (SLG), and submandibular gland (SMG). These exocrine glands secrete either serous (liquid), mucous (slimy and thick substance), or seromucous (a mixture of the two) solutions [19].

The healthy SGs consists of three main cell types: *acinar cells* producing fluids and proteins for saliva (serous or mucus), *ductal cells* making up the transporting ducts for transporting and modifying the saliva, and *myoepithelial cells*. The latter wraps around acinar cells at the terminal ends and the proximal segments of the ducts before appearing to contract, furthering the saliva secretion into lumens (cavity of a tubular structure) of branching networks [20]. Together, the cells make up the generic epithelial architecture of the salivary glands illustrated in Figure 2-3.

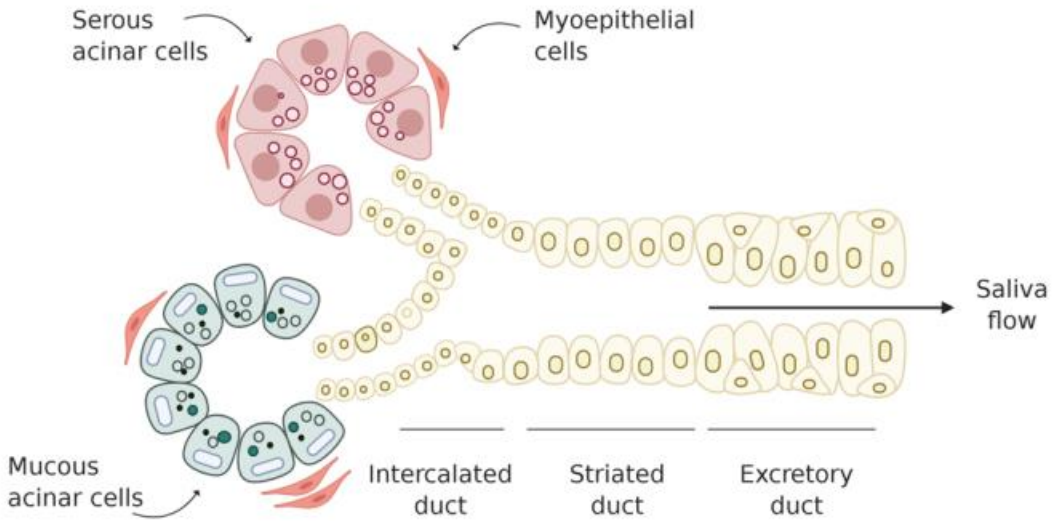


Figure 2-3: The generic epithelial architecture of the salivary glands. Serous or mucous solutions are created by the acini, before being transported into the ductal network for secretion. Myoepithelial cells contracts around the acini furthering saliva flow. Image from [21].

2.1.3.1 Comparing the salivary glands in mice and humans

The SMG consists of a pair of seromucous glands located on the side of the lower jawbone in humans. The produced solution is secreted through excretion ducts known as the submandibular duct (Wharton duct in humans) which opens into a junction at the base of the tongue (sublingual caruncle). While the SMG is the second largest major SG in humans it is the largest among mice making it suitable for translational salivary gland research in mice. While being separate structures in humans, the SL and the SMG are fused in mice [20].

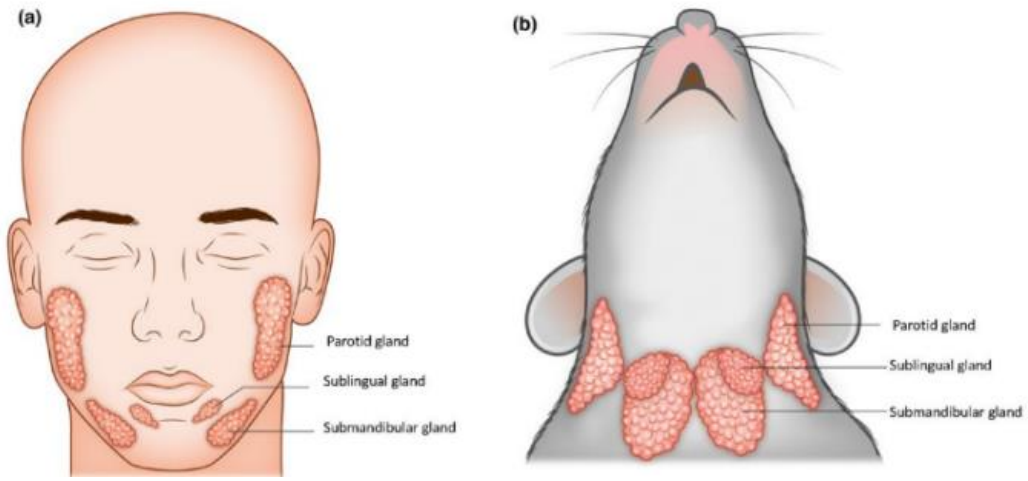


Figure 2-4: Overview of the anatomical position for the salivary glands in humans (left, coronal plane) and mice (right, transversal plane). Image from [20].

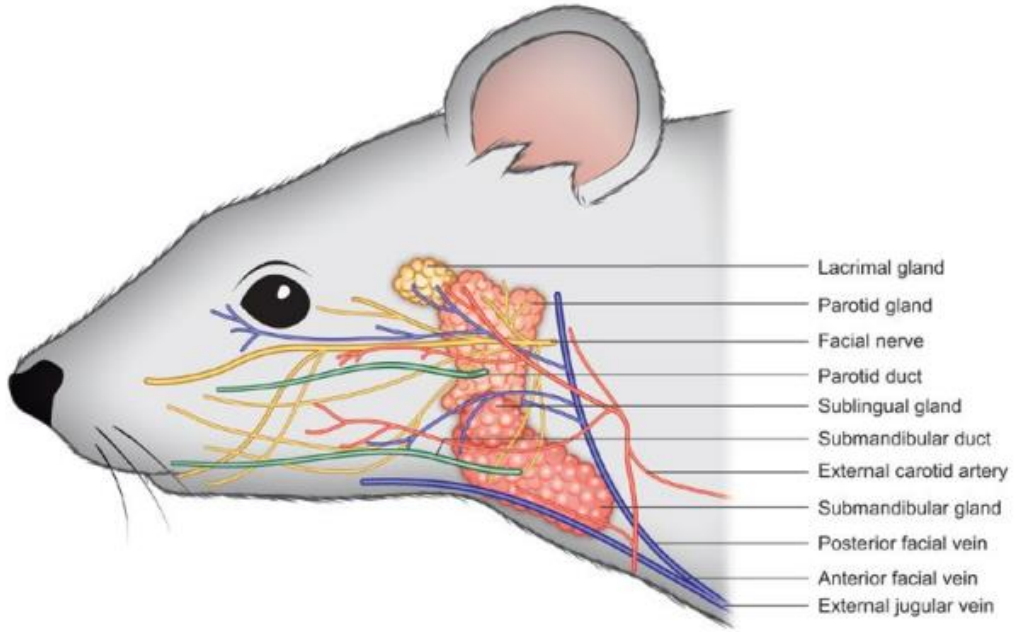


Figure 2-5: Overview of the vascular, nervous, and ductal structures in the HN region of the mouse in the sagittal plane. Image from [20].

2.1.4 Radiation-induced effects on the salivary glands

During RT of HNC the SGs may receive a high dose causing irreversible damage, which may impair their functioning. A common side effect of HNC RT is a reduction in the saliva flow both acutely following RT and as a long-term side effect. Xerostomia (Greek for “dry mouth”) is defined by the Great Norwegian Encyclopaedia as a special case of hyposalivation, having at

least 50% reduction in saliva secretion [22]. This may potentially lead to further severe long-term oral injuries and the quality of life may be severely impacted for HNC survivors. As management of xerostomia rarely is effective, the best measure is prevention [4]. The salivary stem cells mechanisms is a current topic of research, where an increased understanding may have implications for future therapeutic methods to optimize the regenerative abilities of the SGs post-irradiation [21].

A paper from 1992 showed patients having both acute and long-term reduction in saliva in response to RT of HNC, especially for patient receiving doses to the HN region above 45 Gy. If both of the paired SL glands were irradiated patients showed a much higher problem with dryness, compared to only irradiating one [23]. A similar trend for saliva reduction have been observed in rats receiving a single dose of 15 Gy, with a second dip at 180 days maintained at 360 days. Loss of SG mass was also observed in the rats, along changes in epithelial architecture such as a changed proportion between ductal and acinar cells, vacuolization, and interstitial fibrosis (scar tissue in the close-proximity extra-cellular environment) [24].

A 2011 review study article assesses the changes in the SGs post-RT observed in earlier works [25]. Histological analysis confirms the changes in epithelial architecture, with observed loss of acini and increased presence of ductal cells. The increase in ductal cells might either be due to increased proliferation, or a combination of ductal dilations with loss of gland mass leading to an increased number of ducts per unit volume. An increase in adipose (fatty) tissue is observed along interstitial fibrosis, and infiltration of inflammatory cells. The histological changes were reported to be more prominent in the SMGs than the PGs, but the difference in radiosensitivity between the glands is a controversial topic with conflicting results.

By ultrasonographic evaluation the irradiated glandular tissue is seen to exhibit higher heterogeneity relative the healthy tissue, along increased vascular resistance (changes in the blood vessel composition reducing the blood flow). Some evaluations of the SGs by magnetic resonance imaging (MRI) reports a decrease in signal intensity while other studies report an increase, both using T2-weighted MRI (see section 2.2.3). The increase in signal is hypothesized to be related to oedemas (fluid build-ups) damaging the blood and lymph vessels, accumulating interstitial fluids.

2.2 Magnetic resonance imaging (MRI)

Magnetic resonance imaging (MRI) uses non-ionizing radiation in addition a strong magnetic field to produce images with a high contrast between soft tissues of similar densities. This non-invasive imaging technique is an extremely versatile diagnostic tool in medicine and is only transferring heat as the physical by-product in the patient. As the image acquisition process is complex and dependent on an ensemble of factors the unit for pixel intensities in the MR-image are often considered arbitrary.

The following chapters are mainly based on Atle Bjørnerud's compendium in the course FYS-4740: The Physics of Magnetic Resonance Imaging [26].

2.2.1 Nuclear magnetic resonance (NMR) as basis for MRI

The phenomenon known as nuclear magnetic resonance (NMR) provides the basis for all MRI. Proton NMR is the foundation for almost all clinical usage of MRI as hydrogen atoms, having a single proton as its nucleus, are abundant in human tissue consisting of water or fat.

Atomic nuclei with non-zero spin interacts with magnetic fields. If placed in a strong constant external magnetic field $\overrightarrow{B_0}$, the nuclei try to align with the field but due to having non-zero angular momentum they have a precession along the axis of the magnetic field's direction. The precession frequency is referred to as the *Larmor frequency* which follows the linear relationship:

2.2-1.

$$\omega_0 = \gamma B_0$$

The gyromagnetic ratio γ is unique for all nuclear isotopes of non-zero spin, which for the hydrogen isotope of mass number 1 (1H *protium*; the nuclei being a single proton) is $\gamma = 2.68 \times 10^8$ Hz / Tesla.

The wavefunction for a single proton in a static magnetic field consists of a linear combination (superposition) of two possible quantum spin states which is referred to as parallel and anti-parallel states being the low and high energy eigenstates respectively - referencing their alignment possibilities when in a static B-field [27].

While the field interactions affecting the proton spin states is a quantum mechanical phenomenon (each proton wave function collapsing into one of the two aforementioned states when *measured*) the distribution of an ensemble of proton spins affected by the B-field may be described classically using the Boltzmann factor:

2.2-2.

$$\frac{N^+}{N^-} = \exp\left(\frac{\Delta E}{k_B T}\right) = \exp\left(\frac{\gamma \hbar B_0}{k_B T}\right)$$

Where N^+ describes the number of protons being in the low-energy state of spin "up" (parallel to the B-field axis), N^- the number of protons in the high-energy state "down" (anti-parallel). The difference in energy between these energy states is $\Delta E = \gamma \hbar B_0$. $\hbar$ (h-bar) is the reduced Planck constant, k_B is the Boltzmann constant, T is temperature in Kelvin.

The macroscopic (Boltzmann) net magnetization, the sum of all magnetic moment contributions in a sample $\vec{M} = \sum_i \overrightarrow{\mu_i}$ representing the induced polarization of the material, have zero-valued components perpendicular to the static B-field while being in equilibrium due to the incoherence

of the individual magnetic moment precessions cancelling out. Expanding equation 2.2-2. as a first-order Maclaurin series and using that $\mu = \hbar S\gamma = \pm\frac{\gamma\hbar}{2}$ (spin state S is either +1/2 or -1/2) for protons, the component of the magnetization vector parallel to the B-field becomes:

2.2-3.

$$\vec{M} \parallel \overrightarrow{B_0} = \sum \vec{\mu} = N_0\overrightarrow{B_0}\left(\frac{\gamma\hbar}{2}\right)^2 / \, k_B T$$

By sending an electromagnetic pulse in the radio-frequency spectrum (RF-pulse) at the energy $\Delta E = \gamma\hbar B_0$, photons matching the Larmor frequency in equation 2.2-1. will induce a perturbation in the polarized medium. This frequency matching is referred to as *resonance* and "pushes" the magnetization vector from the longitudinal $\overrightarrow{B_0}$-field direction (z) towards the orthogonal plane referred to as the transversal direction (xy). After the pulse the magnetization moves back towards equilibrium, and the change in magnetic flux Φ may be observed as an induced current (Faraday's law of induction) in a receiver coil (Rx) perpendicular to the B0-field direction.

Thus, the potential strength of the measured signal is directly proportional on the strength of the magnetization vector which, following equation 2.2-3., increases with B_0 and γ^2. As 1H are abundant and have the highest γ-factor of signal-yielding isotopes in the human body it is a highly detectable nuclei and is therefore used in all clinical MR-imaging (excluding spectroscopy).

The change in Boltzmann magnetization due to the oscillatory B-field contained in a RF-pulse (B1) in addition to the static B-field (B0) may be derived from the Bloch equation:

2.2-4.

$$\frac{d\vec{M}}{dt} = \gamma\vec{M} \times \vec{B}$$

Immediately following the RF-pulse $\vec{M}$ have some non-zero component in the transversal plane (non-equilibrium), dependent on the angle between $\vec{M}$ and the longitudinal (z) direction known as the *flip angle* α. The observed signal in a Rx-coil perpendicular to z, as $\vec{M}$ moves back towards equilibrium losing its transversal component, is known as the free induction decay (FID). As the individual spin precessions synchronizes and are in phase right after the pulse the summed magnetization vector also have a precession [28], meaning the FID signal oscillates with the Larmor frequency. Due to molecular field variations the exact resonance frequency differs slightly for the signal-yielding protons causing the spins to go back to decoherence. This is known as spin-spin (T2) relaxation. The decay rate of the FID is slightly faster than T2, as it turns out that bulk inhomogeneities in the B0-field (denoted ΔB_0) will amplify the spin dephasing. This measured decay time is known as T2*.

2.2-5.

$$T2^{*} = \left(\frac{1}{T2} + \gamma \Delta B_0\right)^{-1}$$

An additional component of returning to thermal equilibrium, in addition to the loss of transversal magnetization, is the recovery of the longitudinal magnetization. Stimulated emission by having EM-field variations at slightly differing Larmor frequencies cause the excited protons in the high-energy spin down state to dissipate their energy through phonon emissions to the surrounding material (thermal dissipation). This relaxation is therefore referred to as spin-lattice relaxation or simply T1-relaxation.

Solving the Bloch equation (2.2-4.) yields the exponential equations for Mz recovery and Mxy decay with respect to the tissue dependent time constants T1 and T2:

2.2-6.

$$M_{xy}(t) = M_{xy}(0)\exp(-t/T2^{*})$$

$$M_z(t) = M_0\left[1 - \exp\left(-\frac{t}{T1}\right)\right] + M_z(0)\exp\left(-\frac{t}{T1}\right)$$

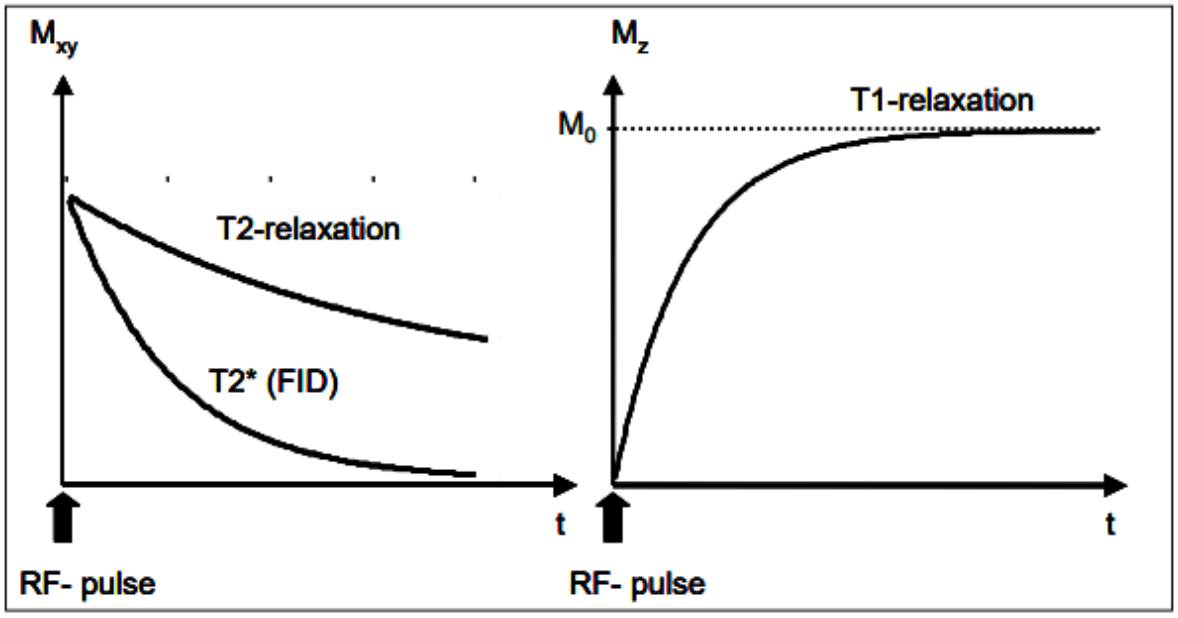


Figure 2-6: Left: decay of transversal magnetization due to spin-spin relaxation, divided into measured (T2, FID) and ideal spin decoherence time (T2, given no bulk inhomogeneities in B_0). Right: Recovery of longitudinal magnetization through spin-lattice (T1) relaxation. Image source: [26].*

2.2.2 Slice selection and k-space

By adding a space-dependent gradient $\vec{G}$ to the static B0-field a certain part of the body may be selected for imaging. The gradient makes the Larmor frequency (equation 2.2-1.) a function of position, such that the resonance condition is only fulfilled within some region determined by the RF pulse. Usually a slice, this region is then the only place where protons become *significantly* affected by the RF-pulse of a certain shape.

The RF pulse, responsible for a time-dependent secondary magnetic field B1, have a bandwidth $\Delta\omega$ resulting in a thickness of the selected slice $\Delta z = \Delta\omega/(\gamma G)$. The gradient's effect on the transverse magnetization due to an excitation pulse, neglecting relaxation effects (the RF pulse duration is only a couple ms compared to T1 and T2 on the time-scale of 100-1000 ms), may be determined using the Bloch equation (2.2-4.):

2.2-7.

$$\frac{d\overrightarrow{M_T}}{dt} = -i\gamma\left(\vec{G}\cdot\vec{r}\right) + i\gamma B_1 M_0$$

The transversal magnetization is now a complex vector $\overrightarrow{M_T} = \overrightarrow{M_x} + i\,\overrightarrow{M_Y}$. Having the RF-pulse last for T seconds, where the gradient $\vec{G}(t) = G_z\hat{z}$ changes polarity after T/2 seconds to correct for phase dispersion, the RF-pulse's effect on the transversal magnetization may be solved by integrating equation 2.2-7.:

2.2-8.

$$M_T(T,z) = iM_0\int_{-k_T}^{k_T}\frac{B_1(k)}{G_z}\exp(ikz)\,dk$$

Using the one-dimensional k-space variable $k \coloneqq \gamma G_z t$ equation 2.2-8. is equivalent to a Fourier transform of B1 [29]. Wanting to keep the RF-pulse effect on M_T as homogenous as possible, i.e. a block function with height $M_0\sin(\alpha)$ within Δz (from -d/2 to d/2), the temporal dependence of B1 (i.e. the shape of the RF pulse signal) is acquired by solving equation 2.2-8. using the inverse Fourier integral:

2.2-9.

$$B_1(t) = G_z\int_{-d/2}^{d/2}\exp(i\gamma G_z t\,z)\,dz = G_z d\,\frac{\sin\left(\gamma G_z t\;d/2\right)}{\gamma G_z t\;d/2}$$

To achieve a perfect block-function shape for M_T the sinc-shaped RF-pulse in equation 2.2-9. would have to span for an infinite amount of time. This is unachievable in practice, meaning the RF-pulse is truncated in time. The transversal magnetization in the excited slice will therefore not be a perfect block function but have some discrepancies, especially at the edges of the slice.

To spatially separate the distribution of signal-yielding 1H nuclei in the selected slice – the spin density $\rho(\vec{r})$, and thus be able to reconstruct the MR image as intensities within some voxel

space, some information for back-tracking the contributing signals must be encoded into the measured signal. This is done using *additional* time-dependent gradients timed with some relation to the excitation pulse (see section 2.2.3). The way such gradients are applied across the RF-pulse and signal acquisition (sampling), differentiates between *frequency-encoding* and *phase-encoding* gradients. The induced phase shift for a certain voxel given a gradient becomes $\alpha(\vec{r},t) = -\gamma \int_0^t \vec{G}(\tau)\cdot\vec{r}\,d\tau$ after the gradient have been on a time t. Defining the k-space variable for multiple dimensions as $\vec{k} = \gamma \int_0^t \vec{G}(\tau)d\tau$ the signal-yielding magnetization M_T may be related to the spin density through the Fourier transform - and vice versa by the inverse transform:

2.2-10.

$$M_T(t) = \iint \rho(\vec{r}) \exp\left(-i\vec{k}\cdot\vec{r}\right) d\vec{r}$$

$$\rho(\vec{r}) = \frac{1}{2\pi} \iint M_T\left(\vec{k}\right) \exp\left(i\left(\vec{k}\cdot\vec{r}\right)\right) d\vec{k}$$

This result is the logical basis for sampling the MRI-signal in *k-space*, before creating the anatomical MR-image as the Fourier transform of the measured signal.

The sampled k-space variable may be interpreted as containing information about spatial frequencies in the acquired MR-image, where the lowest spatial frequencies correspond to the centre of k-space and the highest frequencies to the periphery. Thus, the signal sampled into the central k-space is largely responsible for the contrasts in the image, while the periphery corresponds to high-frequency parts of the image (edges) making up details (see Figure 2-7).

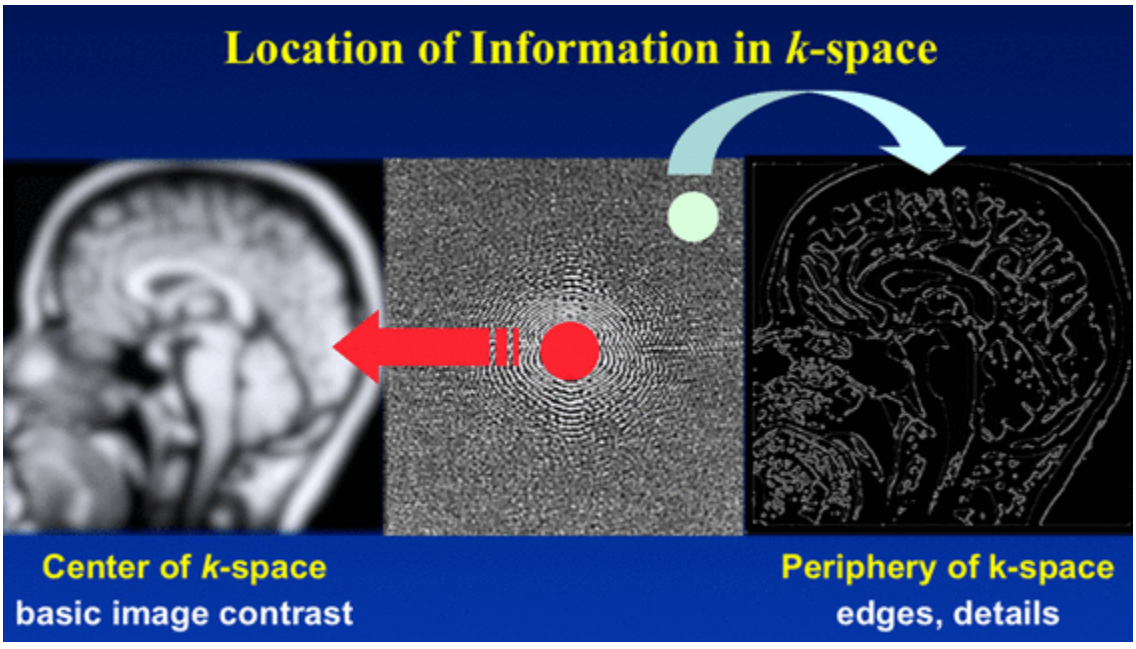


Figure 2-7: Parts of the MRI-signal, sampled into k-space, responsible for contrast (center) and details (periphery) of the MR-image. Image source: [30].

2.2.3 The spin echo pulse sequence and image weighing

In practice two types of spatially encoding gradients (related to the spin density distribution, see section 2.2.2) are often distinguished by how they are designed: to either affect the spin phases, or shift the Larmor frequency. The gradients are independently applied during RF-excitation, during signal sampling, or in-between. Using a gradient to induce a shift in the spin phases right after excitation, before signal acquisition, is referred to as the *phase-encoding gradient* (or *preparation gradient*) [31]. The phase gradient is normally applied with a y-direction dependence, initiating k_y for sampling the acquired signal into k-space. Another gradient is usually applied in the x-direction at the time of acquisition, making the Larmor frequency a linear function of position, relating the k_x value to sample as a function of time. The latter is referred to as the *frequency-encoding* or *read-out* gradient [32].

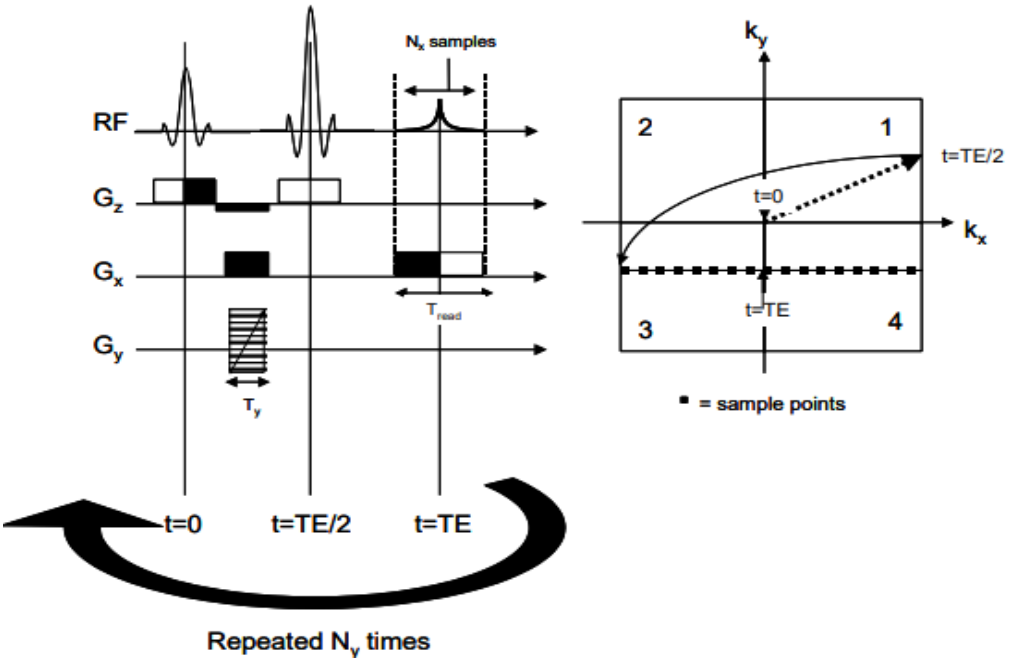


Figure 2-8: The spin echo (SE) sequence. Slice selection is done by having a Gz gradient at both the 90-degree RF pulse for excitation, and the 180-degree pulse responsible for the echo. A phase-encoding gradient Gy in addition to a phase-encoding gradient Gx prepares the signal before sampling all kx values at a ky row during the echo, before re-applying Gx during signal acquisition. This creates a movement to the right in k-space while sampling. Image source: [26].

While one may acquire an MR-image by sending a new RF-pulse for each point in k-space individually, it would be an impractically time-consuming process. To shorten the image acquisition time *pulse sequences* are utilized, enabling a sampling of multiple points in k-space after a single excitation pulse. The spin echo (SE) sequence samples a row in k-space, corresponding to a phase-gradient, for each excitatory RF-pulse. A "echo" of the initial excitation pulse having a 90° flip angle, is created by applying a second RF-pulse with a 180°

flip angle. This causes the spins to rephase a second time after the initial excitation pulse, creating the echo. By applying both a phase-encoding and frequency-encoding gradient before the 180° pulse a point (k_x, k_y) in k-space is initialized. The signal is sampled at the corresponding k_y-row while re-applying a frequency gradient at the read-out – thus sampling all the k_x values at a row corresponding to a single k_y value (Bjørnerud, chapter 6.2 [26]). The SE sequence is illustrated in Figure 2-8.

Biological tissue has varying relaxation times (T1, T2, and T2* as described in section 2.2.1), even while being of similar density, which is the basic fact exploited to achieve contrast between soft tissues in MRI. Depending on the time between each excitation pulse (repetition time TR), and the time from the RF-pulse to the maximal sampled signal at the echo (echo time TE), the longitudinal and transverse components of $\vec{M}$ have reached different values due to tissues having varying relaxation constants (equation 2.2-6.). By applying another excitation pulse before the longitudinal magnetization is back at equilibrium for all tissues, an image contrast between soft tissues based on the relaxation times is achieved. T1 contrast, or weighing (T1w), is achieved by having a long enough TR to ensure tissues with high T1 have not regained full longitudinal magnetization compared to tissues with shorter T1. T2 contrast (T2w) is achieved by having a long enough TE such that the transverse magnetization for some tissues gets close to zero due to spin dephasing compared to tissues having a longer T2.

Concerning the interpretation of MR-images being either T1- or T2-weighted, the T1 images reflects signals from fatty tissue while T2-images reflects signals from both water and fat [33].

A SE sequence *preceded* by a 180-degree pulse is known as inversion recovery (IR), flipping the longitudinal magnetization. The IR sequence may be used for *fat suppression* by applying the 90-degree excitation pulse at the exact time the longitudinal magnetization from the fatty tissue reaches zero during recovery. Assuming fat to have spin-lattice relaxation $T_{1,fat}$, the time to inversion (TI) for fat suppression becomes $TI = \ln(2)\, T_{1,fat}$ [34].

2.2.4 Accelerated k-space sampling by rapid acquisition with refocused echoes (RARE)

As the standard SE sequence only samples a single row in k-space for each RF-pulse, the total acquisition time becomes large when increasing the k-space matrix size or decreasing the slice thickness. By refocusing the echo multiple times after a single excitation pulse, in addition to changing the phase encoding gradient between each echo, multiple rows in k-space may be sampled between each RF-pulse separated by the time TR. The rapid acquisition methods differ mainly in how the multiple echoes are produced. When using gradients for refocusing it is known as Echo Planar imaging, and when using multiple 180° RF-pulses it is known as the fast spin echo (FSE) sequence. The FSE sequence is traditionally known as Rapid Acquisition with Relaxation Enhancement (RARE) [35] (Bjørnerud, chapter 9 [26]).

Concerning the FSE sequence some parameters to notice are the echo train length (ETL) and the *effective* TE (TE_{eff}). The ETL, or the *turbo factor*, is the number of 180° refocusing pulses between each TR – describing the relative decrease in acquisition time relative the standard SE sequence (per definition having ETL = 1). TE_{eff} is the time between the 90° excitation pulse and the echo being sampled closest to the centre of k-space, and is the echo with highest amplitude due to having a minimal (or zero) phase encoding gradient. As the signal sampled into the centre of k-space largely determines the contrast of the image (section 2.2.2) the image weighing in a FSE sequence becomes a function of TE_{eff} (section 2.2.3). It is however important to notice that all rows in k-space contributes to the image contrast in total, and therefore the total weighting. If ETL = N_y (having a k-space matrix of size $N_x \times N_y$), known as a single-shot FSE as all of k-space is sampled following a single excitation pulse, the rows will have largely different TE's. The difference in TE between the rows are all (more or less) contributing to the total weighing of the image, producing an image with some differences in contrast relative a standard SE image.

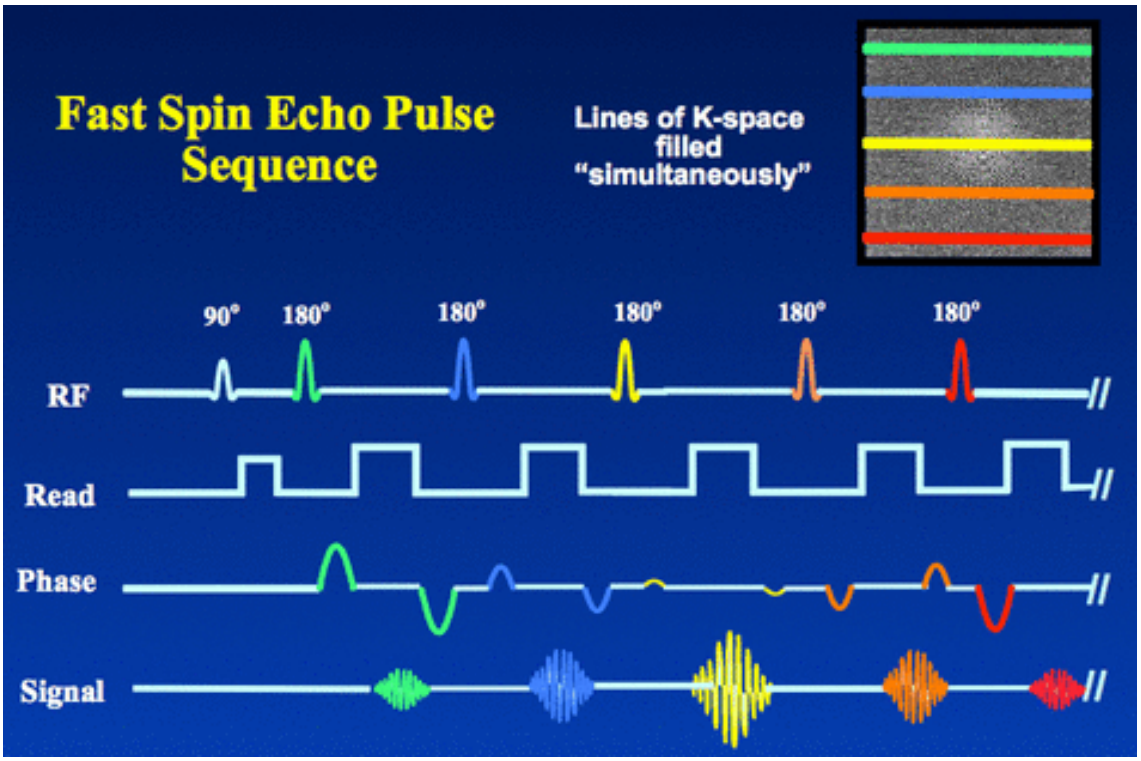


Figure 2-9: Pulse sequences making up the fast spin echo sequence. Following a single 90-degree excitation pulse, multiple 180-degree pulses each creates an echo sampled into various rows in k-space due to varying phase-encoding gradients.

2.2.5 MRI artifacts

The complexity of the MRI acquisition process (including image reconstruction) introduces multiple sources of systematic error. Special cases of such errors are known as *artifacts*. The artifacts may originate from either the patient or the MRI system, and may be visible in the image as geometrical deformations or periodic intensity patterns [36].

Low-frequency periodic error in image intensity is known as a *bias field*, present as changes in image intensities not being related to the imaged tissues. Such an artificial inhomogeneity in the intensities may be caused by one or multiple underlying factors: instrumental factors as inhomogeneities in the static B-field (B0) or the RF B-field (B1), or patient movements during acquisition [37].

Gibbs ringing is an artifact recognized by parallel lines appearing close to high-contrast transitions in the image. In a similar fashion to the truncation of the RF-pulse for slice selection (section 2.2.2) an over- or undershoot occurs at the edge of the discretely sampled signal (as the excited slice is not a perfect block function in time). As lower spatial frequencies in the image are less affected by this sampling truncation, being closer to the "true" signal with less components in the sampled Fourier-series, the artifact is mainly visible at high-contrast edges [38] (Bjørnerud, chapter 5.3 [26]).

Very small perturbations in the effective B-field (ΔB) relative the voxel sizes may cause a dephasing of the magnetization within a single image voxel. This loss of signal is known as intra-voxel dephasing. Similarly, having very large perturbations in the B-field within a voxel may shift the effective B-field to not correspond with the assumed resonance conditions, known as off-resonance effects. Both cases are responsible for *susceptibility artifacts*, which in practice may be caused by a combination of the two cases (Bjørnerud, chapter 12.2 [26]). Compounds in the anatomy affecting the B-field may for example be weak diamagnets (water), or strong ferromagnets (iron in braces). The susceptibility artifacts are usually seen as high-intensities in some areas of the image, or a vanishing signal [39].

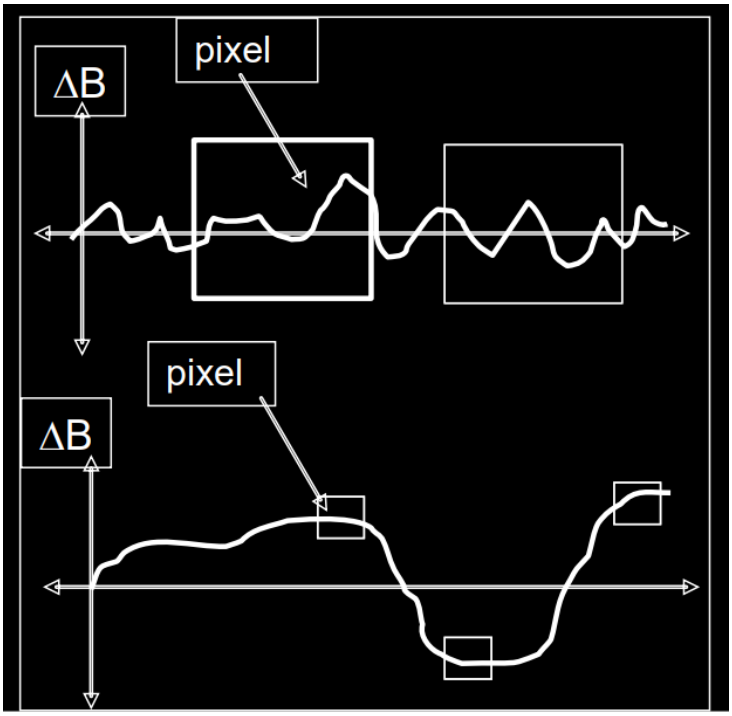


Figure 2-10 Top: Intra-voxel dephasing due to big voxel size (dependent on bandwidth) compared to local field perturbations. Bottom: opposite case causing off-resonance effects. (Source: Bjørnerud, chapter 12 [26])

2.3 Radiomics

Radiomics utilizes machine-learning (ML) concepts to extract and find quantitative information from medical images such as computed tomography (CT) images or MRI, with some relation to a biological phenomenon. Cancer treatment is heading towards a more personalised approach (precision oncology), taking more patient variations into account. While genetic and biochemical markers are considered the driving forces behind precision oncology, radiomic features have shown promise in this context and are increasingly considered in clinical studies [6].

From a two- or three-dimensional *region of interest* (ROI, or *volume of interest* - VOI) *image features* are extracted which may contain phenotypic information (*imaging biomarkers*) that is invisible to the naked eye [40]. As the radiomic features are based on non-invasive medical images, identification of robust imaging biomarkers may limit the use of invasive techniques such as biopsy based molecular assays in the clinical setting [9].

How the features are extracted from the images, the core of the radiomic process, may broadly divide radiomics into two categories: hand-crafted radiomics, or deep radiomics [6]. Where hand-crafted radiomics rely on well-defined mathematical operations and matrices, deep radiomics utilizes neural networks for unsupervised extraction. Due to the flexible nature of deep neural networks the features may take on almost any perceivable shape based on the weighted connections within the neural network, but are therefore harder to interpret than hand-crafted features.

Radiomics is a high-throughput method, providing many features per image - usually a much larger number of features than sample size. The *feature selection* process is therefore integral for identification of important features having some relation to the research question (or clinical decision). Using the selected features the last part of the radiomics pipeline considers predictive modelling of some biological or clinically interesting outcome [41].

Radiomic image features have been shown to reflect the intratumor heterogeneity on a cellular level [42] which have been strongly correlated to patient survival and tumour control (section 2.1.2) [43]. Radiomic studies have been published with an increasing frequency the last 7 years. Radiomic studies in oncology considers features extracted from tumours in the brain, head and neck region, lungs, breasts, gastrointestinal, cervix, prostate, and colon [6]. Changes in radiomic features over time, *delta radiomics*, have shown predictive abilities of outcomes in lung cancer patients such as survival and distant metastases [44]. Delta-radiomics have also been shown to improve the prediction of late xerostomia (section 2.1.4) after RT of HNC using shape-based features extracted from the parotid glands [10].

A big issue within radiomics is the reproducibility and replicability of results, as the majority of radiomic features have been shown to have stability issues. This is often accredited to scanner differences, acquisition protocols, and the software used for feature extraction [45]. What steps, and their order, included in the preprocessing of MR-acquired images have been shown to affect

feature repeatability (stability of features acquired from the same set of data) [46], [47]. The Image Biomarker Standardization Initiative (IBSI) collaboration initiative attempts to address said issues by standardizing the computation workflow (feature extraction) for a set of 169 radiomic features (Figure 2-11)[7]. An extension of this initiative (IBSI 2) attempts to include the standardization of convolutional image filters applied before feature extraction, which may highlight specific characteristics in the images. IBSI 2 is a work still in process [48].

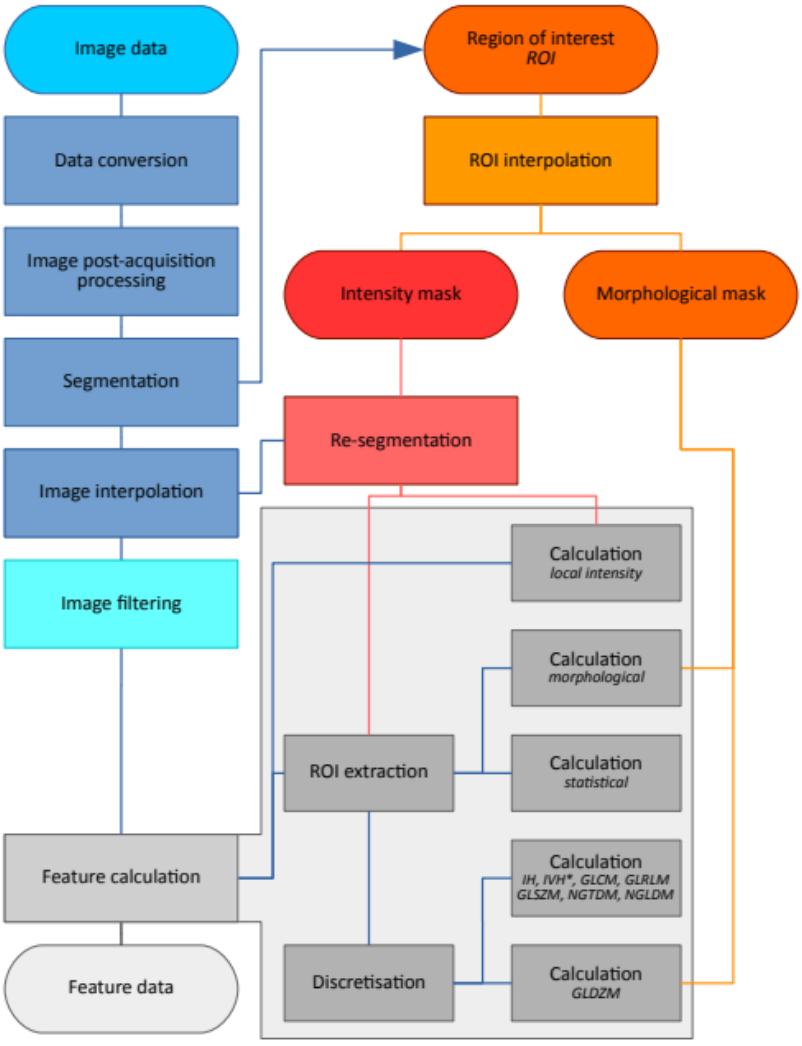


Figure 2-11: The workflow proposed by the second iteration of the image biomarker standardization initiative (IBSI 2) for calculation of features of varying types [48] .

For single-channel 2D images consisting of $M \times N$ pixels, the extraction of d image features per image may be seen as a *dimensionality reduction* from the image space to the feature space (assuming $M \times N > d$). The hand-crafted radiomic features attempts to capture relevant information from the original image, while being interpretable in relation to phenotypic

properties for the tissues / organ delineated by the ROI. The three main classes of hand-crafted image features in radiomics following IBSI are *shape-based*, *first-order statistics*, and *texture-based* [7].

2.3.1 First-order features

Let the grey-levels intensities corresponding to the N_v pixels in the ROI of the image be denoted as $X_{gl} = \{X_{gl,1},\ X_{gl,2}, \dots ,\ X_{gl,N_v}\}$. Features calculated directly from this set are the *intensity-based statistical features*, or *first-order* features, describing the various properties of the image intensity distribution [49].

First-order features include the intensity mean, median, variance, minimum, maximum, and the coefficient of variation (CV). Features based on *percentile values* in X_{gl} include the inter-quartile range (IQR), quartile coefficient of dispersion, and the robust mean absolute deviation. The shape of the intensity distribution is described by features such as the kurtosis or skewness – representing the flatness and asymmetry of the distribution, respectively [50].

Other first-order features require the creation of an *intensity histogram* for calculation. The continuous intensity distribution, from which X_{gl} may be assumed to be sampled from, is divided into N_g bins providing a *discretized* set of gray levels $X_d = \{X_{d,1}, X_{d,2}, \dots , X_{d,N_g}\}$. The frequency of each value from X_d within the discretized image is the histogram. Examples of first-order *intensity histogram features* requiring such a discretization before calculation include the entropy and uniformity.

2.3.2 Shape-based features

The shape-based features represent the morphologic characteristics of the ROI, such as surface area, sphericity, and lengths along the major and minor axis or their proportion (elongation). The features are calculated only using the ROI (denoted morphological mask in Figure 2-11), without any dependence on image intensities.

2.3.3 Texture-based features

Sometimes referred to as the *second-order statistics*, texture-based radiomic features describe the spatial relations and patterns found between pixel intensities in the ROI [50]. The radiomic framework presented here considers texture features calculated from five different matrices. Each texture matrix is calculated from the discretized set of N_g intensity values in the ROI, following their definitions in IBSI [49].

The *gray level co-occurrence matrix* (GLCM) counts the occurrences of each pair of discretized intensity values at a pixel-distance δ in the direction θ. Entry (i, j) in the GLCM is the count of

how many times in the image the discretized intensity $X_{d,j}$ appears δ pixels away from $X_{d,i}$ in direction θ. The set of possibilities for θ is dependent on δ by its connectivity: for a 2D image having $\delta = 1$ each pixel has 8 neighbours and thus possibilities for θ, each corresponding to a differently calculated GLCM [51]. The shape of the GLCM is dependent on the number of discretized gray levels, with dimensions $N_g \times N_g$.

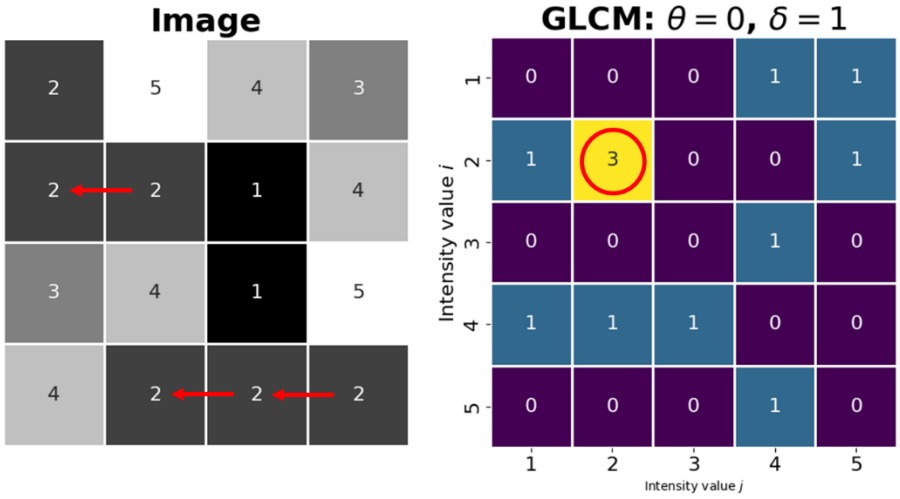


Figure 2-12: Calculating the GLCM for a discretized image of $N_g = 5$ gray levels (left), resulting in a 5×5 GLCM (right). Having 3 instances of gray level $i = 2$ neighbouring the same value ($j = 2$) at distance $\delta = 1$ in the right horizontal direction $\theta = 0$, yields entry $(i, j) = 3$ in the GLCM. Image created by the author using the python package pyRadiomics [52].

The *gray level run length matrix* (GLRLM) quantifies the length of consecutive gray levels (runs) in the direction θ. The (i, j)'th element of GLRLM is a count of how many times gray level $X_{d,i}$ appears next to itself j times in direction θ [53].

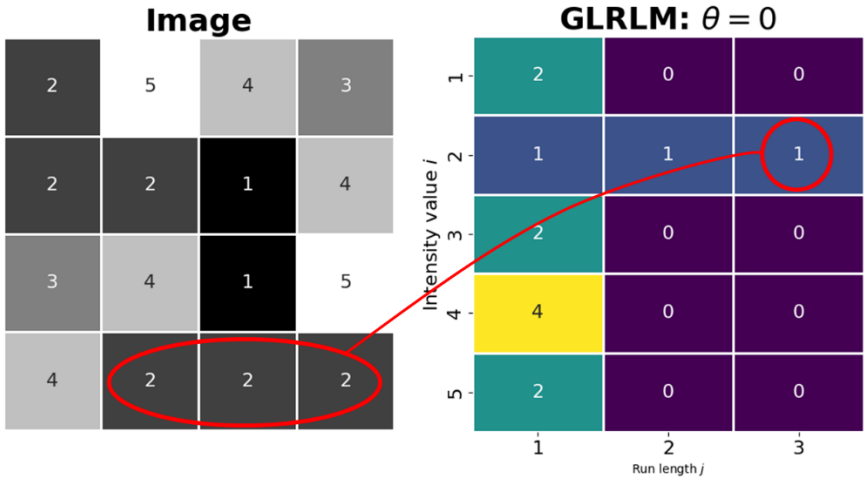


Figure 2-13: GLRLM entry $(2, 3) = 1$ represents the single instance of three consecutive appearances of gray level $i = 2$ in the right ($\theta = 0$) direction.

Defining a zone as the number of connected pixels of the same gray level in all ordinal (8-connected) directions, the *gray level size zone matrix* (GLSZM) quantifies the number of pixels in each zone for a gray value $X_{d,i}$. The matrix is rotationally invariant, i.e. no θ-dependence for the entries, in contrast to the GLCM and the GLRLM [54].

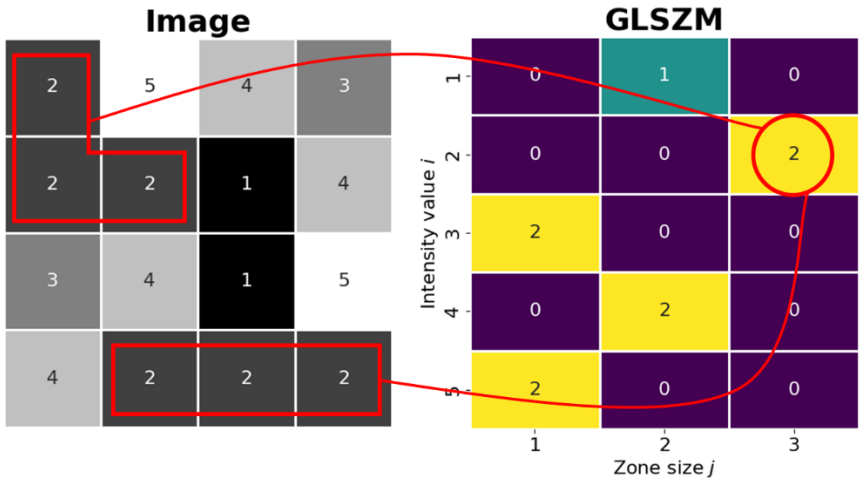


Figure 2-14: GLSZM entry $(2,3) = 2$ represents the two instances in the image where gray level $i = 2$ creates a zone of size $j = 3$.

The *gray level dependency matrix* (GLDM) counts the number of pixels at a distance δ considered dependent, in all directions. Pixel i is defined as dependent on a neighbouring pixel j if $|X_{d,i} - X_{d,j}| \leq \alpha$. Entry (i,j) in the GLDM counts the number of instances where gray level i is dependent on j neighbouring pixels in the image [55].

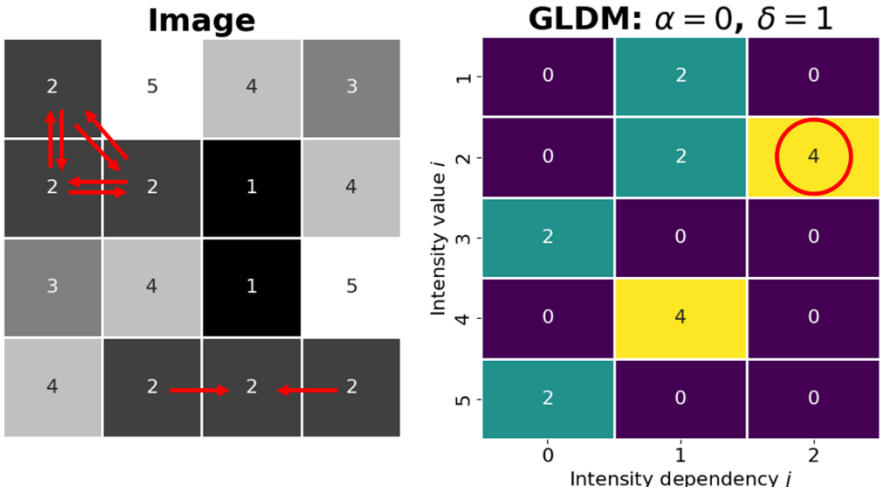


Figure 2-15: GLDM entry $(2,2) = 4$, representing the four instances where gray level $i = 2$ is dependent ($\alpha = 0$) on $j = 2$ neighbours in the nearest vicinity ($\delta = 1$).

Introduced as an alternative to the GLCM, the *neighbouring gray tone difference matrix* (NGTDM) calculates the sum of absolute differences between each discretized pixel of gray level i and the corresponding average of the neighbouring pixels at distance δ [56]. The proportion n_i is the number of instances of gray level i in the ROI, $p_i = n_i/N_v$ is the fraction of said instances, and s_i is the sum of absolute differences between $X_{d,i}$ and their average neighbouring values as seen in Figure 2-16. Features calculated from the NGTDM include coarseness, contrast, busyness, and complexity.

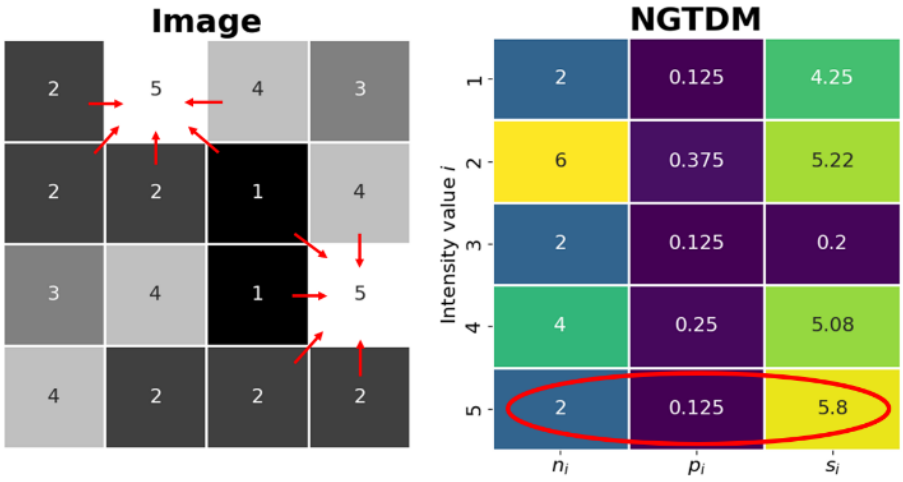


Figure 2-16: Illustration of the NGTDM calculation from a discretized image having 5 gray levels. Gray level $i = 5$ have $n_5 = 2$ instances, yielding the fraction $p_5 = \frac{2}{16} = 0.125$. Their corresponding neighbours at $\delta = 1$ is shown as red arrows, making $s_5 = \left|5 - \frac{2+2+2+1+4}{5}\right| + \left|5 - \frac{4+1+1+2+2}{5}\right| = 2.8 + 3 = 5.8$.

2.3.4 Image filtering in radiomics

All features, except shape-based, may reveal new or amplify existing relationships within the image by calculation after applying some *image filter*. Such features may be referred to as "higher-order statistics" [50], but will for simplicity here be referred to as first-order or texture features calculated after filtering. The following filter definitions are based on the pyRadiomics package [52], unless other is specified.

Some filters which are easy to define using simple mathematics, are seen in equation 2.3-1. The square root and logarithm filters assume $x \geq 0$, having slight variations when $x < 0$ to avoid complex pixel intensities after filtering.

2.3-1.

$$\text{Square filter:} \quad f(x) = (cx)^2 \quad \text{where} \quad c = (\max(|x|))^{-\frac{1}{2}}$$

$$\text{Square root filter:} \quad f(x) = \sqrt{cx} \quad \text{where} \quad c = \max(|x|)$$

$$\text{Logarithm:} \quad f(x) = c\log(x+1) \quad \text{where} \quad c = \frac{\max(|x|)}{\log(\max(|x|)+1)}$$

$$\text{Exponential:} \quad f(x) = e^{cx} \quad \text{where} \quad c = \frac{\log(\max(|x|))}{\max(|x|)}$$

The gradient filter (∇f) describe the rate of change in pixel intensities in across the image, emphasizing edges and details. The gradient filter is further defined in section 3.2.2.

The local binary pattern of an image is a rotationally invariant texture descriptor, emphasizing the relationships between pixels in some neighbourhood. Given P neighbours surrounding the center pixel g_c at a distance R, the local binary pattern becomes:

2.3-2.

$$LBP(P,R) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$$

Where $s(\ldots)$ is the sign operator, evaluating the differences between the center pixel to each neighbour [57].

The wavelet filter is in some sense similar to the Fourier transform, based on a scaling and time-shift factor instead of frequency. The wavelet transform decomposes a signal into lower resolution at multiple levels [58]. In the pyRadiomics package [52], the 2D wavelet filter is differentiated by a high (H) and low (L) version.

2.3.5 Feature selection

The second grand step in the radiomics pipeline, after the image features have been extracted, is the selection of the most relevant features with respect to the research question and outcome. Due to the high quantity of radiomic features extracted per image, usually much more than the number of images in total, this step of dimensionality reduction is critical. The optimal feature selection method reduces the feature set into the smallest possible subset of uncorrelated features, maintaining the highest possible amount of relevant information while excluding the noise. A smaller subset of features reduces the complexity of the models, lowering bias (chance of overfitting, see section 2.4) but increases the interpretability of the models.

In a clinical setting is the identification of the most relevant and robust radiomic features (as potential imaging biomarkers) an important part of radiomic research, in addition to predictive modelling [6].

Using patterns within the feature-space without any relation to outcome for feature selection, is denoted *unsupervised* selection. Examples include multi-dimensional clustering methods as k-means, or principal component analysis (PCA) reducing the feature-set to linear combinations of features maximizing the explained variability [59].

Utilizing some outcome variable in the selection process is known as *supervised* feature selection and may be broadly divided into three categories: *filter, wrapper*, and *embedded* selection methods. Filter methods becomes a part of the preprocessing independent of the machine-learning methods used for modelling, e.g. by ranking the variables univariately by their correlation to the outcome. Wrappers use a specific model to rank the predictive abilities of each feature in relation to some outcome, e.g. by scoring the features univariately by the coefficient of determination (R2, section 2.4.6) following a simple linear regression to a continuous outcome. Embedded methods select the best features within the training process in a ML modelling framework, effectively combining the filtering and wrapper methods. Embedded methods may use a model with some regularization parameters for iterative elimination of unimportant features, such as the LASSO operator or random forests [60]. Random forests are further discussed in detail in section 2.4.3.

While correlation is often used as a metric to describe redundancy in a feature-space, a high correlation between features does not necessarily mean they do not contain complementary information which may be of interest [60] (except of course when having perfectly correlated features).

2.3.6 Earlier work using radiomic features to predict xerostomia

Models predicting xerostomia as a late response following RT of HNC have been in use for many years as the normal tissue complication probability (NTCP) model, assessing patient risk for side-effects (section 2.1.2). NTCP-models for xerostomia utilize dose-volume histogram (DVH) parameters in addition to patient reporting, often being baseline information during dose planning [61]. In a 2018 study van Dijk et al. incorporated more patient-specific responses to the RT, obtained during the treatment in addition to baseline [10]. Using radiomic features extracted from the parotid glands (PGs) in CT images, the relative change from baseline in each feature was calculated at various time-points (delta-radiomics). The delta-features were used to predict xerostomia 12 months after RT, and shape-based features were found to be the strongest predictors (specifically, the relative change in the PG area between baseline and after 3 weeks). However, a 2022 replication study did not manage to obtain univariate significance for the shape features - emphasizing the need for external validation before claiming a generalized result in radiomics-based research [62]. The replication study found instead the maximum Hounsfield units (HU) value within the ROI to be significant for prediction of late xerostomia in patients with intact SMGs.

Another 2018 study used CT-based radiomics on the PGs to create probability models for xerostomia 6 months post-treatment, where clinical variables such as smoking, age, and total delivered dose was included in the model [63]. 5 radiomic features were selected by univariate spearman correlation to the outcome with a threshold at 0.7, being texture-based features in addition to VOI size. Three logistic regression models were created using baseline features, mid-RT delta-features, and a model emphasizing the temporal trajectory of the features. While models using only using image features performed worse than the clinical variables, models combining both performed the best – indicating the radiomic features contained additional relevant information. A third 2018 study also studying PG-features from CT images found organ- and dose-shape features to improve NTCP-models for prediction of xerostomia at multiple time-points post-treatment (0-6 months, 6-15 months, or 15-24 months) [64].

As mentioned in section 2.1.4 the tissue of the salivary glands (SGs) has been shown to become more heterogenous after ionizing irradiation. While not denoted as radiomic research, earlier studies have investigated whether textural image features have any relation to the probability of developing complications in the SGs post-RT. A study by van Dijk in 2016 showed that textural biomarkers calculated from CT images of the PGs and SMGs inhibited predictive performance for xerostomia [61]. While the maximum CT intensity was the best predictor from the SMG (being a first-order feature), the short run emphasis (SRE) calculated from the GLRLM in the PG was shown to increase with heterogeneity. By visual inspection the high-SRE images was assumed to be related to fat saturation in the PGs after irradiation, affecting heterogeneity. Another study concerning HNC patients treated with IMRT, analyzed CT images and found the GLCM correlation and GLRLM run-length non-uniformity to be significantly different for xerostomic patients [65]. The study hypothesized a relation between the texture-features and an increased radiosensitivity, reduction in vascularization, or an increase in adipose tissue.

2.4 Data modelling and statistical learning

Being the foundational framework in machine learning (ML), statistical learning utilizes methods from both statistics and functional analysis. It is broadly categorized into supervised and unsupervised learning. Supervised learning relates the output to the input data using statistical modelling which estimate, or predict, the outcome based on the input. Unsupervised learning have no such “a priori” knowledge of what the input data (*all* the data in the unsupervised case) should relate to, and instead tries to find some underlying patterns within the data [66].

The input data is referred to as input *variables*, *predictors*, *independent variables*, or *features*. In supervised learning the outcome data is referred to as *output*, *outcome*, or *response variables*. It is generally assumed that the output Y follows some pattern relating to the input X, but not in a perfect manner due to noise in the outcome (denoted ϵ in equation 2.4-1)

2.4-1.

$$Y = f(X) + \epsilon$$

Supervised statistical learning is largely concerned with finding a *response function* f, used to *estimate* the response $\hat{Y} = \hat{f}(X)$. Whether the output is a *continuous variable* or restricted to *discrete* values representing *class labels* differentiates this search for f into *regression* and *classification* problems. Assuming the response function to follow some specific functional form (e.g. linear, polynomial, or logistic functions) necessitates the estimation of some *functional parameters*, while other methods makes no such assumption and is therefore known as *non-parametric methods* (e.g. b-splines or decision trees) [66].

When evaluating how well the predictors match the response data for a given estimated function, some quantitative metric is needed to assign the *quality of fit*. For regression problems the *mean squared error* is a commonly used metric, found by summing the average squared differences between the predicted and observed outcome values for all n instances in the data set (equation 2.4-2).

2.4-2.

$$MSE := \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

For classification tasks the *error rate* is defined, using the indicator variable I, as:

2.4-3.

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1 \text{ if } y_i \neq \hat{y}_i \\ 0 \text{ if } y_i = \hat{y}_i \end{cases}$$

$$\text{Error rate} := \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Using a subset of the total available data in the search for the response function $\hat{f}$, while preserving the remaining data for testing, is known as *train test splitting*. Calculating an error metric using the test data is known as *test error*, or *generalization error* as it describes the model's performance on unseen instances. The generalization error may be divided into three contributing factors: *bias*, *variance*, and the *irreducible error*. While the irreducible error relates to the variabilities contained in the data itself (e.g. a broken sensor or natural underlying variations) the balancing between bias and variance is important when choosing and tuning the modelling framework – referred to as the *bias variance trade-off* [67]. Having a higher bias, thus lower variance, imply the model produces more similar and less accurate results when applied to varying data. This is known as *underfitting*. When underfitting the underlying pattern in the data is not properly captured by the model, which may be due to choosing a model with a too low *complexity* compared to the patterns (such as fitting a linear curve to a periodic signal). *Overfitting* is the other end of the trade-off, meaning the model is overly sensitive to small

changes in the data and is characterized by higher variance and lower bias. A model overfitting the data may seem very accurate when evaluated on the training data but will quickly produce less accurate estimates when evaluated on the unseen test data. Choosing the proper model is therefore an optimization problem between these two extremes where the loss of generality due to high variance is not sacrificed for loss of useful information (high bias) – illustrated as the stippled vertical line in Figure 2-17.

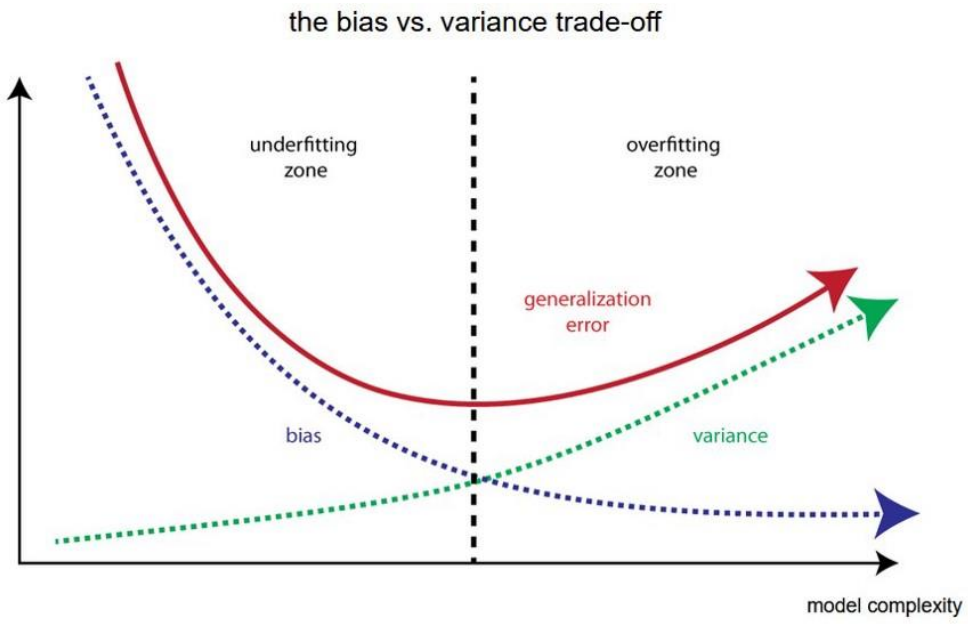


Figure 2-17: Illustration of the bias-variance trade-off. The model variance increases, while the bias decreases, with increased model complexity. The vertical black line represents the optimal model complexity. Image from [68].

2.4.1 Validation methods using cross validation

As discussed in the previous chapter the calculated error for the test and training sets may yield different results, more so if the model is overfitting the data. Methods exists to address this discrepancy and are often more relevant if there is little available data for study. While mathematical methods exist to penalize the error estimate for overfitting (*regularization*) different ways of dividing the available data into training and test sets may also address this issue. If a subset of the training data is held out during the model training, it may be used as a more accurate generalization error estimate for validation of model choice or tuning of model parameters (*hyperparameter tuning*) and is therefore known as the *hold-out* or *validation set*. The *validation error* tends to overestimate the test error and have a high variation, issues which *cross validation* (CV) methods try to mitigate.

K-fold cross validation divides the training data into k approximately equally sized subsets (folds), where k-1 of the folds is used for training and the remaining fold estimates the validation

error. This process is repeated k times such that each fold is used as the validation set once, and the final CV error is calculated as the average of the k error estimates. Simply choosing k = 2 corresponds to the single train / validation split, while k = n (where n is the training sample size) is known as leave-one-out cross validation (LOOCV).

The choice of k is also related to the bias variance trade-off as the training sample size affects the model's bias. LOOCV have a basically unchanged training sample size ($n - 1$) and may therefore be considered to have low bias (see Figure 2-17), while the opposite is true for a single validation split. A choice of $k = 5$ or $k = 10$ is common, which empirically have been shown to lead to an appropriate amount of model bias [66].

The whole k-fold CV process may be repeated multiple times before averaging the results, leading to *repeated cross validation*. Balancing the folds by having similar proportions of some categorical variable (e.g. a binary outcome) is known as *stratified cross validation* [67].

2.4.2 Bootstrapping and bagging

Being an extremely versatile resampling method, the *bootstrap* is a commonly used tool in statistical learning. Having N samples in the original dataset B new resampled sets are created each containing N samples. For each picked observation from the original set to the resampled set the pick is "returned" to the original set, such that instances occurring only once in the original set may appear multiple times in the resampled set. This is known as resampling *with replacement* [66]. As this resampling from the original set mimics drawing new samples from the original population one may argue that the method yields a more accurate picture of the underlying distribution from which the original data is assumed to be drawn – increasing the accuracy of the computed statistic of interest [69].

In a machine learning approach this resampling method may be used to train B independent predictors – one for each bootstrapped set from the training data. The predicted response to an observation is found by averaging over the predictions from the B predictors in the regression case or choosing the most predicted class (*majority vote*) when performing classification. The method of using such an *ensemble* of bootstrapped predictors to achieve a single prediction is known as bootstrap aggregation – or simply *bagging* [70]. This method is practical among models trending towards high variance as averaging a set of observations (or in this case predictions) reduces the variance [66].

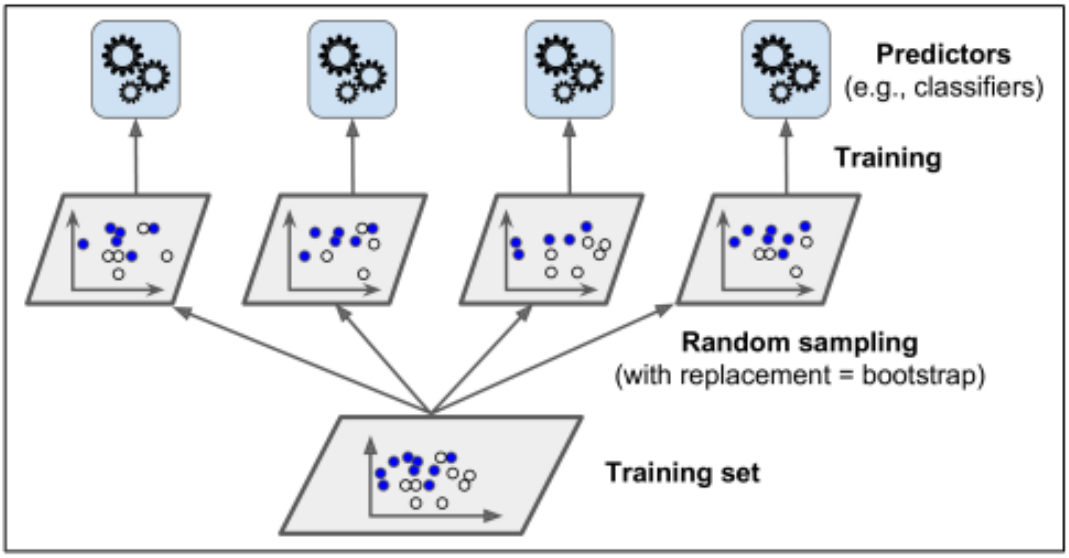


Figure 2-18: Illustrating the bootstrap aggregating (bagging) technique, from [67]. B bootstrapped datasets are created using the original training set, used for training a similar number of models (ensemble).

As each bootstrap sample on average contains about $2/3$ of the original n observations, the remaining $n/3$ observations (known as the out-of-bag (OOB) set) may be used for evaluation of the corresponding bagging predictor – giving an estimate for the bagged model accuracy without the need for evaluating on the testing data. For an instance i in the original training data all predictors having this instance in the OOB set, the $\sim B/3$ predictors not using instance i for predictor training, are used in the aforementioned aggregated way to make a prediction for the instance. Calculating the MSE (for regression, equation 2.4-2.) or classification error (equation 2.4-3.) for the n OOB predictions yields the out-of-bag error (OOB error).

2.4.3 Tree-based methods

Decision trees are the basis for a group of non-parametric modelling techniques which may capture complex non-linear relationships between multiple input variables and the outcome space. The whole decision space may be visualized as a graph, known as a "tree". The predictor space is divided into regions by binary decision boundaries, such as a threshold value for a continuous variable, at the *internal nodes* (vertices in graph theory). Connected by *branches* (edges), the internal nodes keep dividing the predictor space into subregions until the final node representing the predicted outcome is reached – the *terminal node* (or *leaf*) [66]. Such trees may be used for either regression to a continuous outcome variable or classification of discrete outcomes (class labels).

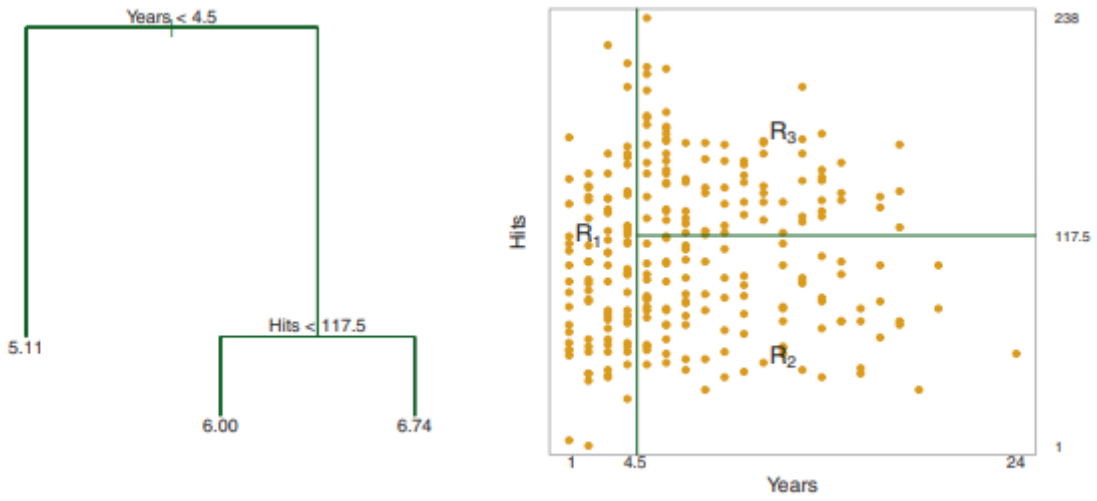


Figure 2-19: Single decision tree shown as a tree-graph visualization (left) with its decision boundaries and resulting subregions in the predictor space (right). Image from [66].

2.4.3.1 Regression Trees

Dividing the predictor space into J high-dimensional rectangular partitions (*boxes*), the optimized subregions $R_1, R_2, .., R_J$ are found by minimizing the residual sum of squares (RSS):

2.4-4.

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_J} \left(y_i - \hat{y}_{R_j}\right)^2$$

y_i is the observed outcome, and $\hat{y}_{R_j}$ is the mean response values from the training data within box R_j. As considering all possible subregions would be computationally demanding, the trees are usually trained using a top-down approach known as *recursive binary splitting*. Beginning with a single *best* split of the input space, the top of the tree, the remaining regions are successively split creating the branches - moving down the decision tree. Considering all the p predictors with their corresponding cut-off values, the combination yielding the lowest RSS is chosen. The process is repeated, considering one of the previously split regions instead of the entire predictor space, in a recursive manner J partitions are reached.

While the trained decision trees may give good predictions onto the outcome space, the trees quickly becomes highly complex and therefore prone to overfitting [66]. To mitigate this a large tree T_0 is initially made before it is *pruned* back to a less complex *sub-tree*. The process of *cost-complexity pruning* involves adding a regularization parameter α to the optimization criterion (RSS in equation 2.4-4.). The values for α are used to find corresponding optimal sub-trees given

this regularization criterion, of which the best may be chosen by k-fold cross validation on the training data.

2.4.3.2 Classification Trees

While regression trees predict outcomes which may have "any" continuous value, the classification tree predictions are restricted to a discrete set of certain values – labels representing the possible outcome classes. *Binary classification* is the simplest case where the outcome variable is either true ($\hat{y} = 1$) or false ($\hat{y} = 0$). For an observation within some input space subregion (defined by the decision boundaries) the predicted class label is simply the label with the most occurrences along the training data within this region. The classification trees are grown similarly as regression trees using the recursive binary splitting technique described in chapter 2.4.3.1, but with some other optimization criterion than the RSS. One of two standard criterion measures is normally used: either the *Gini index* or the *cross-entropy*. Having K classes the measures are defined by the following equations [66]:

2.4-5.

$$\text{Gini index: } G = \sum_{k=1}^{K} \hat{p}_{mk}\,(1 - \hat{p}_{mk})$$

$$\text{Cross entropy: } D = -\sum_{k=1}^{K} \hat{p}_{mk} \log\,(\hat{p}_{mk})$$

$\hat{p}_{mk}$ is the fraction of training samples within subregion m belonging to class k. One may interpret the Gini index as a measure of the total variance across all K classes, referred to as *node purity*. As the Gini index becomes small if $\hat{p}_{mk}$ is close to either 1 or 0, almost all cases at the terminal node falls into a single category (i.e. the decision tree is *pure*). The cross entropy goes to zero as $\hat{p}_{mk}$ approaches 1 or 0, similarly to the Gini index, but using cross entropy as the tree growing criterion tends to yield more balanced trees while the Gini index trends towards isolating the most frequent class in its branches of the tree [67]. If prediction accuracy is the most important goal for the tree-based model one may alternatively use the *classification error* $E = 1 - \max_{\text{k}}(\hat{p}_{mk})$ for pruning.

2.4.3.3 Bagged trees and random forests

As the decision tree methods discussed in the previous chapters are known to have high variance (see bias variance trade-off in chapter 2.4) a natural extension is to consider many trees before averaging the resulting predictions. If each tree is trained on a bootstrapped sample from the original dataset the resulting *ensemble predictor* is known as *bagged trees* (see section 2.4.2). While growing a single decision tree to high complexity may lead to loss of generalization, the

ensemble method allows for higher complexity in the individual trees with less effect on the generalization – leading more robust models [71].

An improvement to this ensemble method is made by *restricting* the predictors considered per tree to a subspace of the total amount of predictors, with a different subspace of similar sizes for each tree. This improved ensemble method is known as a *random forest* and lowers the correlation between the trees allowing for a higher complexity per tree without loss of generality in the ensemble. Having p predictors in total the size of the predictor subset m is normally chosen to be some function of p, such as $m \approx \sqrt{p}$ or $m \approx \log_2(p)$, while using $m = p$ corresponds to the aforementioned bagged trees [66].

While bagged trees and random forest models might be hard to interpret, as they may not be visualized as simply as a single decision tree, they may produce a measure of the importance of each individual input feature in the model (*feature importance*). Using an appropriate model metric as mentioned in the previous chapter (RSS, Gini index) the decrease in the model metric is averaged over all decision splits in the B trees at the same feature, providing a measure of the feature's contribution to the model - its importance [66].

2.4.4 Multiple linear regression

Using multiple explanatory variables to explain a continuous outcome variable as a weighted sum is known as multiple linear regression. Given an p-dimensional observation $\overrightarrow{x_i} = \left[x_{1i}, x_{2i}, \dots, x_{pi}\right]^T$ (the feature-space) the predicted outcome becomes:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \cdots + \widehat{\beta_p} x_{pi} \qquad 2.4\text{-}6.$$

The estimation of the $p + 1$ regression parameters $\hat{\hat{\beta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right]^T$ (where β_0 is the intercept), is done by minimizing the squared error across all observations (the residual sum of squares RSS). This procedure is known as least squared error (LSE) estimation [66]. The RSS in this parametric setting is similar to equation 2.4-4:

$$RSS = \sum_{i=1}^{N} (y_i - \widehat{y_i})^2 \qquad 2.4\text{-}7.$$

Where N is the number of training observations used to fit the regression parameters.

2.4.5 Logistic regression with ridge (l_2) regularization

Multiple logistic regression is a parametric model using a weighted sum of predictors (the *systematic part*) to express a continuous value between 0 and 1. The estimated outcome is often interpreted as the class probability for a binary outcome making it suitable for classification. Given an observation number i consisting of d features, $\vec{x}_i = [x_{i1},\ x_{i2}, \dots, x_{id}]^T$, the systematic component $\eta(\vec{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id}$ utilizes the d features as predictors in a logistic model as:

2.4-8.

$$f(\vec{x}_i) = \frac{1}{1 + \exp\left(-\eta(\vec{x}_i)\right)} = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^{d} \beta_j x_{ij}\right)}$$

The shape of this function is sigmoidal and produces outcomes between 0 and 1 given real values of β_j and x_{ij}. Given a binary outcome the estimated probability that the outcome y_i belongs to one of the classes (usually $y_i = 1$) becomes: $\hat{p}_i = P(y_i = 1|\vec{x}_i) = \hat{f}(\vec{x}_i)$. The distribution of y_i then becomes binomial as $P(y_i|\vec{x}_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$.

The regression parameters $\hat{\vec{\beta}}$ are estimated using the *likelihood L*, defined as the product of the observed distributions:

2.4-9.

$$L = \prod_{i=1}^{n} P(y_i|x_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\Rightarrow \ \ln(L) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)]$$

Maximizing L, or equivalently the natural logarithm of the likelihood $\ln(L)$, to estimate the regression parameters is known as maximum likelihood estimation (MLE) [66].

To mitigate overfitting a regularization parameter λ is introduced, penalizing the usage of more predictors. Adding the penalization term to log(L) seen in equation 2.4-10 is known as ridge regularization, or l_2 regularization as it is proportional to the sum of squared coefficients. The penalization strength depends on both λ and the number of predictors d and overall lowers the estimations of the parameters (shrinkage) [72].

2.4-10.

$$\overrightarrow{\hat{\beta}_{MLE}} = \underset{\vec{\beta}}{\operatorname{argmax}}(\log(L))$$

$$\overrightarrow{\hat{\beta}_{l_2}} = \underset{\vec{\beta}}{\operatorname{argmax}}(\log(L) - \frac{\lambda}{2}\sum_{j=1}^{d} \beta_j^2)$$

After the estimation of regression parameters (training), the fitted model may be used for classification. Given a classification threshold p_{thresh} an observation is assumed to belong to class $y = 1$ if the estimated probability is above said threshold.

2.4-11.

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i = \hat{f}(\vec{x_i}) > p_{thresh} \\ 0 & \text{if } \hat{p}_i = \hat{f}(\vec{x_i}) \leq p_{thresh} \end{cases}$$

2.4.6 Metrics for model assessment & statistical methods

The coefficient of determination (R^2) is a metric describing the proportion of the variability in an outcome variable explained by a regression model. The theoretical maximum $R^2 = 1.0$ means the model estimation follows the outcome perfectly [66]. The general definition of R^2 is:

2.4-12.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

RSS and TSS is the residual and total sum of squares, respectively. Using this definition R^2 may become less than zero, indicating that the model performs worse than a constant guess at the expectation value of the outcome (which would make $R^2 = 0$) [73].

For classification purposes the area under the receiver operating characteristic curve (ROC AUC, or simply AUC) is often used as a metric for evaluating classification models. AUC is always between 0 and 1, where the latter indicates a perfect model. The curve in ROC-space is created by plotting the *true positive rate* ($tpr = \frac{true\ positives}{all\ actual\ positives}$) against the *false positive rate* ($fpr = \frac{false\ positives}{all\ actual\ negatives}$), while varying the classification threshold (such that all prediction probabilities above this threshold is classified as true). As such each point in ROC-space represents a *confusion matrix* describing the goodness of the model's predictions [66]. An AUC score of 0.50 indicates that the model is not able to distinguish between the binary classes in any way and is equivalent to randomly guessing. The predictive performance of radiomics-based models is often evaluated and compared by the AUC [10], [41], [74].

However, the AUC has some drawbacks. The metric does not account for the stability of a predictor in ROC space, as two very different ROC-curves may have equal area. Sensitivity (TPR) and specificity (TNR = 1 – FPR) are also equally weighted which may not always be feasible, such as when considering detection of pathogens with pandemic potential demanding a very high sensitivity will less emphasis on specificity [75]. AUC may also be unfeasible when comparing different types of classifiers, as the metric is based on the intrinsic properties of the classifier [76].

The brier score (BS) is a probability-based metric for evaluating classification models, equivalent to the MSE for binary outcomes. For N binary outcomes y_i and estimated class probabilities $\hat{p}_i$, the BS is defined as:

2.4-13.

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$$

Independent of the prevalence of y_i the BS obtains a score of 0.25 given constant probability estimates at 0.5 (random guesses). A lower BS indicates better model performance. The model assumes no threshold-based classification inference, and is as such used much is probabilistic forecasting such as weather prediction [77]. While not heavily emphasised as a scoring metric in radiomics publications, it is utilized in [78].

3 Methods

As part of the research convergence environment PROtons Contra CAncer (PROCCA [79]) a pre-clinical study was performed to evaluate the long and short term effects of X-ray irradiation in mice. The radiomics-based image analysis in this thesis was performed using MR-images obtained in this study, with measured saliva production from the same mice used as a functional endpoint. Regions of interest of the left and right unit of the submandibular gland were segmented using a semi-automatic segmentation procedure. As method development was a key part of the study, it was warranted to include some discussion in this chapter on the various steps in the pipeline.

All programs created for this work is found in the GitHub repository:

https://github.com/mrbrodude/PROCCA

3.1 Data acquisition

The following chapter about how the data used in the radiomic investigation was acquired is written using excerpts from the study publication [80] where a complete description of all methods in the study is given in more detail. The study consisted of four experiments, including a pilot, with some variations in the study design. Data from all experiments are used in this study.

Mice from the same genetic line (C57BL/6J) was kept in a 12h dark / light cycle, pathogen-free conditions, and fed a commercially available fodder with no restrictions on water availability. The mice were 12 weeks old at the onset of each experiment, referred to as baseline at either day

-7 or day -3 in the experiments, and at this time baseline measurements were sampled. The time when the irradiation schedule was initiated is referred to as day 0 for all experiments – meaning the mice received the first irradiation fraction(s) on this day (see Figure 3-1). All experiments were approved by the Norwegian Food Safety Authority and performed in accordance with directive 2010/63/EU on animal protection for scientific purposes [81]. The mice were anesthetized using Sevoflurane 4% in O_2 gas when delivering each irradiation fraction, as well as when magnetic resonance imaging (MRI) was performed. During saliva sampling a subcutaneous injection of Zoletil-mix was used as anaesthesia.

The mice belonged either to an irradiation group receiving varying doses or a control group where no irradiation was delivered. The irradiated mice were either given 1 or 2 fractions per day for 5 or 10 days, such that all mice received 10 fractions in total, with doses varying from 3.0 to 8.5 Gray per fraction (Gy / f). The delivered X-rays was either generated with (1) 180 kV / 10 mA and 0.3 mm Cu filter and 0.65 Gy / min dose rate, or with (2) 100 kV / 15 mA and 2.0 mm Al and 0.75 Gy / min. Absolute calibration of the X-ray delivery system was performed using an FC65-G ionization chamber (IBA Dosimetry, Germany) with a MAX-4000 electrometer (Standard Imaging, USA) following standards for dose to water (section 2.1.1). The anaesthetized mice were positioned on their right side, and a custom-build lead collimator ensured a 25 x 20 mm radiation field covered the oral cavity, pharynx, along the major salivary glands of the mice.

Magnetic Resonance Imaging was performed using a 7.05 T Biospec scanner (Bruker Medical systems, Germany) at baseline (day -7 or -3), in quick succession after the irradiation, and at time points in the follow-up period varying with the experiments. T2 weighted images were acquired for 62 of the mice at varying time points, where 29 belonged to a control group receiving no irradiation. For 31 individuals T1 weighted images were also acquired, where 15 belonged to a control group. All series with T1 imaging had a corresponding T2 series. For some mice imaging both before and after pilocarpine injections (used for measuring saliva production) were acquired. An overview of the number of images taken at each time-point is seen in Table 3-1. The T2 imaging protocol was a fast spin echo (FSE) sequence (TurboRARE) with echo time TE = 31 ms and repetition time TR = 3100 ms having an echo train length ETL = 8 (see sections 2.2.3 and 2.2.4). The T1 protocol also used a FSE sequence (RARE) with TE = 8 ms, TR = 1500 ms and ETL = 4. Both the T1 and the T2 protocols produced images with a resolution of 256×256 pixels, but with varying amounts of slices (between 8 and 30). The pixel spacing for both the T1 and T2 images is isotropic in-plane with a distance of 0.12 mm between each pixel in every image slice, while the voxel spacing was 0.7 mm between the slices. Only images taken of the sagittal plane was used for analysis in this thesis, which constituted most of the images. The body temperature of the mice was monitored and maintained at 37°C using a feedback-regulated fan.

	Tot	# of mice (#control)	Before / after pilocarpine	Day →	-7	-3	5	8	12	26	35	56	70	105
T1	100	31 (15)	79 / 21		6	21	18	0	10	0	17	0	10	18
T2	233	62 (29)	151 / 82		61	21	38	26	10	18	17	14	10	18
Both	333	62 (29)	230 / 103		67	42	56	26	20	18	34	14	20	36

Table 3-1: Number of MR-images acquired grouped by weighting (T1, T2), before and after pilocarpine injection, and time (day 0 is the first day of irradiation).

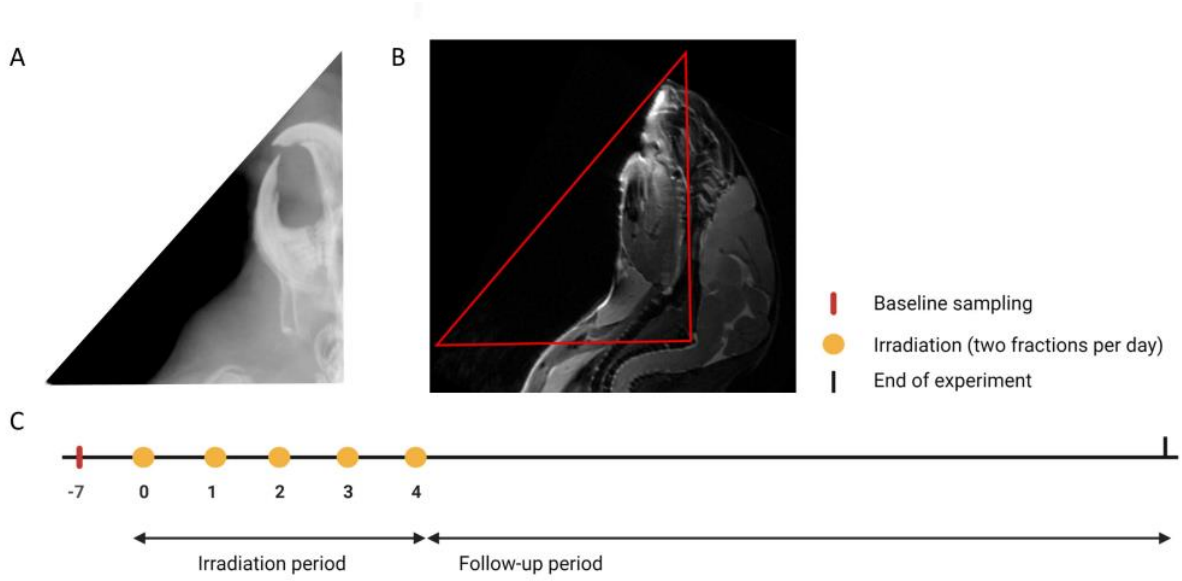


Figure 3-1: Radiation field (A and red triangle in B) covering the HN region of a mouse in a similar fashion to a clinical field. Figure C shows the timeline for baseline sampling (day -7), irradiation (day 0 to 4), and the follow-up period. Image from [80].

Measurements of saliva was taken at the baseline time, immediately after the irradiation period, and at some later time-point in the follow-up period. Of the 112 individual mice from which saliva were measured 40 belonged to a control group, and the remaining 72 were in a group receiving irradiation. A pilocarpine dose of 0.375 mg / kg (mg solution per mouse weight in kg) was given as an intraperitoneal injection to mice placed under anaesthesia. Saliva was collected using a cotton swab for 15 minutes, centrifuged at 7500 g at $4°C$ for 2 minutes, before the obtained volume was measured and stored. The number of saliva measurements taken over time is seen in Table 3-2.

Day	-7	-3	3	5	8	12	26	35	56	75	All days
# of measurements	72	40	31	39	31	40	10	51	15	18	347
# control	30	10	9	19	9	10	2	19	5	5	118
# irradiated	42	30	22	20	22	30	8	32	10	13	229

Table 3-2: Number of saliva measurements per day relative to start of irradiation (day 0), for 112 individual mice. All mice had baseline measures taken (at day -7 or -3), and in total 391 measurements were acquired.

3.2 Image segmentation

Defining the 2D region of interest (ROI), or 3D volume of interest (VOI), from which radiomic features are extracted is known as *segmentation* or *delineation*. The segmentation process in a medical setting considers a tumour or organ within the image(s) to be isolated – masked - for study. As the boundary of the tumour / organ to be delineated may be hard to observe on medical images, e.g. by being of similar density as the surrounding tissue in a CT image or having similar relaxation constants in a MR image (2.2.3), the segmentation process is sensitive to both intra- and inter-observer variations. Many of the radiomic features have been shown to be strongly affected by the ROI / VOI variations and are therefore considered non-robust with respect to segmentation [82].

While IBSI incorporates image segmentation after the post-acquisition processing in the workflow shown in Figure 2-11, the segmentation pipeline utilized here (described in section 3.2.4) considers very different preprocessing steps than the feature extraction process (section 3.3).

Manual segmentation is when an expert (oncologist, radiologist, medical physicist) uses experience and medical knowledge to decide where the boundary is. *Automatic* and *semi-automatic segmentation* are differentiated by the need of any user input, or manual corrections, in the process [8]. Dividing the image into regions satisfying some homogeneity criterion is known as *region-based* segmentation. Alternatively the segmentation boundaries may be defined by characteristics at edges within the image, such as high local changes in intensity represented by a high gradient, i.e. an *edge-based* segmentation [83].

The SMG, being the ROI, is divided into a left and right unit (section 2.1.3), with the center of each unit ending up in different sagittal image slices. Thus for each set of images acquired, the central slices of the left and right SMG were defined, largely by maximizing the area of the glands.

3.2.1 Histogram equalization

Histogram equalization techniques aim to increase the image contrast by transforming the intensity values in the image based on its histogram (see section 2.3.1). The simplest form of histogram equalization applies such a transform that the histogram becomes as flat as possible - and the cumulative histogram becomes a straight line [83]. This global method, considering all intensity values in the image, tends to enhance the noise of the image in addition to the contrast especially in regions of similar intensity (Figure 3-2). Contrast limited adaptive histogram equalization (CLAHE) addresses this issue by basing the transformation on several region-based histograms enhancing the local contrast (adaptive histogram equalization) while limiting the mapping function and as such the achieved contrast amplification [84].

Figure 3-2 is created using the python [85] implementation of OpenCV [86] having built-in functions for both linearizing the cumulative histogram (central column) and CLAHE (right column). The input image (left column) is required to have 8-bit unsigned integer gray values (uint8), i.e. the intensities are scaled (see equation 3.3-3) and discretized to integer values on the interval [0, 255] (as unsigned 8-bit values have 2^8 = 256 possibilities). This scaled uint8 image is the basis for all the subsequent steps in the segmentation pipelines.

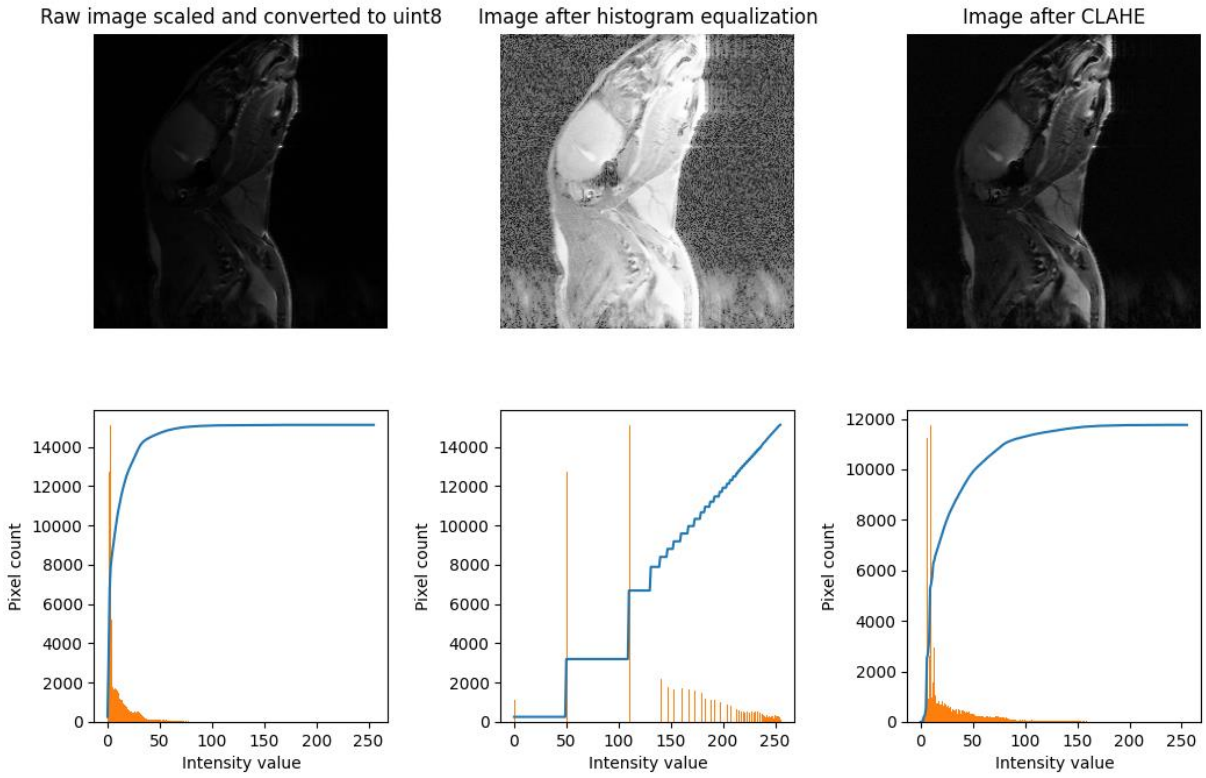


Figure 3-2: Image histograms (yellow lines, bottom row) for scaled image (left), after global histogram equalization (middle), and after CLAHE (right). Cumulative histograms seen as blue lines.

3.2.2 Rank- and kernel-based filters

Rank filters transform the image pixel intensity values based on the histogram calculated within a neighbourhood of each pixel. The considered neighbourhood is based on a binary structuring element – kernel – where the transformed pixel corresponds to the centre of the kernel matrix [87]. The python package scikit-image [88] function disk(radius=1) returns the 3×3 kernel $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, meaning the histograms are computed using the 4-connected nearest neighbours in all cardinal directions in addition to the central pixel. With an arbitrary integer-valued radius r the function returns a kernel matrix of size $(2r + 1) \times (2r + 1)$ increasingly resembling a circle of 1's within the square matrix.

Median filtering is a method used for noise suppression in an image, where all pixel values are transformed to the median value of intensities defined by the kernel [83].

Gaussian filtering is another method of noise suppression. Instead of considering all pixels in the neighbourhood defined by the kernel equally, as in the median filter, the method uses a non-binary kernel which weighs the contributions of each neighbour pixel with the gaussian function $g(x, y) = \frac{1}{\sqrt{2\,\pi}\sigma} \exp(-\frac{d^2}{2\sigma^2})$ where σ is the standard deviation. The distance from the central pixel (x_0, y_0), to be transformed, to a pixel in the neighbourhood at position (x, y) is defined as $d = \sqrt{(x - x_0)^2 + (y - y_0)^2}$ [89].

The *local gradient image* describes the change in pixel intensities within a local region of the image, computed by the difference between the maximum and minimum pixel values within the kernel-based neighbourhood [90].

The median filtered and gradient images computed in the segmentation pipeline (3.2.4), and in Figure 3-3, are created using scikit-image [88]. The gaussian filter used in the bias field correction (3.3.1) is created using OpenCV [86].

Figure 3-3: Gaussian (left), median filtered (middle) and gradient (right) of uint8 converted image after CLAHE. Median and gradient images are computed using scikit-image with kernel disk(radius=2), while the gaussian image is created using OpenCV with a 3×3 gaussian kernel.

3.2.3 Watershed

Watershed is a region-based segmentation algorithm, using the image morphology as a topologic basis to divide the image into regions. The regions are created in analogy to drainage basins, regions defined by added water (e.g. from precipitation) ending up in the same geographic region, divided by watershed lines separating such regions. Necessitating initial markers (seeds) within each object in the image to segment, and some user input for selection of the region corresponding to the wanted ROI, makes it a semi-automatic algorithm [83].

An intuitive watershed algorithm is the *simulated immersion* approach. In analogy to a two-dimensional surface the regional minima, valleys, are punctured with holes before the whole surface is sunk into a great water body. Water fills up the basins surrounding the valleys until water coming from two different punctured valleys would merge – where a dam is built. This process is continued until all regional minima are surrounded by dams – the *watershed lines* are delimiting each *catchment basin* [91].

While proving to be a versatile tool for segmenting 2D images, the formalism may be extended to more dimensions or grid patterns and have been shown to work on three-dimensional medical images [92].

3.2.4 Segmentation pipeline SMG

The methods mentioned in the previous subchapters are implemented in a watershed-based semi-automatic segmentation procedure (Figure 3-4). The segmentations are done on each 2D slice in the MRI series individually before the resulting ROI's are aggregated to a 3D matrix of similar shape as the image series. The ROI to be segmented is the pair of submandibular glands (SMG) in the mice (section 2.1.3.1). As the SMG pair contains two "central" slices when viewed in the sagittal plane (see Figure 2-4) the indices corresponding to said slices (left + right) in the image matrix was saved for extraction of 2D features.

The first step is to process each image slice in the image series, extracted from a DICOM file using pydicom [93], with a MR-specific bias field correction (N4 – see section 3.3.1)) to account for low-frequency nonuniformities across the whole image (see section 2.2.5). The corrected image intensities are then scaled and discretized to uint8 values in the [0, 255] range such that a CLAHE may be performed (see section 3.2.1). A denoised image is then created with a disk kernel having radius *mediandisksize*. A gradient image is created from the denoised image with a disk kernel having radius *gradientdisksize* (see section 3.2.2). A separate gradient operation is again done on the denoised image with a disk kernel of radius *markerthreshsize*, and pixels values below the *markerthresh* value is set to 1 – else 0. The regions are numbered by their connectivity (4 or 8 nearest neighbours) making the markers image. Watershed is then done between the first mentioned gradient image and the markers image (section 3.2.3), from which the watershed regions are manually chosen whether to be included in the slice ROI. A morphological closing operation is done each time a region is added to the ROI such that no

"holes" remains. This process is repeated for all slices until a 3D volume of binary values (1 if in ROI, 0 else) is created.

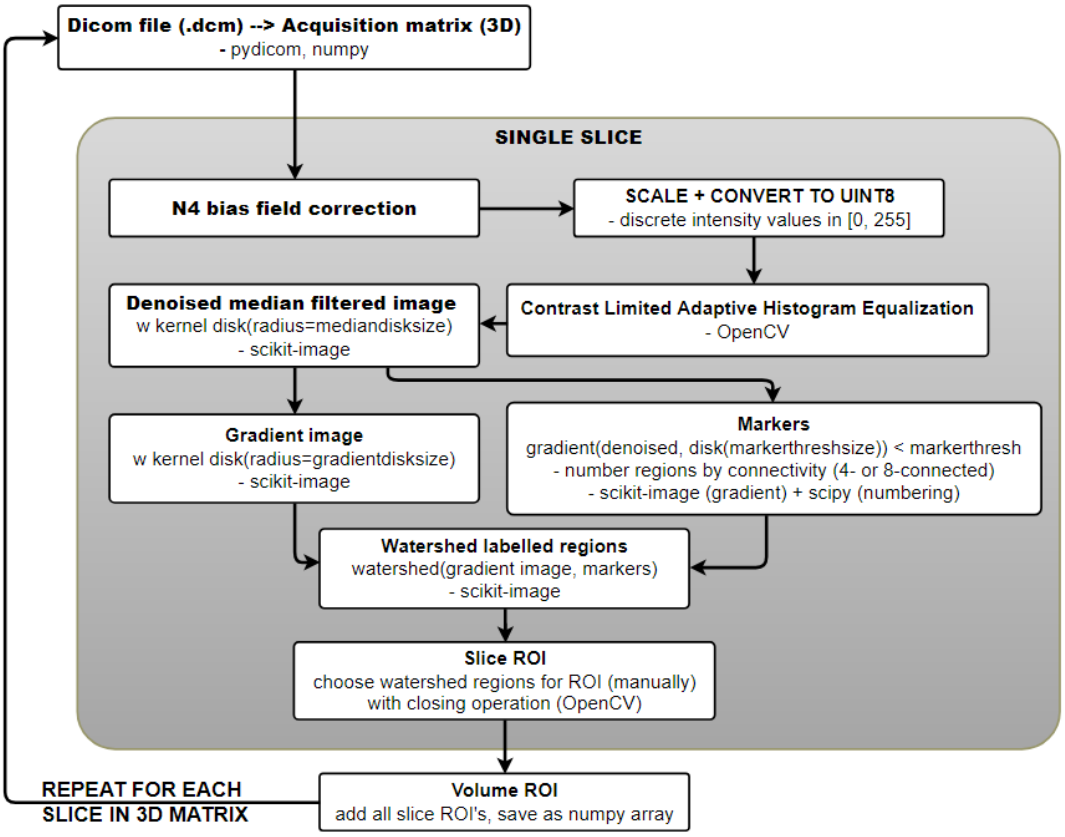


Figure 3-4: The ROI segmentation pipeline.

The segmentation pipeline is semi-automatic as user input is needed to select what watershed regions to include in the ROI, while the initial markers (watershed seeds) are created automatically.

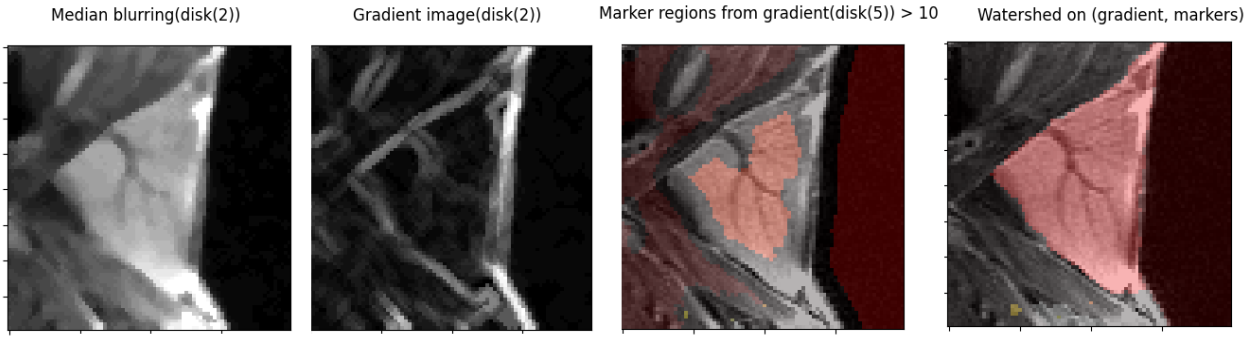


Figure 3-5: Illustrating the four steps used for segmenting the SMG.

The python packages used in the pipeline are as described in the corresponding methods subchapters, and the final 3D ROI is saved as a numpy array [94]. The segmentation pipeline is

heavily influenced by [95]. The implementation is found in *segmentation_algorithm.py* and *watershed.py* in the *extraction pipeline* folder.

3.2.5 Background identification by Otsu thresholding

Being a region-based segmentation method *global thresholding* divides an image into regions having gray levels either above or below some threshold intensity value T. A common application may be to divide an image into foreground (1) and background (0) using the step function [83]:

3.2-1.

$$g(x,y) = \begin{cases} 1 & \text{if } I(x,y) > T \\ 0 & \text{if } I(x,y) \leq T \end{cases}$$

Where $I(x,y)$ is the pixel intensity value at position (x,y).

Otsu's method attempts to find the optimal value for such a global threshold value by maximizing the distinction between the resulting classes [96] – with each pixel being assigned to either 0 or 1 in the binary case. Having a normalized histogram of N_G gray levels, such that $\sum_{i=1}^{N_G} p_i = 1$, the probability that an arbitrarily chosen pixel belongs to class 0 given a thresholding value k becomes $P_0(k) = \sum_{i=1}^{k} p_i$. Similarly, the probability of a pixel belonging to class 1 becomes $P_1(k) = \sum_{i=k+1}^{N_G} p_i = 1 - P_0(k)$. The average value of all discretized pixels up to, and including, gray level k in the image is defined as $m(k) = \sum_{i=1}^{k} ip_i$, and the global mean becomes $m_G = m(N_G) = \sum_{i=1}^{N_G} ip_i$. The mean value of all pixels belonging to class j may then be calculated as $m_j(k) = \frac{m(k)}{P_j(k)}$. From these results the *inter-class variance* m_B, a metric of the separability between the classes, may be calculated as:

3.2-2.

$$\sigma_B^2 = P_1(m_1 - m_G)^2 + P_0(m_0 - m_G)^2 = \frac{[m_G P_0(k) - m(k)]^2}{P_0(k)[1 - P_0(k)]}$$

By finding a $k^* \in [1, N_G]$ such that $\sigma_B^2(k^*) = \max_{0 \leq k \leq N_G} \sigma_B^2(k)$, the largest class distinction is found. Thus, by having $T = k^*$, one may use equation 3.2-1. to segment the image into two mathematically optimal classes. The thresholding process is often combined with some image smoothing to improve the segmentation performance [90].

A python implementation of Otsu's method is found in the SimpleITK library [97], which is used in the automatic background identification pipeline shown in Figure 3-6. The number of bins for creating the normalized histogram is set to 128. First the image is scaled (using min-max feature scaling, section 3.3.2.1) to the range [0, 255] before all intensity gray values are converted to a uint8 data type. The image is then blurred by a gaussian filter with a 9×9 kernel (3.2.2) before a histogram equalization is applied (3.2.1). The background and foreground are then segmented

using Otsu's method, and lastly a morphological closing operation is applied. Figure 3-7 shows an example of the identified background by this method.

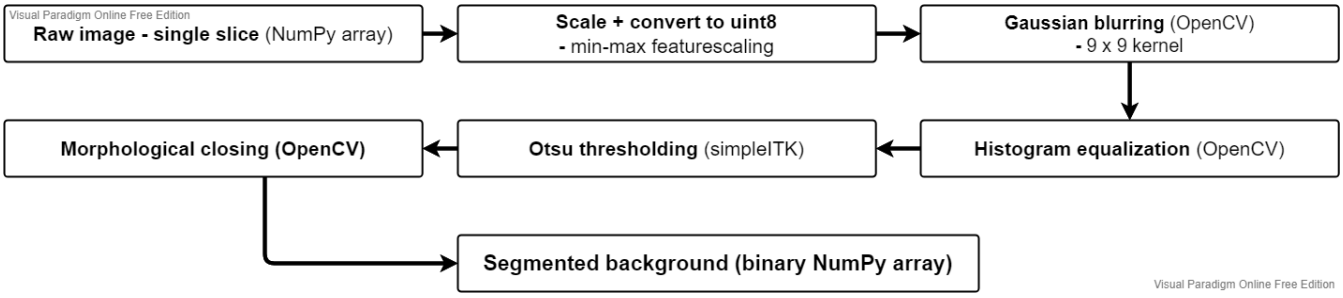


Figure 3-6: Pipeline for identifying and masking the background using Otsu's method. Preprocessing includes gaussian blurring and histogram equalization.

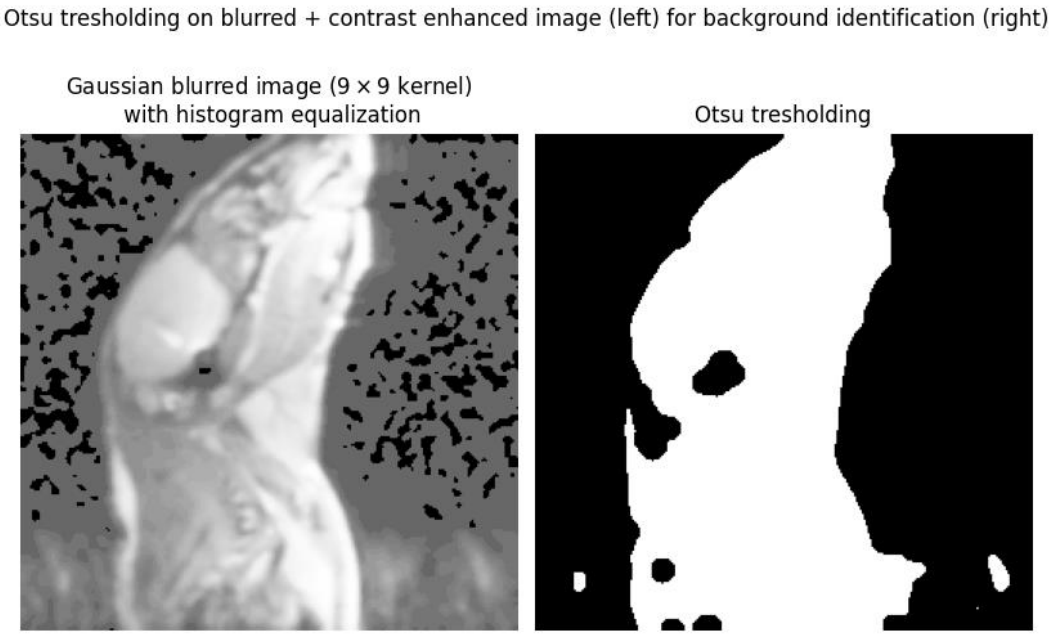


Figure 3-7: Otsu thresholding with 128 bins on 9x9 gaussian blurred image plus histogram equalization. Created using simpleITK [97].

3.3 Image preprocessing for radiomics

Whether or not, and how, to apply various image processing methods before extraction of radiomic features is a hot topic of discussion and research as they are major components of the radiomics process. While the image biomarker standardization initiative (IBSI) does not cover *post-acquisition* processing such as inhomogeneity correction methods and normalization, the

initiative provides standardized methods for interpolation, re-segmentation (exclusion of outliers in the ROI), and image discretization [7]. Preprocessing is considered a major contributor to the reproducibility of features, and the pipeline should be suited to the image sequence in use (T1 and T2 weighted MRI in the case of this thesis) [7], [8], [45], [47].

The order of application on the methods used for preprocessing in this thesis follows the results from a study concerning the interplay effects between post-acquisition preprocessing methods (bias field correction before normalization) [98], while the remaining steps follows the IBSI pipeline as seen in Figure 2-11. The order the methods are presented in this chapter is the same as is done in the extraction pipeline.

Due to having a much larger slice distance (0.70 mm) than in-plane pixel spacing (0.12 mm) no interpolation was done to an isotropic voxel environment, and as such all extraction of features was done from 2D images.

What normalization procedure to be used was chosen on a feature-specific basis, further described in section 3.4.1.

3.3.1 Nonuniform intensity normalization: N4 bias field correction

As discussed in section 2.2.5 MR images often contains some low frequency non-uniformity artifact across the image known as a *bias field*. The *improved nonparametric nonuniform intensity normalization for bias field correction* (from now on referred to as N4 in order to save paper) uses B-splines to iteratively approximate such non-uniformity artifacts as low-frequency fields covering the image [99]. Assuming the acquired image to be a function of the bias field without any noise one may model the image as:

3.3-1.

$$v(x) = u(x)f(x)$$

$$\Rightarrow \quad \hat{v}(x) \coloneqq \log\big(v(x)\big) = \log(u(x)) + \log\big(f(x)\big) = \hat{u}(x) + \hat{f}(x)$$

Where u is the assumed bias-free image, f is the low-frequency bias field, and v is the acquired image. After n iterations the corrected image becomes:

3.3-2.

$$\hat{u}^n = \hat{u}^{n-1} - \hat{f}_r^n = \hat{u}^{n-1} - S^*\{\hat{u}^{n-1} - E[\hat{u} \mid \hat{u}^{n-1}]\}$$

Where $\hat{f}_r^n$ is the estimated residual bias field between iteration n and $n-1$ equal to the B-spline estimator S^*. The B-spline estimator is made publicly available by the creators of the algorithm, and detailed derivations for the expectation value of the bias-free image given the current previous iteration of the corrected image ($E[\hat{u} \mid \hat{u}^{n-1}]$) is found in the paper describing the algorithm [100].

The B-spline is a generalized version of the Bezier curve, a method of nonparametric curve parametrization which may follow arbitrary data points in a smooth continuous way [101]. No a priori knowledge, or assumptions, about the underlying function which the bias field to estimate (f) follows is needed.

The python implementation of N4 in the simpleITK library recommends using a mask identifying the foreground in the image such that the background is not considered when estimating the bias field [102]. The workflow applying Otsu thresholding described in section 3.2.5 is used for this purpose. An example of the resulting bias field estimated from the identified foreground is seen in Figure 3-8.

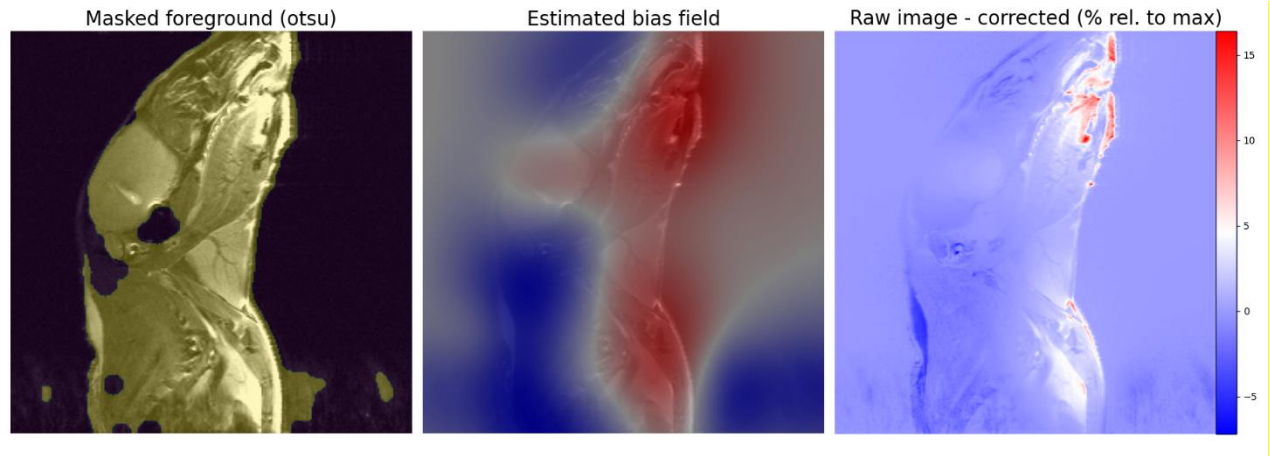


Figure 3-8: Illustration of the N4 bias field correction pipeline. Left: foreground identified by Otsu thresholding. Middle: estimated bias field. Right: difference in image before and after N4 was applied, as percentage values relative to the maximum intensity in the (masked) original image.

3.3.2 Image intensity normalization

As the intensity values in a MR-image has rather arbitrary scaling (see section 2.2.5) IBSI recommends including some procedure of homogenization of grey values between the images, as this has been shown to affect the subset of selected features as well as the reproducibility between MR-acquisition protocols [7], [46], [47]. Many methods of achieving such a homogenization have been proposed for use in radiomic analysis, and commonly occurring techniques include the standard score normalization and Nyul normalization.

Selecting the proper method of normalization is dependent on many intrinsic properties of the considered data but some general conditions should be satisfied: the normalization should not destroy information such as the inter-relationships between neighbouring pixel intensities (textural information) while achieving some standardization of the tissue-dependent intensity values to achieve a meaningful comparison.

Studies have shown that especially standard score normalization and Nyul normalization significantly increases the repeatability of radiomic features extracted from T2-weighed MRI images of cheese [103], the brain [46], and OAR's surrounding the prostate [47].

3.3.2.1 Min-max feature scaling

While not often considered a robust normalization procedure in radiomics, linearly transforming the intensities within the image from the original range to an arbitrary scale is of practical interest when preparing the images for various segmentation procedures. One example is when intensity value inputs are required to be of an unsigned 8-bit integer (uint8) data type having values on the interval [0, 255]. The method linearly scales the intensities in the image from the scale [min(x), max(x)] to a new interval [a, b] using:

3.3-3.

$$x' = \frac{(x - \min(X))(b - a)}{\max(X) - \min(X)} + a$$

Where $\min(X)$ and $\max(X)$ is the minimum and maximum intensity value in the image, respectively. The method makes no assumptions about the underlying intensity distribution, and as such achieves no centering – only scaling and shifting (if $a \neq 0$) [67].

3.3.2.2 Shifted standardization

Assuming the intensity distribution to be normally distributed one may rescale the intensities within an image to have zero mean and a standard deviation of one - as for the standard normal distribution. The user guide for pyRadiomics recommends this method of preprocessing, but with a small modification: instead of centering the mean of the transformed intensity values around 0, the mean is shifted such that the intensity values are centered around 3σ. This will ensure that no division with zero occurs when computing the first-order and texture based features (see section 2.3.1 and 2.3.3) [52], given that the ROI have been re-segmented to exclude such outliers (section 3.3.3). Standard score normalization, or *standardization*, with such a shift is referred to as standard score normalization in this thesis. The normalization function then becomes:

3.3-4.

$$x_{norm} = \frac{x - \mu}{\sigma} + 3\sigma$$

Where the mean μ and the standard deviation σ are computed from all pixel intensities in the image. The standard score normalization is considered more robust to outliers than min-max scaling [67].

3.3.2.3 Nyul normalization

Proposed by Nyul et al. [104] this MR-specific histogram-matching technique aims to standardize the intensity distributions for a set of images of the same body region and acquired using the same protocol (e.g. T2-RARE) to achieve a similar tissue-related intensity dependence. Achieving such a tissue-dependent "global" (in relation to a set of MR-images) intensity value is highly desirable for extraction and comparison of quantitative information in the radiomic procedure.

The method consists of two parts: first *training* a "standard histogram" at specified landmarks (reference points at percentile values in the histogram), which is then used as reference for transforming each image histogram linearly from the original image scale to the standard scale - using a separately acquired transformation function for each landmark interval.

Given a set of MR-images $\{v_j\}$ acquired by protocols $\{P_i\}$ of body parts $\{D_j\}$ a range of intensities of interest (IOI) is selected from the considered minimum and maximum percentiles pc_1 and pc_2 respectively. Along with l other percentiles they make up the landmark percentiles $L = \{pc_1, \mu_1, \mu_2, \ldots, \mu_l, pc_2\}$ corresponding to intensity landmark values $\{p_{1j},\ \mu_{1j},\ \mu_{2j},\ \ldots, \mu_{lj}, p_{2j}\}$ in the histogram of image v_j. Let m_{1j}, m_{2j} denote the minimum/ maximum of all intensity values in the image.

Finding the values constituting the standard histogram $\{s_1, \mu_{1s}, \ldots, \mu_{ls}, s_2\}$, corresponding to the landmark percentile set L, is done by mapping $[p_{1j},\ p_{2j}]$ onto $[s_1, s_2]$ (linearly) defining a transformation function τ_j for each image histogram within the IOI. All landmark values are then transformed by $\tau_j(\mu_{kj}) = \mu'_{kj}\ \forall\ k \in 1, 2, \ldots, l$ which yields each standard histogram landmark by calculating the mean over all the transformed values from the training set: $\mu_{ks} = \frac{\sum_{j=1}^{l} \mu_{kj}}{l}$.

Transforming (normalizing) each image is done by defining mapping functions for each landmark interval $[\mu_{ki}, \mu_{k+1\,i}]$ of the image histogram H_i onto the standard scale intervals $[\mu_{ks}, \mu_{k+1\,s}]$ by (linear) interpolation, creating a set of transformations making up the standardizer τv_i. The first and last map transformations are defined for the intervals $[p_{1i}, \mu_{1i}]$ and $[\mu_{li}, p_{2i}]$ onto $[s_1, \mu_{1s}]$ and $[\mu_{ls}, s_2]$, respectively. Concerning values outside the IOI the mapping function corresponding to the closest interval is used by expanding its domain interval to include either $[m_{1i}, p_{1i}]$ or $[p_{2i}, m_{2i}]$ while expanding the codomain interval by adding either $[s'_{1i},\ s_1]$ or $[s_2, s'_{2i}]$ by (linear) extrapolation as shown in Figure 3-9.

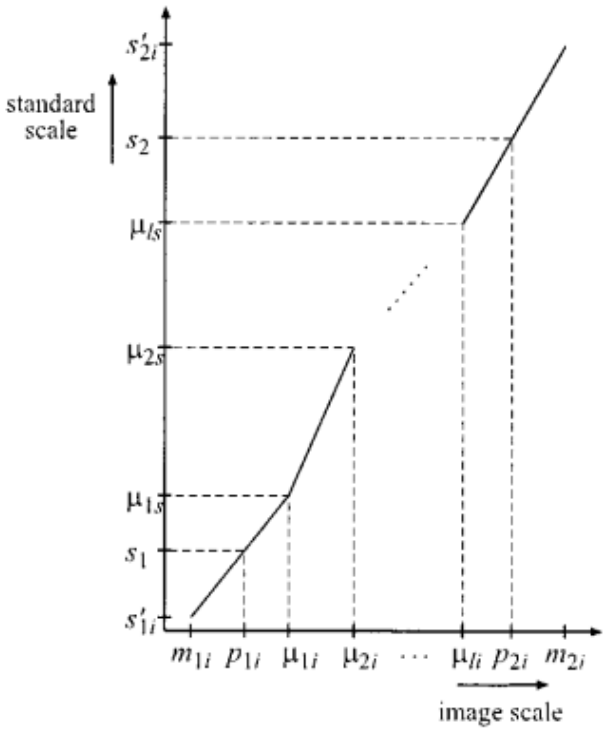


Figure 3-9: Linear intensity mapping from the image histogram scale to the standard histogram scale, defined for each landmark interval. Figure from Nyul et al. (2000, [104]).

As the choice of pc_1 and pc_2, defining the IOI, directly affects the (transformed) shape of the normalized image histogram along the choice of corresponding standard scale values s_1 and s_2 some conditions should be fulfilled to ensure the intensity relations between pixels are preserved after the transformation. The quantities

3.3-5.

$$\mu'_{min} = \min\{\mu'_i\} \qquad \mu'_{max} = \max\{\mu'_i\}$$
$$\mu_l - p_{1l} = \min\{\mu_{1i} - p_{1i}\} \qquad \mu_L - p_{1L} = \max\{\mu_{1i} - p_{1i}\}$$
$$p_{2r} - \mu_r = \min\{p_{2i} - \mu_{li}\} \qquad p_{2R} - \mu_R = \max\{p_{2i} - \mu_{li}\}$$

should satisfy the conditions

3.3-6.

1: $$\mu'_{min} - s_1 \geq \mu_L - p_{1L} \qquad s_2 - \mu'_{max} \geq p_{2R} - \mu_R$$

2: $$s_2 - s_1 \geq (\mu_L - p_{1L} + p_{2R} - \mu_R) \times \max\left\{\frac{\mu_L - p_{1L}}{\mu_l - p_{1l}}, \quad \frac{p_{2R} - \mu_R}{p_{2r} - \mu_r}\right\}$$

3: $\tau v_i(x_1) < \tau v_i(x_2)$ if and only if $x_1 < x_2, \quad \forall \quad x_1, x_2 \in v_i$

The conditions in equation 3.3-6 ensures a one-to-one mapping from the original intensity distribution onto the standard scale, meaning no information is lost. Condition 3 should be evaluated last as is assumes condition 1 is true.

As the images to be normalized have a varying number of pixels belonging to the background, which may affect the histogram percentile values in both the training and transforming steps, a background identification is suggested to be included. While the paper by Nyul recommends using a thresholding method based on the image mean value, the background identification based on Otsu thresholding (section 3.2.5) is used for this purpose as it performs better on the acquired images.

A self-developed python implementation, found in nyul_histogram_matching.py, is used for all Nyul normalization purposes which includes a validation function evaluating the equations in 3.3-6. The standard scales attained from training, used for transformation (i.e. normalization), are created independently for the T1 and T2 images as well as for the images after pilocarpine injections (i.e. the images are split into four groups having separate standard scales). The considered histogram landmarks used are the ten deciles in addition to $pc_1 = 2$ and $pc_2 = 98$ (the 2nd and 98th percentiles), with the corresponding minimum / maximum used for the standard scale histogram is $s_1 = 1$ and $s_2 = 50\ 000$.

3.3.3 Re-segmentation

As the segmented ROI may contain some pixels not belonging to the *actual* regions of interest, the submandibular glands, a second mask is created which attempts to omit such pixels. The new *re-segmented* mask is referred to as the intensity mask by IBSI, while the preserved original ROI is referred to as the *morphological mask* – used for calculation of shape features [7].

As MR-images are on an arbitrary intensity scale the commonly used method of re-segmentation is by excluding intensity outliers using the standard deviation. Values outside the range $[\mu - 3\sigma, \mu + 3\sigma]$ are considered outliers, where μ is the image mean and σ the standard deviation [8], [103].

3.3.4 Discretization

The calculation of certain radiomic features require the images to be discretized (see sections 2.3.1, 2.3.3). The two approaches considered for this purpose in radiomics is either using a fixed number of bins (fixed bin count, FBC), or having a fixed bin width (FBW). Both methods have a noise suppressing effect. FBC have some normalizing properties, but destroys the relationship between image intensity values and the physiological tissue [7]. As this relationship is not well-defined in the MR-images IBSI recommends using FBC for MRI radiomics. However, newer studies have shown that discretizing using FBC on MR-images reduces the feature reproducibility in relation to the inter-observer variability, and recommends using FBW after

some normalization procedure [47], [105]. Using Nyul normalization, which potentially establishes a meaningful relationship between tissue and intensity levels (section 3.3.2.3), might thus make FBW a viable choice.

Having a lower number of bins, i.e. having a lower amount of discretized gray levels N_G, reduces the complexity and size of certain textural matrices such as the GLCM (section 2.3.3). This, in turn, reduces the number of computations needed to calculate the corresponding texture features.

Given pixel x with intensity gray value I(x) the FBW discretization function defined by IBSI is:

$$I_{FBW}(x) = \left\lceil \frac{I(x)}{\Delta I} \right\rceil - \min\left(\left\lceil \frac{I(x)}{\Delta I} \right\rceil \right) + 1 \qquad 3.3\text{-}7.$$

Where the bin width ΔI is the width of the equidistanced bins, and $\lceil a \rceil$ is the ceiling function. The plus one is added to keep the lowest bin value above 0, which is necessary to avoid division with zero for some feature calculations. When discretizing using a FBW approach the user guide for pyRadiomics recommends having a bin width such that the number of gray levels in the ROI ends up in the range [30, 130] [52]. The optimal bin width is as such dependent on the normalization choice. Based on the results in section 4.3.1 the bin widths used for the three considered normalization methods (including no normalization as a "null hypothesis") are seen in Table 3-3. N4 correction was always performed regardless of normalization.

Normalization method	**No normalization**	**Standard score**	**Nyul**
T1 images	100	0.050	800
T2 images	100	0.075	950

Table 3-3: Optimal bin widths for FBW discretization after bias field correction and various normalization procedures.

3.4 Extraction of radiomic features

Radiomic features were extracted from the central slices of the left and right unit of the SMG, for all image series, with a feature-specific normalization choice for the T1 and T2 images separately (section 3.4.1). 828 two-dimensional features were extracted per image using the python package pyRadiomics [52] of which 9 were shape-based features, 162 first-order, and the remaining 657 were texture-based (Figure 3-10). For each image filter, 91 intensity-based features were extracted. The 100 counted for no filter includes the 9 shape features. Two versions of wavelet filters were applied, high (H) and low (L).

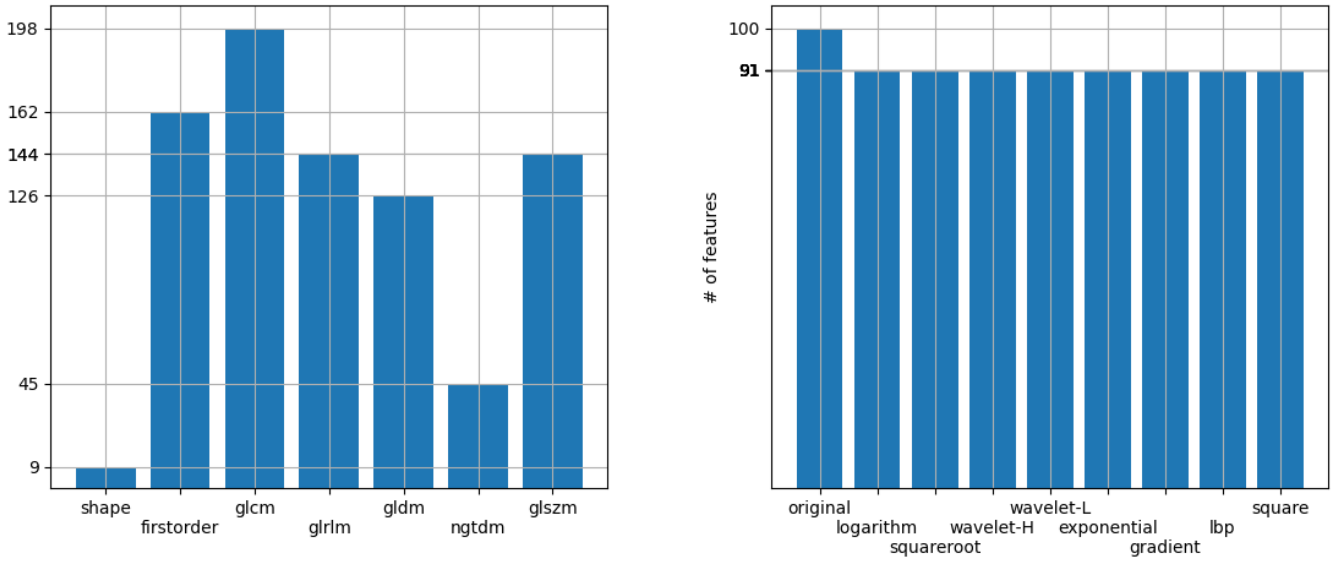


Figure 3-10: Number of extracted features belonging to the different feature classes (left), and with image filters applied (right).

3.4.1 Feature-specific preprocessing selection

Inspired by a paper published in Nature by Fave et al. (2016, [44]) the method of normalization was chosen on a feature-specific basis. Before extraction of all combinations of feature types and possible image filters (828 in total) three normalization procedures was applied: no normalization, standard score normalization (3.3.2.2), and Nyul normalization (3.3.2.3). The procedure of selecting the best normalization on a feature-specific basis is referred to as feature-specific preprocessing selection (FSPS). N4 corrections were applied on all images before any normalization was done (including no normalization).

Only MR-images from baseline instances were considered in the FSPS to establish an initial relationship between features and saliva measurements from the same time-points, acquired either 7 or 3 days before the first irradiation day (day 0). All images acquired after pilocarpine injection was omitted. This resulted in 55 images having both T2-MR data and saliva measurements for the same mouse at the same time, and 24 for the T1 weighted MR-images.

The T1 and T2 images were analysed independently, and both the central left and right SMG slices was included in the images for all considered instances.

For all features a spearman rank correlation was calculated between the measured saliva amounts and the feature corresponding to each normalization group. The correlation was considered significant if the p-value was below a set threshold. If only a single normalization group was significant this was selected as the best normalization procedure for the feature. If multiple groups were significant, the normalization which produced the *lowest* correlation between the feature and ROI area was selected (as this is a shape feature easy to interpret). If none was

significant the feature was dropped from the feature set for further analysis, and as such FSPS becomes the first step of feature selection.

The FSPS procedure was performed four times in total: separate for the image features from T1- and T2-weighted images, and whether the features from the left and right SMG unit were averaged (LR-average) or treated as separate features by being aggregated as new data columns (LR-aggregated). As the shape-based features (9 for 2D extraction) are unaffected by the normalization procedure (2.3.2) they were not included in the FSPS process and simply added to the remaining features.

Using a threshold of 0.05 for significance between saliva measurements and features, the T2 feature set size was reduced from 828 to 562. A threshold of 0.15 was used for the T1 images resulting in 127 remaining features. Spearman rank correlation and its p-value, for both saliva correlation and area dependence, was calculated using the python library Scipy [106].

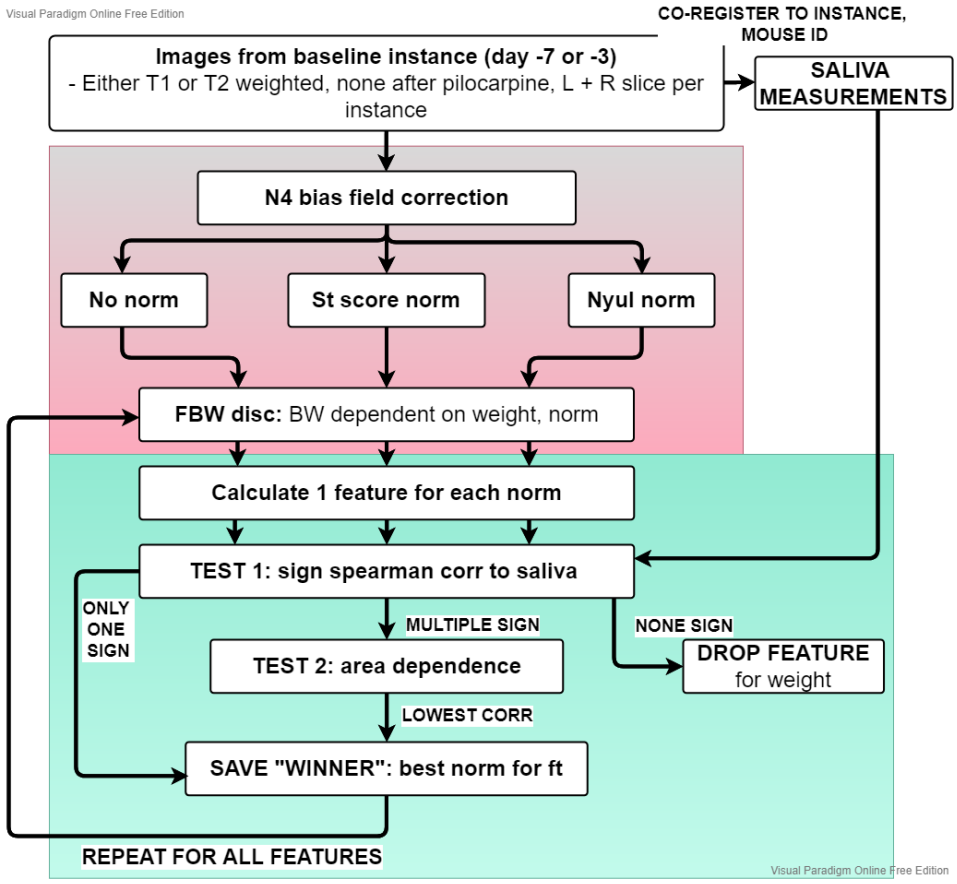


Figure 3-11: The feature-specific preprocessing selection pipeline. Red box represents the preprocessing steps, and the green box the normalization selection for each radiomic feature. The whole selection process is done for the T1 and T2 weighted images separately.

3.5 Delta-radiomics

A recent development in radiomic research, *delta-radiomics*, evaluates the changes in feature values over time on an individual basis. These *delta-features* have shown clinical promise in oncology for prediction of treatment responses and evaluation of side-effects [107], [44], [74]. Following two recent papers the relative net change was calculated between the baseline features and features from images acquired after irradiation using equation 3.5-1.

$$\Delta feature = relativeNetChange = \frac{ft_{after\ irr} - ft_{baseline}}{ft_{baseline}} \quad 3.5\text{-}1.$$

Of the 62 individual mice from which MR-images were acquired (section 3.1) 7 had no images taken at baseline. Of the 55 remaining mice 13 had no saliva measurements taken at later time-points than day 5, and 4 had no images taken after irradiation. Of the remaining 38 individuals available for delta-feature calculation with prediction of late saliva responses, 10 mice had images taken at day 35 with saliva measurements at day 75. The rest had images after irradiation taken at day 5, 8, or 12 with latest saliva measurements at day 26, 35, or 56.

Whether to include these 10 mice in the set of mice for delta-radiomics calculations, or not, becomes a trade-off between two options: either minimizing the variation in time-points considered as late saliva measurements by omitting these 10 individuals, or maximizing the sample size by allowing a temporal overlap between the late saliva measurements and some of the images taken after irradiation. Due to having a small sample size the latter option was chosen, such that all 38 available mice were included (Figure 3-12).

Only T2 images were considered for this analysis as only 21 mice having T1 images taken fulfilled the criterion described above. All images taken after pilocarpine injections were omitted for this analysis.

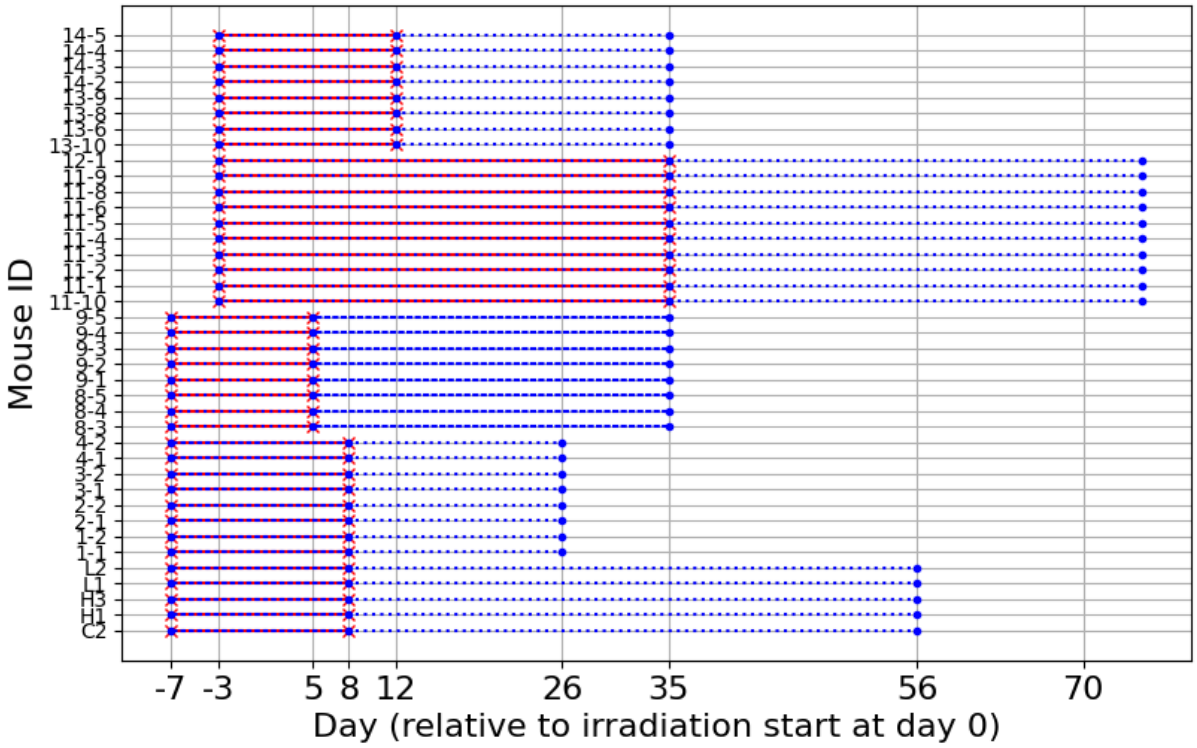


Figure 3-12: Times when the T2-w MR-images (red crosses), and saliva measurements (blue dots), were acquired for the N=38 individual mice available for delta-radiomic analysis. Delta-features are calculated between the images before (day -7 or -3) and after irradiation for prediction of the latest saliva measurement for each mouse.

3.5.1 Delta-P features

As a way of utilizing the images taken after pilocarpine injections, inspired by the aforementioned delta-feature calculations, the relative net change in features from images taken before and after pilocarpine injections and saliva extraction (referred to as before p / after p) were calculated. Instead of being a measure of the relative difference in time it becomes a measure of the effects on radiomic features due to saliva extraction and other eventual effects due to the pilocarpine drug. The relative net change is calculated using 3.5-2.

3.5-2.

$$\Delta\text{P-feature} = \frac{feature_{after\ P} - feature_{before\ P}}{feature_{before\ P}}$$

Only T2-w images are considered for this analysis as there are 36 individual mice which have been imaged before and after pilocarpine injections, while only 9 individuals when considering the T1 images.

3.6 Feature selection and modelling

3.6.1 Splitting the data into training and test sets

To estimate the generalization error of a trained model, evaluation of the model is done on unseen data – the test data (section 2.4). In a single train-test split a subset of the available data is held-out for such testing, a common approach in radiomic studies [10], [74], [108]. Resampling methods is often used in addition to, or instead of, this single split. Shayesteh et al. (2021, [74]) used 1000 bootstraps of the training data for hyperparameter (HP) tuning, and 1000 bootstraps of the test data for evaluation. Crombé et al. (2019, [108]) used a 10-fold cross-validation (CV) on the training data for HP-tuning and training before evaluating on the test data. Fave et al. (2016, [44]) had a somewhat different approach: instead of a single division into train and testing, the feature-selection, HP-tuning, and training was incorporated into a LOOCV which calculated the validation error using the held-out instances across the CV.

3.6.1.1 Single train / test split with respect to available mice for various feature-spaces

Three main methods of radiomic analysis are considered in this thesis, consisting of four feature spaces: using the extracted features only from images before pilocarpine injections (no-p) being either T2- or T1-weighted, using the delta-features (only T2), and lastly the delta-p features (only T2). Before any further dimensionality reduction of the four feature spaces beyond the already omitted features from the FSPS procedure, the data available for each method is split into a training set and a hold-out set for testing. The individuals chosen for the test set may affect the prediction and accuracy measures of the models, and as such the test sets are created to have a as big overlap as possible between the data available for each feature-space. This will allow for a more accurate comparison of the models, since a “global” test set is unfeasible as the individuals available for the different feature-spaces vary. As a second priority each test set is stratified by individuals belonging to the control group. The test set size is attempted to be about 20% of all available data for each method. The number of individual mice with data for each feature space is seen in Table 3-4, along the balance of control / not control individuals in all the train / test sets.

Feature type →	Standard radiomic features (no-p)			Delta-features			Delta-p features		
WEIGHT ↓	N_{tot} (%ctrl)	N_{train} (%ctrl)	N_{test} (%ctrl)	N_{tot} (%ctrl)	N_{train} (%ctrl)	N_{test} (%ctrl)	N_{tot} (%ctrl)	N_{train} (%ctrl)	N_{test} (%ctrl)
T1	29 (48%)	23 (48%)	6 (50%)	21 (48%)			9	-	-
T2	59 (46%)	47 (47%)	12 (42%)	38 (42%)	31 (42%)	7 (43%)	35 (43%)	27 (44%)	8 (38%)

Table 3-4: Number of individual mice for each train / test split. N_{tot} is the total available individuals for a given imaging weight (T1, T2) and feature type (no-p, delta, delta-p). Split into N_{train} training individuals and N_{valid} hold-out individuals for model evaluation, with percentage of the individuals being a control individual (having received no dose). Green cells indicate the feature type and weight is used for modelling, red cells indicate a too small sample size to be used.

3.6.1.2 Three train / test splits of individuals having both T1 and T2-weighted images

In order to compare the predictive abilities of image features from T1- and T2-weighted images, the instances (same mouse at same day) where no-p data (excluding all images after pilocarpine injections) was available for both instances was divided into three splits. As such the outcome data, measure saliva amounts, is the same for both sets of no-p features allowing for a more robust comparison given any outliers in the outcome data (which may vary when looking at different instances as in the split in section 3.6.1.1). Delta-features and delta-p features are not registered for this analysis, given very small T1-weighted features available as seen in Table 3-4. Three splits were made for three different prediction modes: either predicting saliva amounts acquired at the same time as the images (simultaneous), or predicting saliva measured at the latest time-point for each individual mice (days 35-75) using either image features from baseline (day -3 or -7) or right after irradiation (after irr, days 5-35).

For prediction of simultaneous saliva data some individuals may have both T1 and T2 images at multiple time-points, leading to a larger data set than the number of individuals. For prediction of late saliva each individual represents a single data point (being T1 and T2 features plus a single saliva measurement). When predicting late saliva using images from right after irradiation two individuals are not in any test set, thus only used for training (seen as the difference when summing up the test rows and comparing to all individuals in Table 3-5).

	All individuals (# control)	Test split 1	Test split 2	Test split 3
Simultaneous	29 (14)	9 (5)	8 (3)	12 (6)
Late baseline	24 (10)	8 (4)	8 (3)	8 (3)
Late after irr	26 (14)	8 (4)	8 (4)	8 (4)

Table 3-5: Number of individuals having both T1 and T2 images taken at the same time, for prediction of saliva amounts taken at the same time (simultaneous) or at the latest time-point using image features from baseline times or right after irradiation. The data is split into three train / test sets which are analysed independently.

3.6.2 Binary grouping of saliva measurements by xerostomia thresholding

As an alternative to making regression models for the measured saliva amounts, thresholding was applied dividing the outcome space into two groups suitable for binary classification: having xerostomia or not. To account for mouse growth, and thus an assumed natural increase in saliva production, a linear regression was done on all saliva measurements from any individual not having received any dose – the control individuals merged with all baseline measurements (N=190).

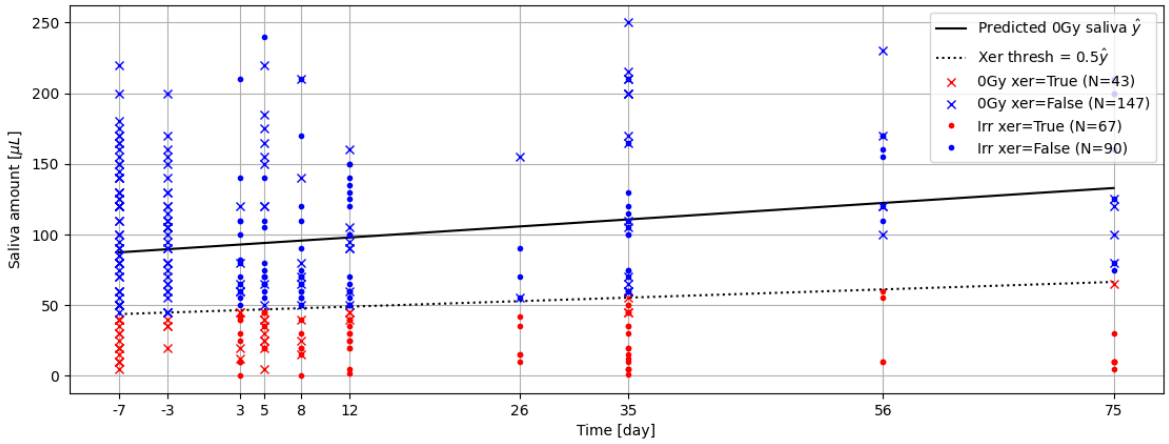


Figure 3-13: Thresholding of all saliva measurements. Xerostomia is assumed true if a measurement falls below 50% of the expectation value for the same day, following a linear regression on control + baseline data (i.e., all measurements from mice having received no dose).

Following the definition of xerostomia as a 50% reduction in saliva production (section 2.1.4) all measurement below half the expectation value from the regression line at a time-point was defined as xerostomia (see Figure 3-13).

3.6.3 Maximum relevance minimum redundancy

Being a method for feature selection (section 2.3.5), the *maximum-relevance minimum-redundancy* (MRMR) algorithm attempts to overcome the correlation between features in the selection process by iteratively selecting features while evaluating each pick to the already chosen subset. As the method works independently of the machine learning model choice it is considered a *filter method*, and as the method (considering all combinations) iteratively increases the pool of selected features it is known as a *forward selection* method.

For each iteration all remaining features available for selection are ranked by some scoring metric and the highest ranked is chosen. The scoring function should reflect the relevance of each feature with respect to the outcome, while penalizing redundancy with respect to the existing subset of chosen features. The F-test correlation quotient (FCQ) is such a scoring metric.

The F-score between an unselected feature and outcome values is calculated, and divided by the average correlation to previously selected features [109].

3.6-1.

$$FCQ_i = \frac{F(ft_i,\ y)}{\frac{1}{n}\sum_{j=1}^{n} corr(s_j, ft_i)}$$

ft_i is non-selected feature i, y are the outcome values to select with respect to, and s_j is the j'th of n already selected features. $Corr(X_i, X_j)$ is the Pearson correlation coefficient, and $F(X_i, Y)$ the F-statistic. Besides choosing the scoring metric is the only input parameter in MRMR the number of features to select, i.e. for how many iterations to run the algorithm.

A python implementation of MRMR is used to select the best k features in the training sets [110], individually for classification and regression purposes.

3.6.4 Hyperparameter tuning and bootstrapped model evaluation

All classifiers and regression models are trained using the training data from each feature spaces (Table 3-4), individually. First the training data is used to find optimal hyperparameters (HPs) such as forest size and optimal criterion measure for random forest models (section 2.4.3.2 and 2.4.3.3), or regularization strength λ for logistic regression models (section 2.4.5). A 5-repeated 2-fold cross validation (CV) evaluator was used for each hyperparameter optimization, resulting in 10 fits per hyperparameter combination. The hyperparameters resulting in highest average accuracy (for classification tasks), or highest average coefficient of determination (for regression tasks), using the CV evaluator was chosen.

4 Results

4.1 Comparing saliva production between control and irradiated groups

Of all 347 saliva measurements 118 were from individuals belonging to a control group and 229 from irradiated individuals. Qualitatively it is difficult to make any separation between the control and irradiated samples seen in Figure 4-1, with no obvious pattern visible.

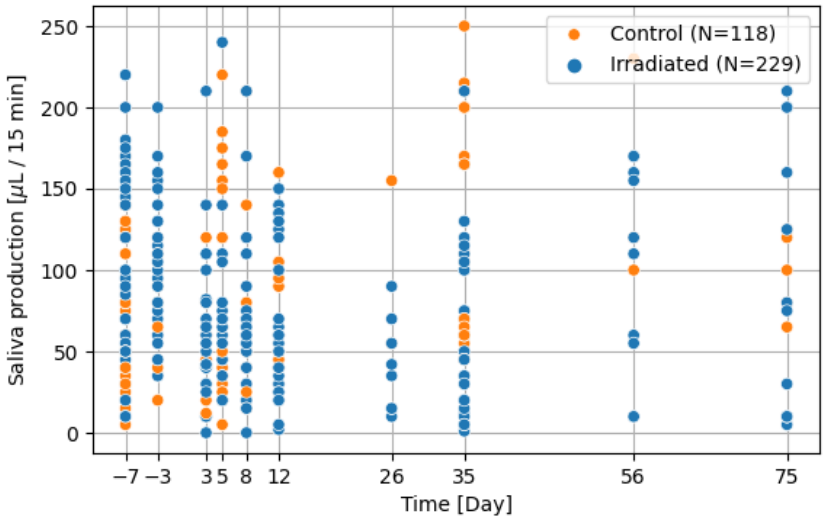


Figure 4-1: All longitudinal salivation data from control and irradiated individuals.

However, a Pearson correlation of -0.34 with a significant p-value (1.9×10^{-7}) is observed between dose and saliva amount when only the irradiated individuals are considered, as seen in Figure 4-2. While the correlation between time and saliva amount is not significant for the irradiated individuals (p = 0.29), a significant correlation of 0.29 for the control individuals (p=0.001) indicate an increase in saliva production over time when undisturbed by irradiation.

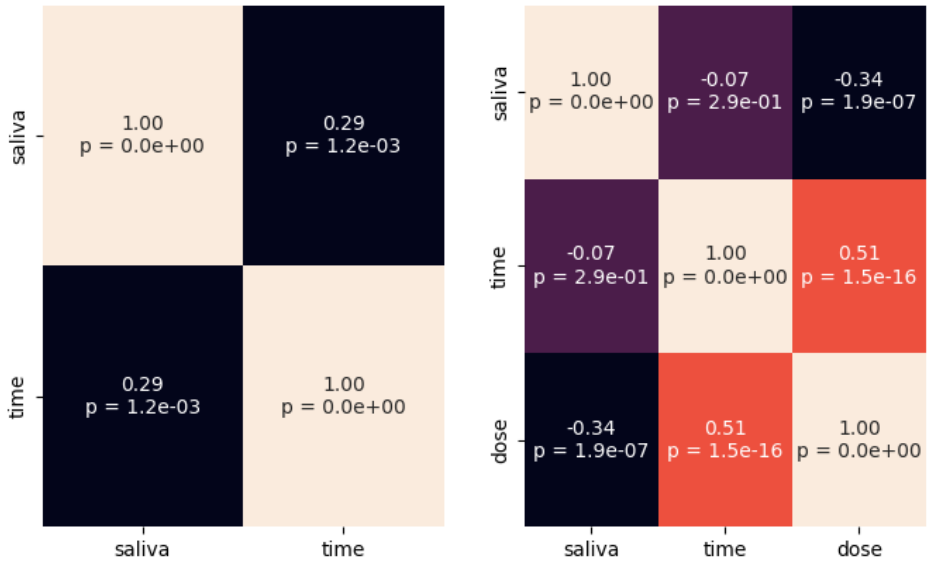


Figure 4-2: Correlation (first number) and p-value (second number) between saliva production, time, and dose for control (left) and irradiated individuals (right).

4.1.1 Longitudinal analysis of saliva measurements

The measurements were further divided into three groups by time of acquisition: baseline, acute, and late. Baseline being the first measurement taken before start of irradiation (day -7 or -3 relative to start of irradiation at day 0), acute being measurements taken close to the last irradiation day (days 3 – 12), and at later days (from day 26 to 75). Number of measurements for each group along summary statistics are seen in Table 4-1. The coefficients of variation for all time-groups, in both control and irradiated, are between 0.50 and 0.71 indicating a high spread between the measurements in all groups. Boxplots of the measurements from each time-group belonging to controls or irradiated individuals are seen in Figure 4-3.

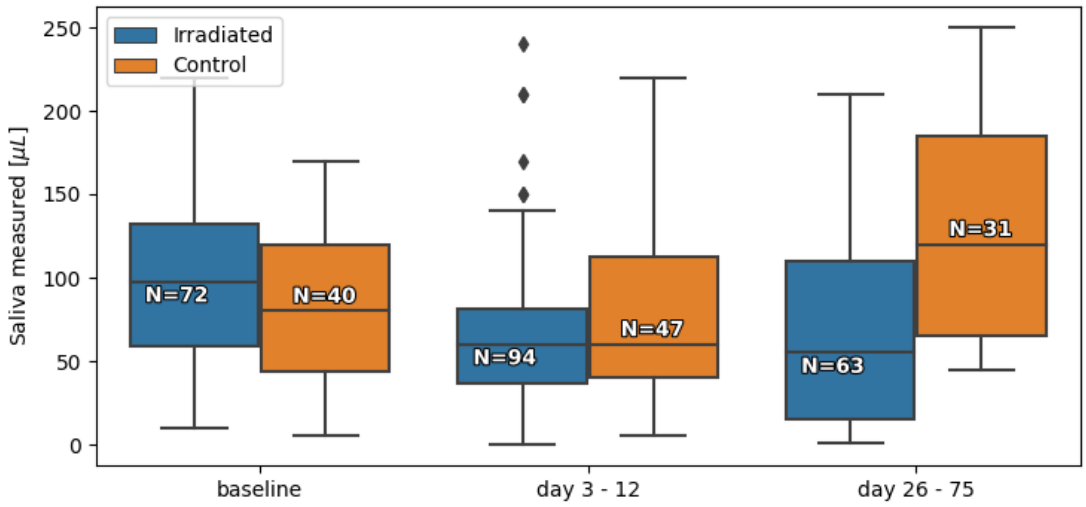


Figure 4-3: Box plot of all saliva measurements, grouped by time intervals: baseline is day -3 or day -7. Blue boxes are measurements from controls, orange from irradiated individuals.

	Number of measurements			Mean		Median		Standard deviation		CV = std / mean	
	Total	Ctrl	Irr	Ctrl	Irr	Ctrl	Irr	Ctrl	Irr	Ctrl	Irr
Baseline	112	40	72	81	98	80	98	44	50	.54	.51
Acute: days 3-12	141	47	94	78	67	60	60	56	47	.71	.70
Late: days 26-75	94	31	63	127	71	120	55	64	60	.50	.85

Table 4-1: Summary statistics for measured saliva values (µL / 15min) for control (ctrl) and irradiated (irr) individuals. The measurements are grouped by baseline times (day -7 or -3), after irradiation (day 5 to 12), and later measurements (day 26 to 75). All values are rounded to the nearest whole number, except the coefficient of variation (CV) which is rounded to two decimal places.

To evaluate whether the mean values of the control and irradiated samples are different in each time-group, Welch's t-test for independent samples were performed due to the groups having different means and sample sizes. Only for the last time-group there was a significant difference between controls and irradiated individuals as seen in Table 4-2, reflecting the largest difference seen in Figure 4-3. Looking at the distributions of saliva measurements for the various time-groups in Figure 4-4, a shift towards lower values is seen for irradiated individuals relative to controls.

	Baseline	Acute	Late
Sample sizes control / irradiated	40 / 72	47 / 94	31 / 63
Statistic t	-1.88	1.08	4.09
P-value	0.064	0.28	1.4×10^{-4}

Table 4-2: Welch's t-test between control and irradiated saliva measurements from the three time-groups.

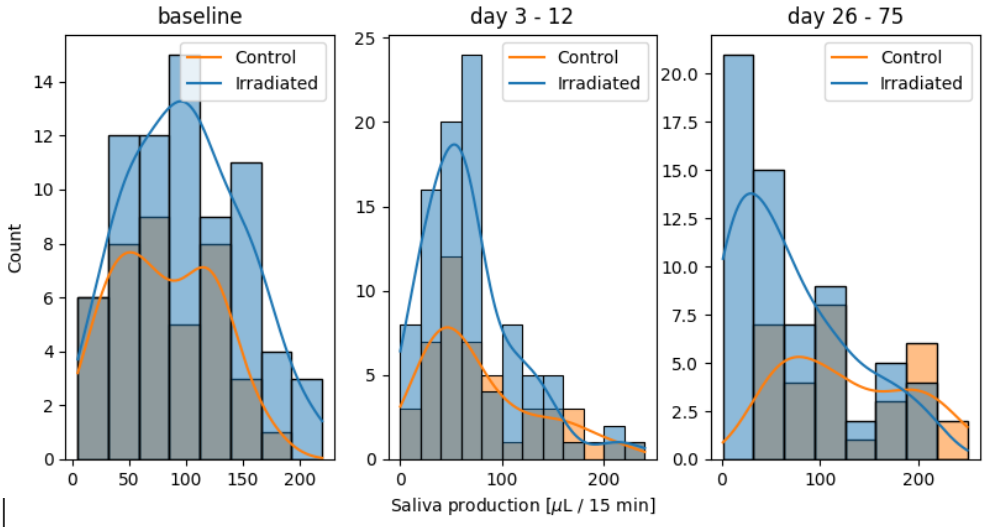


Figure 4-4: Histograms with kernel density estimated curve for saliva measurements from baseline (left), right after irradiation (middle), and the last days(right).

Three paired t-tests were performed to evaluate significantly different means between two time-groups: either comparing baseline measurements with acute measurements, comparing baseline with late, and comparing acute with late. The three paired t-tests were performed on measurements from either control or irradiated groups, separately, making up six paired t-tests in total as seen in Table 4-3. As not all mice had measurements taken at all times, the number of available pairs varied between the tests. If an individual had multiple measurements in the acute or late time groups, the latest measurements were chosen for this analysis.

	Baseline - acute		Baseline - late		Acute – late	
	Control	Irradiated	Control	Irradiated	Control	Irradiated
# of paired values	38	72	26	50	26	50
T-statistic	0.12	4.14	-5.12	1.83	-3.11	-0.26
P-value	0.90	9.4×10^{-5}	2.8×10^{-5}	0.07	4.6×10^{-3}	0.79

Table 4-3: Paired t-tests on longitudinal data for three different time groups, performed on data from control and irradiated individuals separately. Individuals having measurements in both compared time groups are used for each t-test, thus varying the number of compared measurements in each test.

4.1.2 Xerostomia thresholding

Deciding whether a mouse is said to have xerostomia for a measured saliva amount at a given time-point follows the method described in section 3.6.2. Xerostomia is said to be present if the measurement is below 50% of the expectation value of a simple linear regression line from all control data, including all individuals at baseline, over time (see Figure 3-13). The regression line was found to be increasing (91.25 + 0.56 × day), with a p-value of 0.008 and a coefficient of determination equal to 0.037 (see Figure 4-5).

Of all 347 saliva measurements 110 (32%) is thus considered xerostomic, being 33 of the 118 (28%) measured saliva values from control individuals and 77 of the 229 (34%) measurements from irradiated individuals.

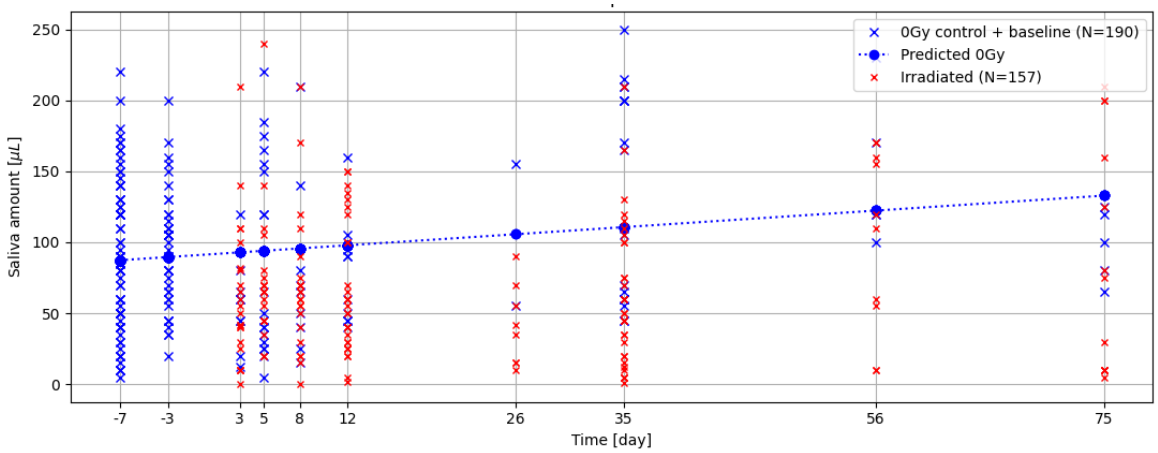


Figure 4-5: Simple linear regression (blue line) using all saliva measurements where no dose had been delivered (N=190), having time as the only input variable. The xerostomia threshold is defined as half the regression line.

4.2 Evaluating the segmented ROIs

The parameters for the watershed-based segmentation algorithm may be described by a tuple of integer values: (median disk-size, gradient disk-size, marker disk-size, marker threshold). All except the last threshold value describe the kernel sizes used in various image filters applied (see section 3.2.2).

The optimal choice of parameters varies between each segmented instance (mouse at a given day) together with image sequence (T1 or T2) and whether the images was taken before or after pilocarpine injections for saliva extractions (no-p or after-p). The count of each combination of segmentation parameters used is seen in Appendix A. All segmented ROIs for the left and right SMG unit at baseline is seen in Appendix B. Although not validated against a ground truth, it is clear that the segmentations varied somewhat in quality (Figure 4-6).

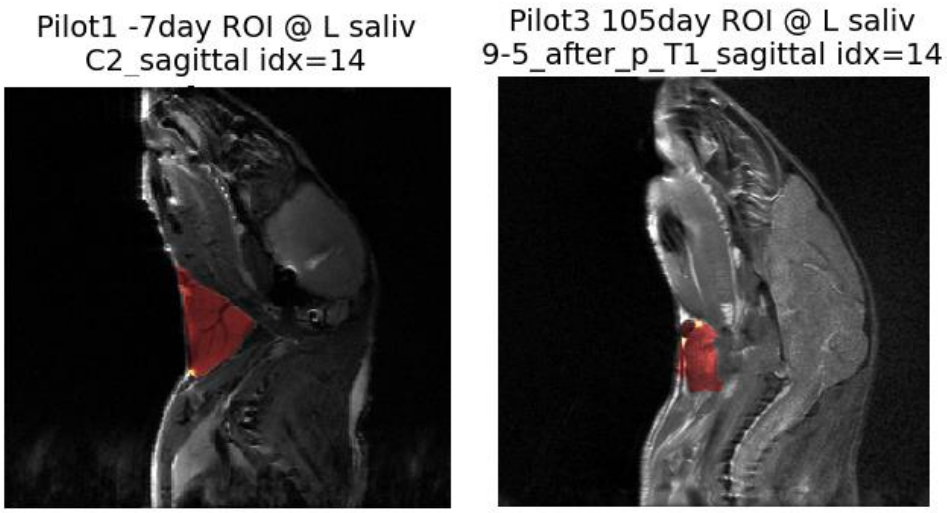


Figure 4-6: Example showing a good ROI segmentation (left) and a less good (right).

4.2.1 Comparing the segmented ROI to SG areas

The area of the left unit of both the sublingual gland (SLG) and the submandibular gland (SMG) was measured using surgical specimen from 20 mice after termination at day 105. The Pearson correlation coefficient between the measured SLG and SMG areas was 0.465 ($p = 0.06$). 9 of the mice also had MR-images taken at day 105, both T1 and T2 weighted. The highest significant correlation (0.71; $p < 0.05$) was between the SLG area measurements, and the T2 ROIs taken before pilocarpine injections (no p), as seen in Figure 4-7.

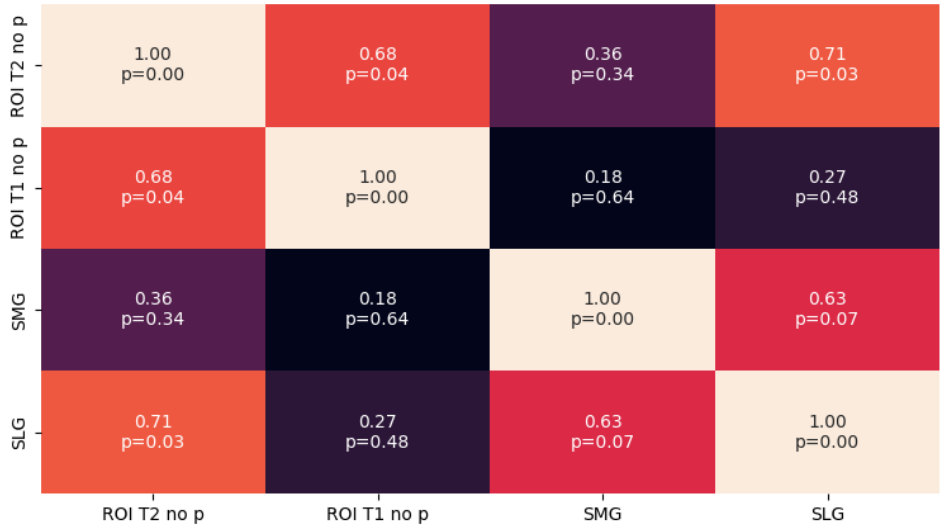


Figure 4-7: Person correlation matrix between segmented ROI size (T1 + T2 before pilocarpine, no p) and measured SMG or SLG areas for 9 mice.

4.2.2 Image-category variability between segmented ROIs

All images were categorized by being: T2-weighted before pilocarpine injections (T2 no-p, N=151), T2 after pilocarpine (T2 after-p, N=77), T1-weighted before pilocarpine (T1 no-p, N=79) and T1 after pilocarpine (N=19). Between these four image categories the variability in ROI size for images belonging to similar individuals at the same days was determined.

In total 326 of all 333 images was evaluated by only considering IDs and acquisition days present among the T2 before pilocarpine images (see Table 3-1). An example of all four sets of images for the same mouse and time points are illustrated in Figure 4-8.

Figure 4-8: Segmented image slices with segmented ROI (central left SMG). All images are of individual 8-3 taken at day 105.

A matrix of Pearson correlation coefficients was calculated between the ROI sizes (number of pixels) for the four image sets, done for both the left and right SMG slices, seen in Figure 4-9. A higher value indicates a lower variability between segmented ROIs.

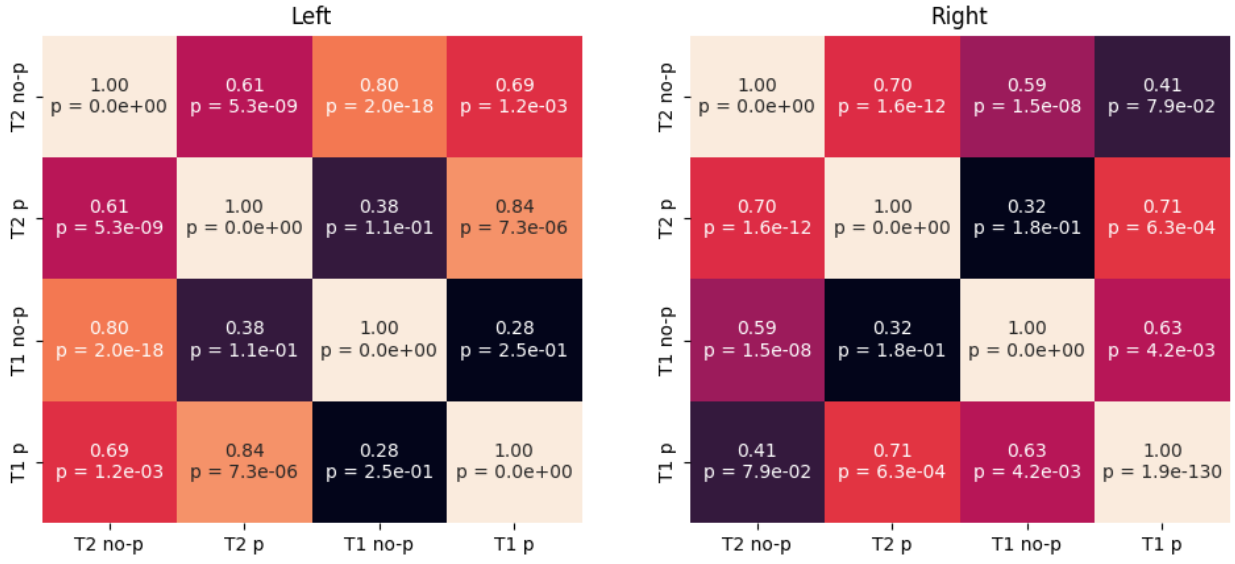


Figure 4-9: Pearson correlation between number of pixels in segmented left + right central SMG, for T2 and T1 images split into before (no-p) and after (p) pilocarpine injections.

4.2.3 Temporal evolution of ROI sizes

Each 2D ROI created for both the left and right unit of the SMG, for all image types (T1 or T2, before or after pilocarpine), were aggregated and split into the same three time-groups as in section 4.1.1. Among the 666 ROIs (333 images times with a left and right SMG unit), the biggest difference in size between the ROIs belonging to control and irradiated individuals is seen in the last time-group (days 26 – 75), where the irradiated SG ROIs have a slightly higher median than control as seen in Figure 4-10. The difference in means is 151 pixels, which was significant under an unpaired t-test ($p < .000$).

No difference in ROI size between the left and right SMG units is observed across the time-groups.

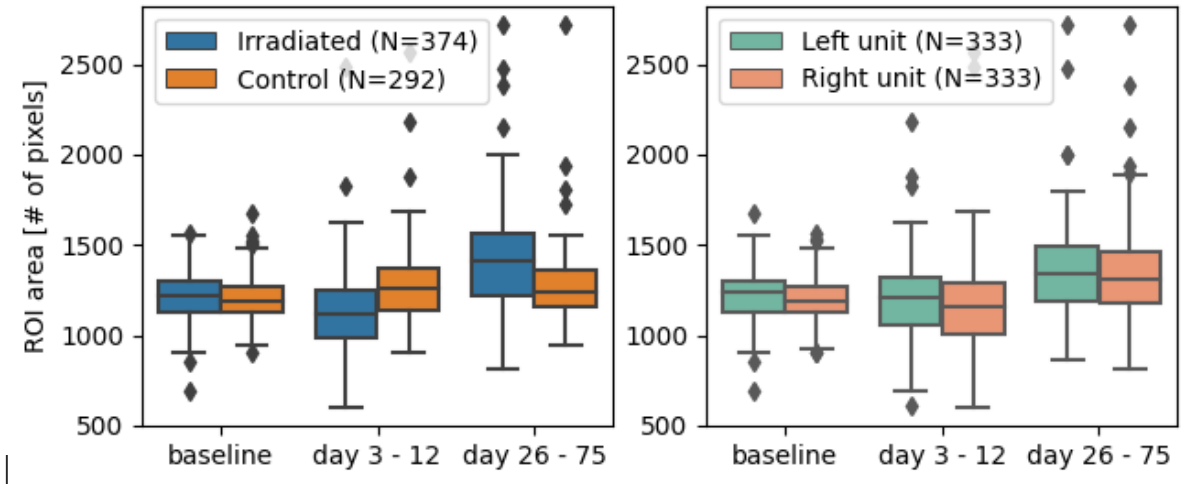


Figure 4-10: Size for all segmented 2D ROIs in three time-groups, split between control and irradiated individuals (left) or ROI's corresponding to the left and right SMG unit (right).

4.3 Preprocessing results

4.3.1 Intensity distribution variability in the ROI after normalization

The distributions of pixel intensities in the ROIs of all images after some or no normalization is seen as kernel density estimated lines (KDE lines) in Figure 4-11. N4 corrections were applied before all normalizations. Comparing the distributions after no normalization to Nyul normalization one may observe a lowered dispersion between the KDE-lines from individual images for the latter. The effect on distributions after standard score normalization is difficult to interpret in the figure due to the 3σ shift (section 3.3.2.2), which varies between each image and as such shifts the center of each distribution.

Using the mean pixel intensities in the ROIs $\{\mu_i\}$ a coefficient of variation of the means (cv_μ) was calculated as the standard deviation of the means (σ_μ) divided by the mean of the means ($\bar{\mu}$) (i.e., $cv_\mu = \sigma_\mu \,/\, \bar{\mu}$). cv_μ calculated for different image subsets after various normalization techniques are seen in Figure 4-12. cv_μ is overall lowered when only applying the N4 correction to the raw images (comparing no norm to raw), further lowered by applying either Nyul or standard score normalization. For all normalizations cv_μ is lower for the T1-weighted images compared to the T2-weighted images. The difference in cv_μ between the images before and after pilocarpine injections (no p / p, respectively) is small with no trend across the normalizations.

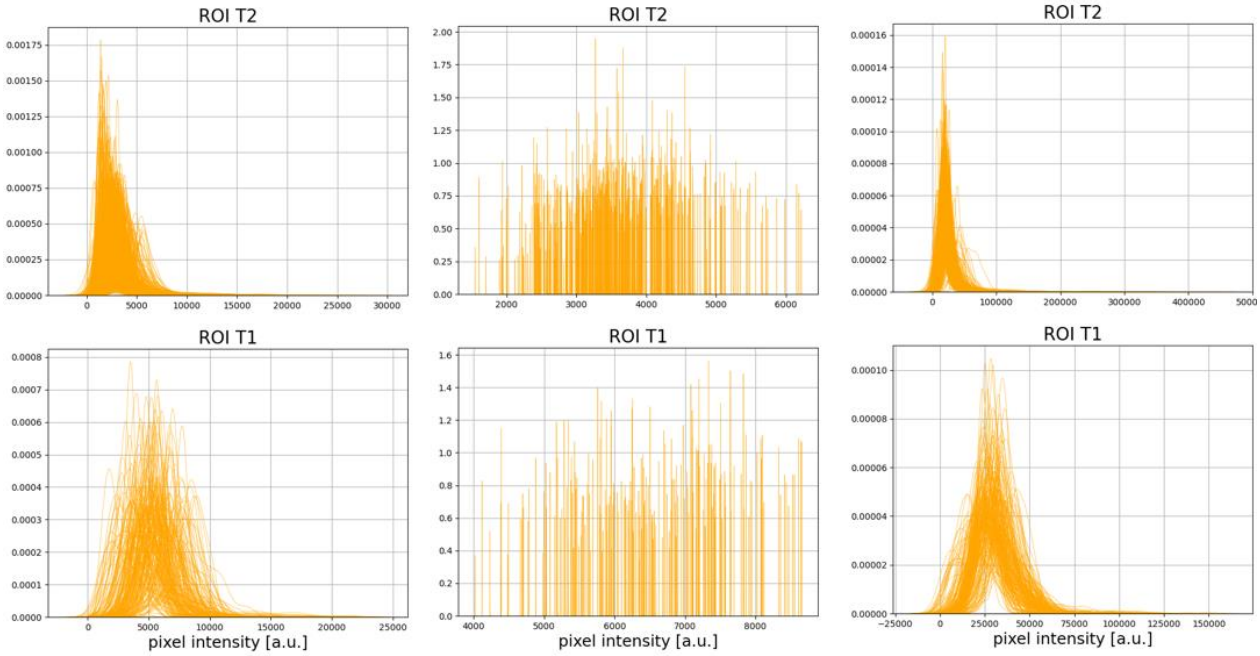


Figure 4-11: KDE lines of intensity values in the ROIs from T2- (upper row) or T1-weighted images, after various normalizations. From left: no normalization, standard score normalization, and Nyul normalization.

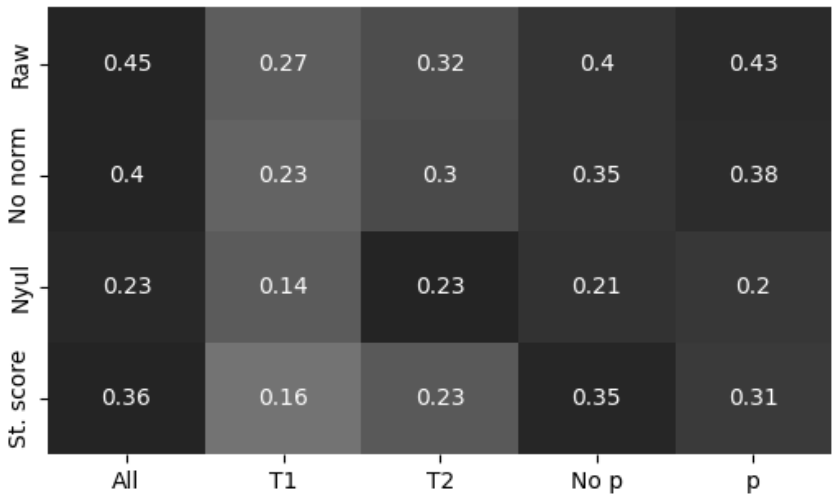


Figure 4-12: Coefficient of variation (CV) of mean pixel intensities in the ROI for various image types (columns) and normalization techniques (rows). N4 corrections were applied before any or no normalizations, except for raw where no preprocessing was done.

4.3.2 Number of bins in the ROI after FBW discretization

As the optimal choice of bin width in a fixed bin width discretization heavily depends on the values and variability of the pixel intensities in the ROIs, the bin width must be chosen in accordance with the former preprocessing steps. As the normalization procedure in this work varies between features the optimal bin width must be chosen accordingly. The intensity distributions in the ROI vary between the T1- and T2-weighted images (see Figure 4-11), such that the optimal bin width also should depend on the MRI-weightings. No split is made between before and after pilocarpine injections as the acquisition protocol is unchanged. All images are N4 corrected before the normalization is done.

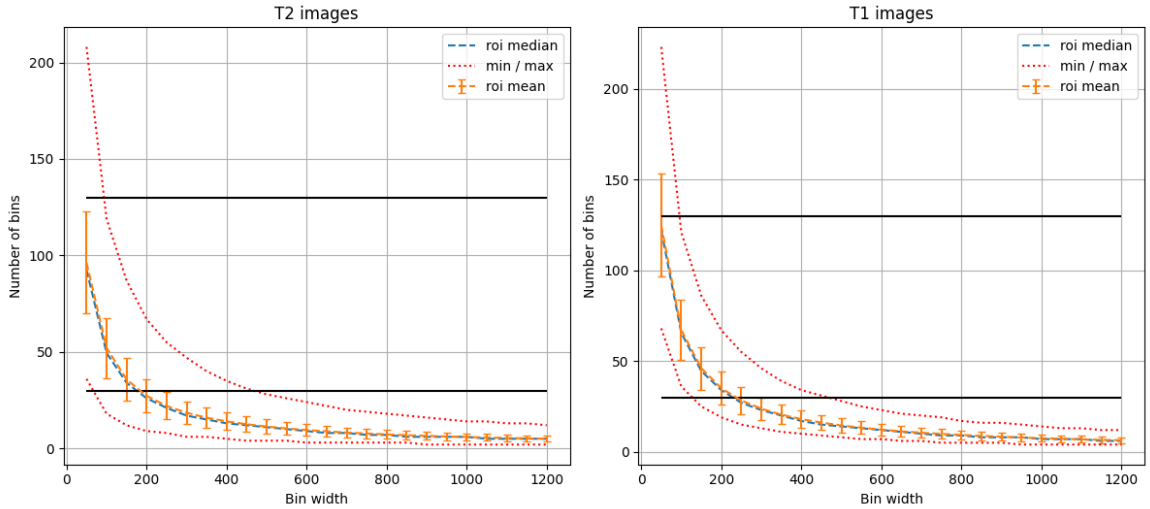


Figure 4-13: Resulting number of bins in the ROIs after FBW discretization (no normalization).

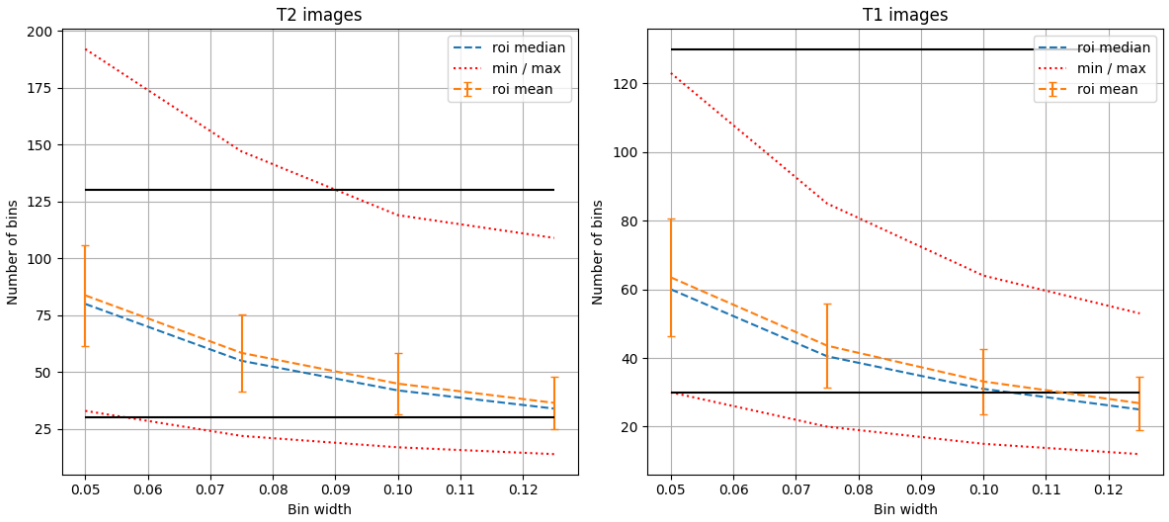


Figure 4-14: Resulting number of bins in the ROIs after FBW discretization with shifted standard score normalization.

As seen in Figure 4-13 a bin width of 100 seems to be a reasonable choice for both the T1 and T2 images after no normalization have been applied.

Looking at the min and maximum number of bins for the T2 images after standard score normalization in Figure 4-14, it seems hard to get all bins in the [30, 130] range. A bin width choice of 0.075 seems reasonable as the minimum and maximum bin counts are only slightly out of this range. This is not as big an issue for the T1 images, where a bin width of 0.05 seems optimal.

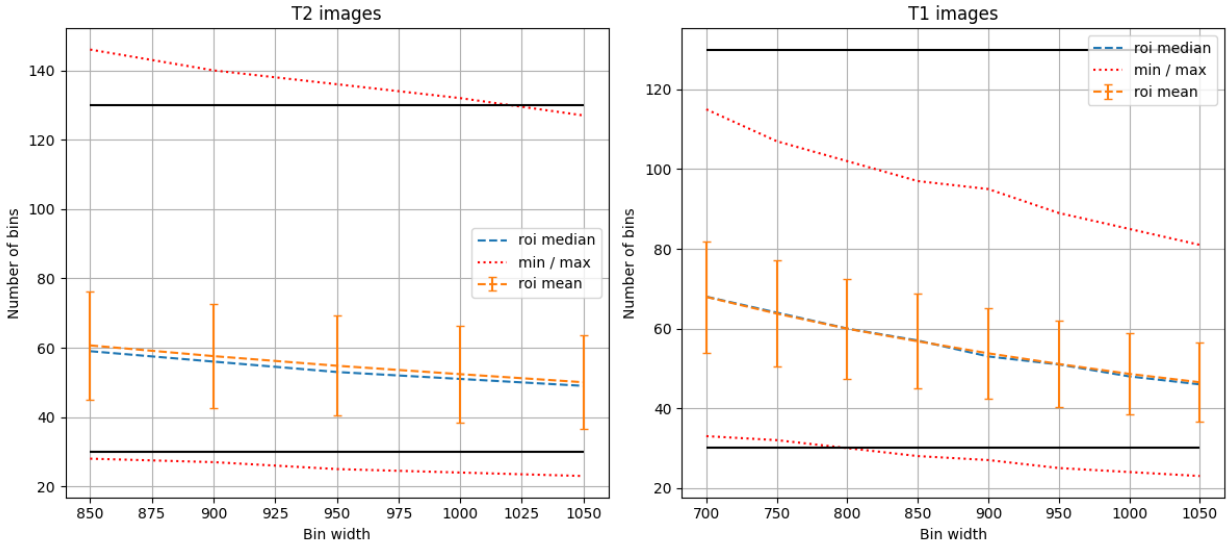


Figure 4-15: Resulting number of bins in the ROI after FBW discretization with Nyul normalization.

Following the Nyul normalization seen in Figure 4-15, the bin width is set to 800 for the T1 images and 950 for the T2 images.

Normalization	No normalization	Standard score norm	Nyul norm
Optimal bin width T1	100	0.050	800
Optimal bin width T2	100	0.075	950

Table 4-4: Optimal bin width for FBW discretization of T1 and T2 images, following various normalization methods.

4.3.3 Feature specific preprocessing selection

Using the methods described in section 3.4.1, feature-specific preprocessing selection was done for each combination of LR-modes and MR-weight. A threshold of 0.05 was used for all T2 images, and 0.15 for all T1 images. The number of remaining features between the three normalization modes are seen in Table 4-5.

	T1 LR-agg	T2 LR-agg	T1 LR-avg	T2 LR-avg	T1 LR-split	T2 LR-split
# fts before FSPS	1656	1656	828	828	828	828
# fts after FSPS	227	694	63	387	115	550
No norm	45 (20%)	43 (6%)	4 (6%)	22 (6%)	24 (21%)	68 (12%)
St. score	84 (37%)	277 (40%)	46 (73%)	85 (22%)	74 (64%)	277 (50%)
Nyul	98 (43%)	374 (54%)	13 (21%)	280 (72%)	17 (15%)	205 (37%)

Table 4-5: Summary of image features (excluding shape-based) remaining after FSPS, for each combination of weights T1, T2 and LR-modes aggregated, average, split. Each remaining feature was from an image either not normalized (no norm) or normalized using shifted standard score or Nyul-normalization.

4.4 Regression analysis

4.4.1 Time and dose as explanatory variables

Using all available data of saliva measurements, time and dose were used as explanatory variables to fit regression models with the saliva amount as the outcome. A multiple linear regression model was fitted to and evaluated on all available data (N=347) or using a 5-fold cross validation (CV). The estimated intercept and coefficients with corresponding p-values, the F-statistic p-value for the models, and the coefficient of determination (R2) for the models evaluated on the training and test data are seen in Table 4-6. In addition to the linear model a random forest (RF) regressor was fitted to the same data with R2-values for the training and test data seen as the bottom two rows. Hyperparameters (HP's) for the RF regressor were tuned on a

2-fold CV of the training data, repeated 5 times, to mitigate overfitting to the training data (Figure 4-16). HP-tuning was independent between each fold in the 5-fold CV. No HP-tuning was performed before fitting all data.

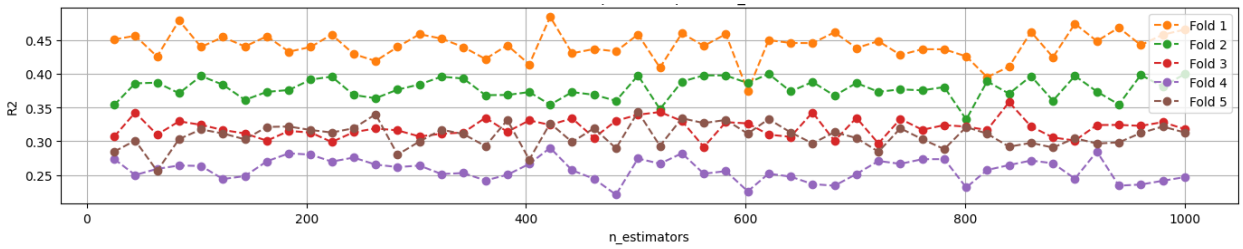


Figure 4-16: R2 as a function of number of estimators in random forest regression models, i.e. the number of decision trees in the random forests. For each fold the training data was evaluated by a 5-repeated 2-fold cross-validation and evaluated by the average R2 seen in this plot.

The RF model scored a higher R2 than the linear regression model when evaluated on the training data for each fold, and on the models using all data. For all test folds the regressors performs oppositely with the linear regression models having higher R2 than the RF models. The models perform differently on the test data across the folds with R2 ranging from -0.41 to +0.18 for the RF models, and from -0.04 to +0.23 for the linear regression models. Fold 4 yields the highest R2 on the test data for both regressors along the lowest R2 on the training data, and the highest p-value between the linear regression models.

The estimated intercept in the linear model is seen to be stable across the folds with low p-values (< .000), while the estimated coefficient for the dose-variable varies more – but maintains low p-values (< .000). Time, however, is seen to be the most unstable explanatory variable with coefficient estimates ranging from 0.29 in fold 4 to 0.87 in fold 5. The estimated time-coefficient is barely significant in fold 4 with p = 0.046. The feature's importance in the RF models is split approximately 50 / 50 between time and dose in all folds, where fold 5 is the only fold where time is more important than dose – corresponding to the higher estimated time coefficient in the linear model for this fold.

	All data (no split)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test set size (control)	347 (118)	70 (30)	70 (21)	69 (28)	69 (20)	69 (19)
P-value model	5.65×10^{-9}	1.87×10^{-8}	4.11×10^{-8}	7.83×10^{-10}	1.52×10^{-4}	7.26×10^{-7}
Intercept (p-value)	93.21 P < .000	92.21 P < .000	95.91 P < .000	94.46 P < .000	90.79 P < .000	92.88 P < .000
Coef. Time	0.43 P = .002	0.45 P = .002	0.40 P = .006	0.55 P < .000	0.29 P = .046	0.87 P = .006
Coef. Dose	- 0.60 P < .000	- 0.59 P < .000	- 0.62 P < .000	- 0.76 P < .000	- 0.49 P < .000	- 0.60 P < .000
R2 train	0.10	0.12	0.12	0.14	0.06	0.10
R2 test	-	0.00	0.01	- 0.09	0.23	- 0.04
RF R2 train	0.26	0.22	0.20	0.25	0.13	0.19
RF R2 test	-	- 0.04	0.00	- 0.16	0.18	-0.41
RF % importance time / dose	50 / 50	50 / 50	44 / 56	44 / 56	47 / 53	56 / 44

Table 4-6: Results from multiple regression models using time and dose as explanatory variables to predict saliva production as outcome. The three bottom rows are results from random forest regression models, the rest from linear regression. The models were trained and evaluated on either all data (left column), or a 5-fold split into training and test.

In addition to the 5-fold CV, a leave-one-out cross-validation (LOOCV) was performed. For each observation left out a RF regressor was fitted 100 times alongside a linear regression model fitted once. No HP-tuning was done for the RF model.

	LOOCV linear reg	LOOCV RF reg	5-fold test average linear reg	5-fold test average RF reg
R2	0.09 ± 0.00	0.08 ± 0.00	0.02 ± 0.11	-0.08 ± 0.20
MSE	2799 ± 0	2825 ± 6	2974 ± 633	3348 ± 1059

Table 4-7: LOOCV and average 5-fold test results for linear regression and RF regression on saliva data using time and dose as predictors. Top row: coefficient of determination (R2), bottom: mean squared error (MSE). The uncertainty for the LOOCV results is due to the small variation in repeated RF modelling, while the higher variation for the 5-fold CV is the standard deviation of the results across the folds.

Comparing the mean squared error (MSE) and coefficient of determination (R2) from the LOOCV to the averaged results of the 5-fold CV, seen in Table 4-7, LOOCV yields better scores (higher R2 and lower MSE) than the averaged 5-fold for both the RF- and linear regressor. This is expected due to the relation between low bias and the LOOCV (section 2.4.1). Linear regression also performs better on unseen data in the LOOCV as is the case for 5-fold CV.

4.4.2 Image features as explanatory variables

Training data from three feature spaces (no-p T1 & T2 along T2 delta-p) was used to fit regression models using saliva amounts as the continuous outcome variable, measured on the same days as the images were taken. The features from the left and right SMG unit were aggregated column-wise and thus treated as separate features (LR-aggregated).

An exception was made for images taken at day 70, where saliva measured at day 75 was used for prediction being the closest in time. The regression models used were either a multiple linear regressor or a random forest (RF) regressor.

The single split between training and test data was based on individual mice as described in section 3.6.1.1 summarized in Table 3-4, with the corresponding sample sizes seen in Table 4-8. Individuals may have both imaging and saliva data from multiple days which was treated as separate observations within each split, but were all in either the training or test set (e.g. two time-points for mouse C2 would both be in the training set).

	All data (control)	Training (control)	Test (control)
No-p T1	69 (33)	55 (26)	14 (7)
No-p T2	140 (61)	113 (50)	27 (11)
Delta-p	69 (28)	54 (23)	15 (5)

Table 4-8: Number of data points for the three feature-spaces used in regression models predicting simultaneously measured saliva. Number of control individuals in parenthesis.

Either only time and dose were used as predictors in the model (td-model), or the best 5, 10, or 15 image features selected by MRMR (3.6.3) from the various feature-spaces. Lastly all image features were used in the model without any selection (excluding the features dropped in FSPS, section 3.4.1). The coefficient of determination (R2) was calculated using the fitted training data and the hold-out test data, separately, with the resulting values seen in Figure 4-17.

Figure 4-17: Coefficient of determination, R^2, calculated from regression tasks done on various features spaces using multiple linear or random forest regression. The model was evaluated on the data used for training (left), and the hold-out testing data (right). The models were created using either only time and dose (td), or the best 5, 10, 15, or all, image features.

While both the linear regression and random forest models managed to fit the training data with some high R2 scores close (or equal) to 1.0 in Figure 4-17, the results on the test data were seen to be below or close to zero for almost all models. Across all three subsets of the predicted outcome (varying between the feature-spaces) the td-models had the highest R2 when evaluated on the test data. In some cases the R2 was higher for test than training data (linear regression no-p T2 td, and both regressors for delta-p td).

Using all N data-points available (left column in Table 4-8) combinations of regression models (random forest or multiple linear regression), feature spaces (no-p T1, T2, or delta-p), and LR-modes (aggregated or averaged features between the LR-subunits) were evaluated using leave-one-out cross-validation (LOOCV). For each instance left out the remaining $N - 1$ data-points were used for the selection of 5 image features, hyperparameter tuning, and training. The predicted saliva amounts for both the td- and feature-model were saved for all left-out instances and used to calculate the coefficient of determination for each model (Figure 4-18). Due to the random nature of the random forest regressor (bagged decision trees, see section 2.4.3.3) each RF model was fit 100 times to the training data, from which the average and standard deviation of the predicted values were used to create an uncertainty interval for the R2 scores.

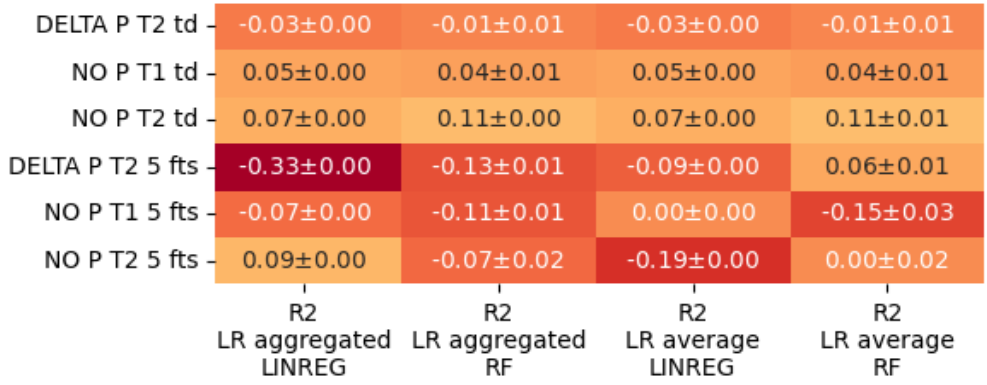


Figure 4-18: Coefficient of determination (R^2) for leave-one-out cross-validated regression models (linear regression or random forest), using various feature spaces.

Overall, the R2 values were close to or below zero with the highest score at 0.11 belonging to the no-p T2 RF td-models (LR-aggregated or average makes no difference with respect to time and dose). The no-p T2 image features max out at 0.09 using LR-aggregated features in a multiple linear regression model, and the delta-p features obtained R2 = 0.06 using a LR-average RF model. The no-p T1 features tops out at $R^2 = 0.00$.

4.4.3 Testing the added predictive ability with best radiomic features

To evaluate the significance of any potentially increased performance of using image features in combination with time and dose, compared to only time and dose, new models were created using the top k features from each LOO cross-validated model in section 4.4.2. The top k features were defined as being picked among the top 5 (by MRMR) in at least 50% of all feature selections across the LOOCV, in a similar fashion as done in [44]. The top 5 and 4 features for the two top-performing models across the LOOCV are seen in Figure 4-19, with similar plots for all models in Appendix E.

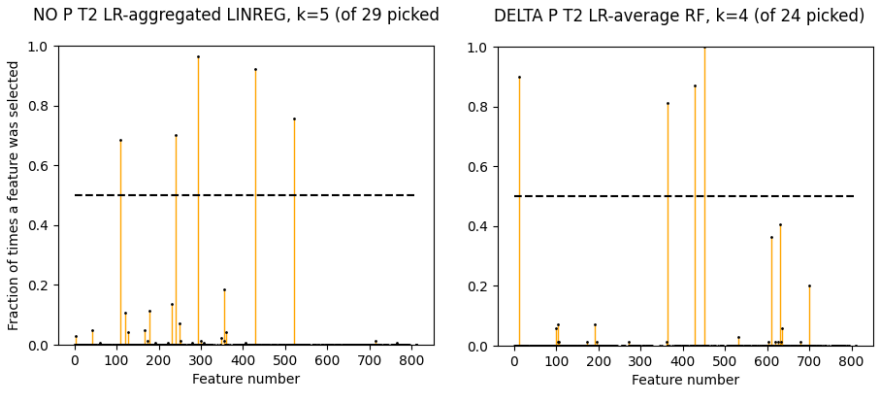


Figure 4-19: The k best features for the two best performing image feature-models in the LOOCV. The k best features are defined by having been selected in at least 50% of the left-out instances across the LOOCV (black stippled horizontal line).

For each 12 combinations of feature-spaces (3), LR-modes (2), and regressors (2), $k + 1$ models were created in addition to the td-model. A model using all best k features, and k models using time and dose in addition to one of the k features. A LOOCV was performed for each model, predicting continuous saliva production for each hold-out observation across the CV. The squared error between the model predictions and the ground truth (measured saliva production) was calculated for each observation.

For each set of squared errors, corresponding to a feature model, a one-tailed paired t-test was performed against the null hypothesis that the MSE for the td-model was lower or equal to the MSE of the given model ($H_0: MSE_{td} \leq MSE_i$). Of 52 tests performed 4 was significant, seen in Table 4-9. The MSE from the LOOCV where 5 features were selected by MRMR for each observation left out (from section 4.4.2) is also seen as MSE MRMR, for comparison.

Under the Bonferroni correction the significance threshold for each t-test is scaled from $p < 0.05$ to $p < \frac{0.05}{m}$ (m being the number of tests performed) to account for random rejections of null hypothesis – being more probable with increasing m. Using the scaled threshold $p < 9.6 \times 10^{-4}$ none of the tests were significant.

	Ft-space	LR-mode	Regressor	MSE MRMR	k	MSE td	MSE	Relative change MSE	p-value MSE td > MSE
Td + ft297R	Delta-p	Aggregated	Linear reg	3707	4	2888	2452	-0.151	.028
Td + ft12	Delta-p	Average	Linear reg	3042	4	2888	2314	-0.199	.044
Td + ft522R	No-p T2	Aggregated	Linear reg	1829	5	1871	1694	-0.095	.039
Td + ft240L	No-p T2	Aggregated	Linear reg	1829	5	1871	1606	-0.142	.018

Table 4-9: MSE for time + dose models (MSE td) compared to time + dose + a single feature (MSE), by a one-tailed paired t-test.

The four features significantly improving the td-models all had the same sign for the estimated regression coefficients across the LOOCV. The regression coefficients had a coefficient of variation below 0.1 for all four features, indicating a stability of the features.

4.5 Classification of binary xerostomia outcomes

All measured saliva values were divided into one of two groups with xerostomia either being true or false using the thresholding method described in section 3.6.2. Using this as the binary outcome variable, classification tasks were performed using random forest classifiers. The

explanatory variables were either only time and dose, or a subset of the image features, when not specified otherwise. The potential image features used in each model were all surviving features following the feature-specific preprocessing selection (FSPS, section 3.4.1), varying between being features from T1- or T2-weighted images, and whether the features from the left and right unit of the SMG were aggregated as separate features (LR-aggregated) or averaged (LR-average).

The best 5, 10, and 15 features were selected from the training data for each feature space using the maximum relevance minimum redundancy (MRMR) algorithm described in section 3.6.3. The available feature spaces used for modelling were either T1 or T2 features before pilocarpine injections (no-p T1 / T2), delta-p features (section 3.5.1), or delta-features (section 3.5).

4.5.1 Prediction of simultaneous xerostomia using time and dose

Similarly to the regression to all saliva measurements (N=347) in section 4.4.1, time and dose are now used as explanatory variables to predict the probability of xerostomia. A logistic regression model and a random forest (RF) classifier were initially fit to all data without any split into training and test. In addition to models using time and dose as variables, a second-degree interaction term between time and dose was added for two models - making up four models in total. Classification metrics with respect to the ground truth, in addition to various p-values, is seen in the left part of Table 4-10 with the right part being results from a leave-one-out cross-validation (LOOCV) evaluation of the models. In the LOOCV each fit was repeated 100 times to estimate variations in the RF classifier. Uncertainties in each metric is reported as the max difference to the mean metric using the 25^{th} and 75^{th} percentile of the estimated class probabilities. The area under the receiver operating characteristic curve (AUC), and the brier score (BS) were calculated for model evaluation.

Using the log-likelihood ratio (LLR) between the null model (having only intercept) and the full model (with estimated coefficients) as a z-statistic, the p-values for the logistic regression models were calculated (p-val LLR). Both models, with and without the interaction term, were significantly ($p < .000$) better performers relative the intercept-only null models. For each estimated coefficient the p-values were calculated for both logistic models, showing significant dose coefficients ($p < .05$) and insignificant time coefficients. While the interaction term in the logistic regressor yielded better model scores both when fit to all data and in the LOOCV relative no interaction term, the difference is negligible for the random forest classifier.

	All data (no split)							LOOCV		
	Acc	AUC	BS	P-val LLR	P-val time	P-val dose	P-val interaction	Acc	AUC	BS
Logreg	.69	.65	.20	<.000	.697	<.000	-	.68	.55	.21
Logreg w/ inter	.72	.64	.20	<.000	.162	.027	.031	.70	.61	.20
RF reg	.77	.80	.16	-	-	-	-	$.70 \pm .03$	$.67 \pm .01$	$.20 \pm .00$
RF reg w/ inter	.77	.80	.16	-	-	-	-	$.71 \pm .03$	$.67 \pm .01$	$.20 \pm .00$

Table 4-10: Metrics and p-values for td-models classifying xerostomia. Left: models trained and evaluated on all data. Right: models evaluated by LOOCV.

The RF model scores lower AUC than logistic regression when fit to all data and across the LOOCV. The BS was lower for the RF models when fit to all data but changed little between models in the LOOCV.

4.5.2 Prediction of simultaneous xerostomia using image features

Image features of various types were used to predict xerostomia corresponding to saliva measured at the same day the MR-images were taken (referred to as simultaneously), in a similar fashion as in section 4.4.2. Models were created using either no-p T1, no-p T2, or delta-p features with a random forest classifier.

Each model was trained using the feature-space specific training data following the split in Table 3-4. Evaluating each model, based on either feature space, was done by bootstrapping the test data 1000 times (inspired by [74]) – resulting in a similar amount of estimated class probabilities. The average of the corresponding AUCs $\pm$ the standard deviation of the AUCs for each model is seen in Figure 4-20. Across the feature-spaces and LR-modes the td-models have the highest average AUC, except for the no-p T1 model using 5 selected LR-average features. Models using LR-average features is seen to generally have higher AUCs than using LR-aggregated features. The number of available data points for each model, being the same for LR-average and LR-aggregated feature-based models, is seen in Table 4-11.

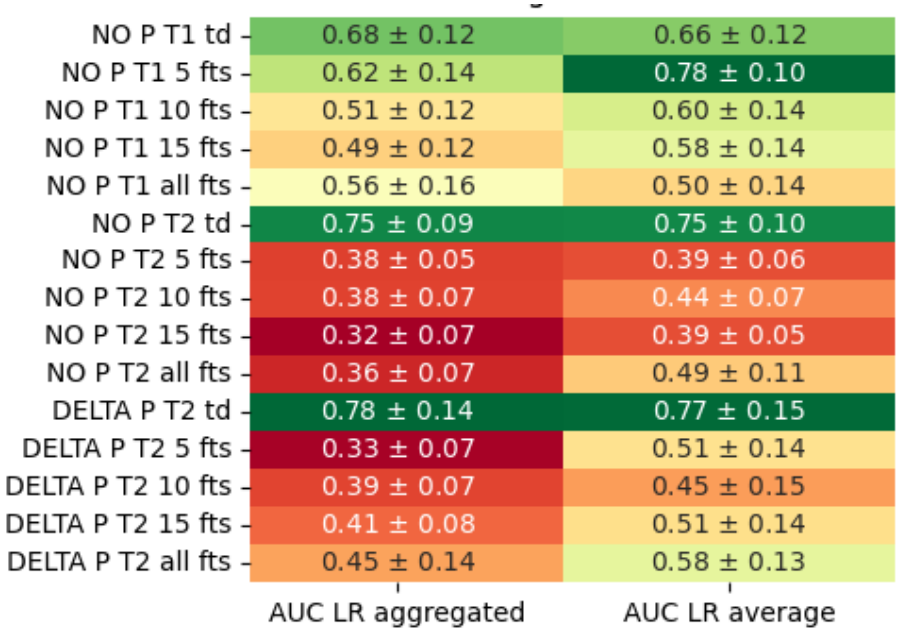


Figure 4-20: $\mu_{AUC} \pm \sigma_{AUC}$ for random forest models evaluated on the test set, bootstrapped 1000 times. Binary xerostomia values corresponding to saliva measured at the same day as the images were acquired, were used as the outcome variable.

	All data	Training data			Test data		
	Total (xer)	Total (xer)	# control (xer)	# irr (xer)	Total (xer)	# control (xer)	# irr (xer)
No-p T1	69 (31)	55 (24)	26 (7)	29 (17)	14 (7)	7 (3)	7 (4)
No-p T2	140 (58)	113 (48)	50 (16)	63 (32)	27 (10)	11 (4)	16 (6)
Delta-p T2	69 (26)	54 (22)	23 (9)	31 (13)	15 (4)	5 (2)	10 (2)

Table 4-11: Number of available data points (image features with corresponding binary xerostomia outcome) for training and testing for each feature-space. Number of xerostomic individuals in parenthesis.

LOOCV were used for model evaluations in addition to the single split into training and test data. Feature selection and hyperparameter tuning was done using the $N - 1$ training observations, repeated for all N left-out observations in the LOOCV. To estimate the standard deviations of the models a 1000 repeated bootstrap of each training set was used for model fitting before class probability estimation, producing the results seen in Figure 4-21.

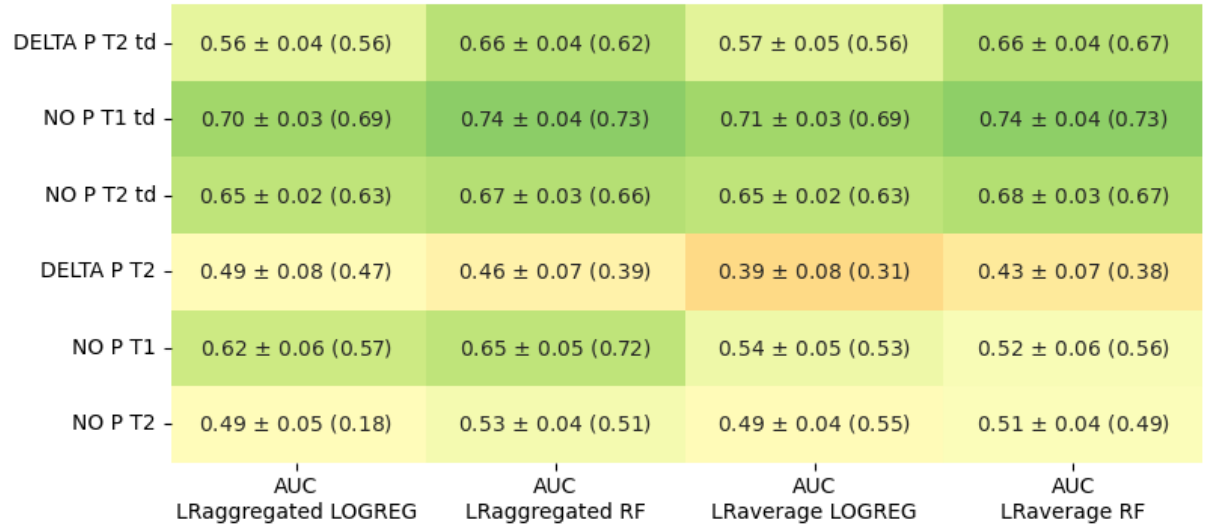


Figure 4-21: ROC AUC scores for LOOCV using image features or time + dose (td) as explanatory variables to predict simultaneous xerostomia. For each observation left out the training set was bootstrapped 1000 times from which $mean(AUC) \pm sd(AUC)$ were calculated. In parenthesis: AUCs from a LOOCV without any bootstrapping or repetitions.

4.5.3 Prediction of late xerostomia using features from earlier days

Random forest classification models were created using radiomic features extracted from MR-images taken either before any irradiation (baseline), after irradiation (after-irr), or a combination of the two (delta-features). The binary xerostomia outcome data for classification is now the *latest* saliva measurement acquired for each individual (varying between day 35 and 75), for evaluation of the radiomic features' ability to predict xerostomia forward in time. The models were created using either LR-aggregated or LR-average image features from one of four feature-spaces: either T2 or T1 no-p features, T2 delta-p features, or T2 delta-features. The top features were selected from each feature-space (after FSPS) using MRMR, and used as predictors in the RF models. Time and dose were omitted before any selection and used in a separate model for comparison (td-models). Concerning the delta-feature based td-models, the time variable is meaningless - leaving only dose as the explanatory variable.

The number of available data points for training and testing each model is seen in Table 4-12, based on the split on individuals (Table 3-4) stratified with respect to control as described in section 3.6.1.1. The sample sizes vary between the feature spaces and prediction modes.

	All data		Training data		Test data	
	All	Control	All	Control	All	Control
No-p T1 baseline	24 (14)	10 (3)	19 (12)	8 (3)	5 (2)	2 (0)
No-p T1 after irr	26 (12)	14 (3)	21 (11)	11 (3)	5 (1)	3 (0)
No-p T2 baseline	54 (29)	23 (7)	42 (23)	18 (6)	12 (6)	5 (1)
No-p T2 after irr	42 (20)	19 (3)	35 (18)	16 (3)	7 (2)	3 (0)
Delta-p T2 baseline	27 (11)	12 (4)	20 (8)	9 (3)	7 (3)	3 (1)
Delta-p T2 after irr	18 (6)	7 (0)	14 (5)	6 (0)	4 (1)	1 (0)
Delta-features T2	38 (19)	16 (3)	31 (17)	13 (3)	7 (2)	3 (0)

Table 4-12: Number of observations (individuals) available for prediction of late xerostomia, from each feature-space and prediction mode, along numbers of observations reserved for training and testing. Number of xerostomic observations seen in parenthesis.

The average ROC AUC $\pm$ standard deviation of the AUCs for each model is calculated from 1000 evaluations on bootstrapped test sets using the top 5 features, seen in Figure 4-22. The same metric is shown for delta-feature models, with a varying number of selected features, in Figure 4-23. Additional models using more than 5 features may be found in Appendix C.

Image features from baseline is seen to be generally worse predictors than from after irradiation. The no-p T2 features have the worst performance with maximum AUC = 0.60, compared to no-p T1 and delta-p (both having AUC > 0.80). Some predictive ability is seen for baseline features on the T1 data, as both the td- and image feature-models have average AUCs close to 0.70.

Using features from after irradiation both T1 and delta-p image features have average AUCs above 0.80 indicating good performance. The td-model using the T1 data have a very high AUC (0.95), with the td-models using the T2 and delta-p data being a little lower but above 0.80. The LR-aggregated delta-p features have exactly the same average AUC as td. The LR-averaged T1-features have a somewhat higher performance than aggregated (average AUC at 0.88 and 0.75), while the LR-aggregated delta-p features perform much better than LR-average (AUCs at 0.82 and 0.51).

The delta-features are seen to have the worst predictive abilities in the single split, with maximum AUC at 0.50.

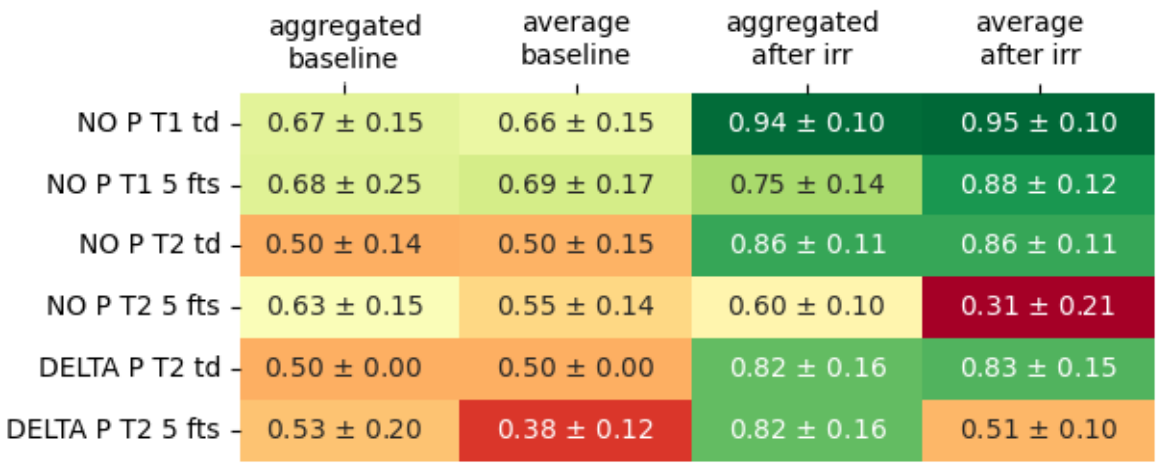


Figure 4-22: ROC AUC values from evaluating each model on the test set, bootstrapped 1000 times. The presented values are the average AUC ± the standard deviation of the AUCs, from predictions made on the bootstrapped test sets for each model.

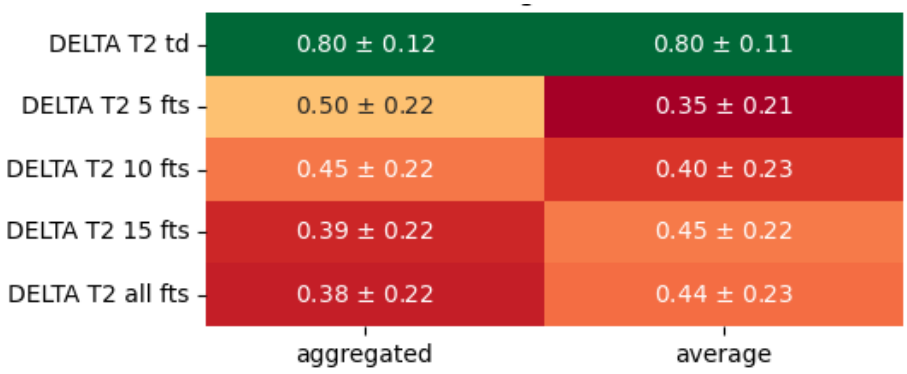


Figure 4-23: ROC AUC values for evaluating delta-feature based models on the test set, bootstrapped 1000 times. The heatmap values are the average AUC ± the standard deviation of the AUCs, for each model.

As an alternative to only using a single train / test split, LOOCV was used to evaluate models based on the four feature-spaces. Each observation is left out of training and used for testing once, repeated for all observations (left column in Table 4-12).

For each observation left out for testing the remaining data were used for feature selection, hyperparameter tuning, and training. The number of times a feature was selected across the LOOCV for a model was counted. Either a random forest classifier or logistic regression classifier with l_2 penalization was used (section 2.4.5), from which the classification probability given each test observation across the LOOCV was saved.

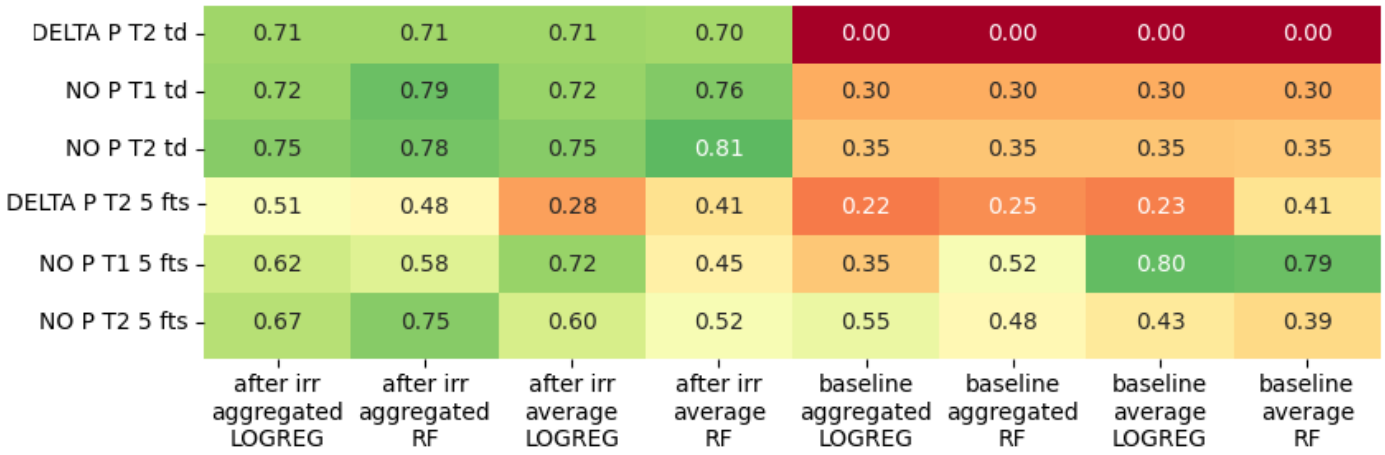


Figure 4-24: ROC AUC for LOOCV evaluated models using image features from days right after irradiation, or baseline, to predict late xerostomia.

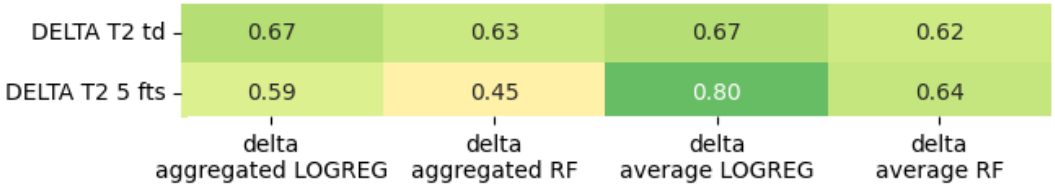


Figure 4-25: ROC AUC for LOOCV evaluated models using delta-features to predict late xerostomia.

Generally across the LOOCV-evaluated models, using baseline features is seen to yield lower AUCs than using features from after irradiation (Figure 4-24). An obvious exception is for the baseline no-p T1 LR-average models, having AUC = 0.80.

The td-models outperform all models using after-irr features, except for the delta-feature LR-average logistic regression model having AUC = 0.80 (Figure 4-25). Among the after-irr models, no-p T2 LR-aggregated with a RF classifier performs the best with AUC = 0.75, followed by no-p T1 LR-average with a logistic regressor (AUC = 0.72). Using delta-p features scored AUCs equal to or below 0.51 in the LOOCV.

4.5.4 Comparing T1- and T2- based feature models on the same subset of data

To allow for a fair comparison between the predictive abilities of radiomic features from T1- or T2-weighted MR-images, the instances (same individual mouse at the same time) where images of both modalities were acquired were registered and split into three test / train sets as described

in section 3.6.1.2 and summarized in Table 3-5. Only no-p features were considered, with the feature from the left and right SMG units aggregated as separate features (LR-aggregated).

The features were evaluated on predicting simultaneous xerostomia, or late xerostomia using baseline or features from after irradiation. The number of available data for training and testing varies between the prediction modes as seen in Table 4-13.

	All data		Test split 1		Test split 2		Test split 3	
	Tot	Control	Tot	Control	Tot	Control	Tot	Control
Simult.	69 (31)	33 (10)	23 (13)	13 (6)	23 (10)	9 (2)	23 (8)	11 (2)
Late baseline	24 (14)	10 (3)	8 (5)	4 (1)	8 (4)	3 (1)	8 (5)	3 (1)
Late after irr	26 (12)	14 (3)	8 (4)	4 (1)	8 (4)	4 (1)	8 (3)	4 (0)

Table 4-13: Size of the data sets available where both T1 and T2 images are present (all data), and the three test sets, for prediction of simultaneous or late xerostomia. Number of xerostomic observations in parenthesis.

Models were trained on features from either T1 images, T2 images, or using both (aggregated as separate features, denoted T1 + T2 or *combined*). Time and dose were omitted before any feature selection and used in a separate model (td). The top 5 features from each feature-space were used in a random forest classifier model. The time + dose, T1, T2, and combined models were evaluated on the same test set for a given prediction mode and split bootstrapped 1000 times. The average ROC AUC with the standard deviation of the AUCs, combined for the three splits, for the models is seen in Figure 4-26. The AUCs for the three train / test splits separately is seen in Appendix D.

When predicting xerostomia using image features from simultaneous time-points, only using time and dose as explanatory variables produces the highest average AUC across all splits at 0.79. Only using T2 features have the lowest AUC, and an improvement is seen when using both T1 and T2 features compared to only T1 features. Between the 1000 AUCs calculated from each model, due to the bootstraps, paired t-tests were performed (paired as the bootstrapped test was the same for the models). The simultaneous T1 model had significantly higher AUC than the T2 models ($p < .000$).

Predicting late xerostomia using baseline features yields lower AUC's than using features from after irradiation in every model. While baseline models using T1 features and time and dose have average AUCs above 0.5, the distributions of AUC's have a high variability with a maximum at 0.5 for the td-model as seen in Figure 4-26.

Regarding predicting late xerostomia using features from after irradiation the td- and combined model perform the best with an average AUC between all splits of 0.81, which is significantly greater than the T1- and T2- based models when comparing all AUCs with paired t-tests ($p < .000$). The T1-model have a somewhat higher average AUC than the T2-model ($p < .000$), even though the distributions look very similar (Figure 4-27). The difference in average AUC between

the top-performing td- and combined models is not significant with $p = 0.18$, indicating a similar performance across the bootstrapped test sets.

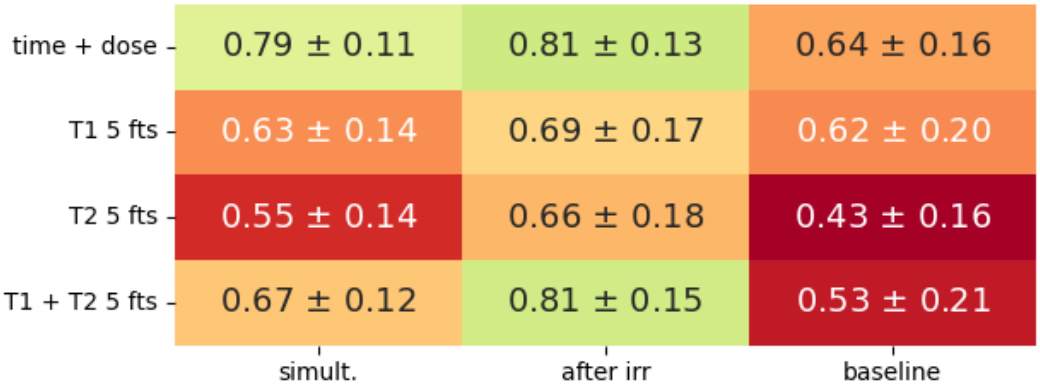


Figure 4-26: Performance of varying models applied to the three splits for each prediction mode, given by the AUC for the corresponding model on the given test set (bootstrapped 1000 times). Each cell value is the average of all AUCs from the three splits ± the standard deviation of the AUCs.

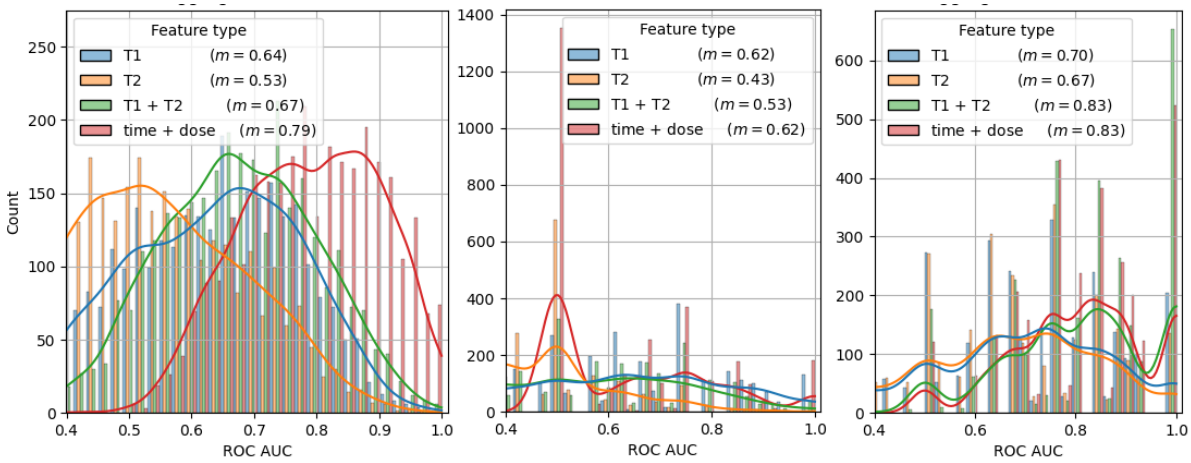


Figure 4-27: Distributions of all AUCs from 1000 bootstrapped test sets in all three splits. Prediction of simultaneous xerostomia in left plot. Prediction of late xerostomia using baseline features in middle plot, and features from after irradiation in the right plot. Median of all AUC's given a feature type is seen in the legend as m.

In addition to the 3-fold split, LOOCV was used to evaluate the models. The results are seen in Figure 4-28 as ROC AUC and BS. Paired t-tests were performed between the models, for each prediction mode, using the squared differences between the estimated class probabilities and the ground truth for each observation (which, when divided by the number of observations is equal to the BS). The BS for the T2-models was significantly different ($p < 0.05$) than all other models for simultaneous prediction. For late predictions using baseline only the BS between the T2- and td-model was significantly different. While the T2-based and combined models are seen

to perform the best using after-irr features (highest AUC and lowest BS), the brier scores were not significantly different to the T1- or td-model.

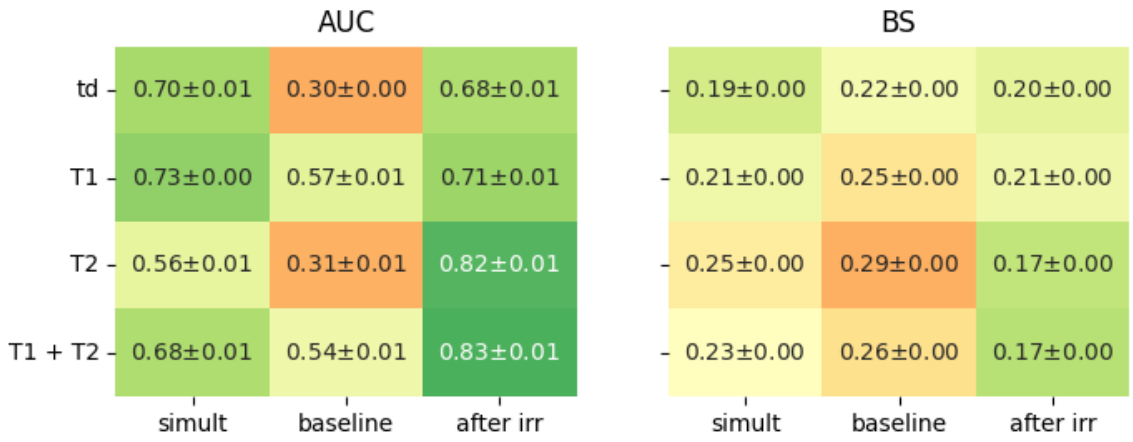


Figure 4-28: ROC AUC and BS for LOOCV evaluated models using either T1, T2, or both feature types to predict simultaneous or late xerostomia. All models used a random forest classifier and LR-aggregated features.

4.5.5 Testing the added predictive ability of radiomic features to time and dose for xerostomia classification

For each combination of feature-spaces and LR-modes a logistic- and RF-classifier was trained on either only time and dose (td), or time and dose in addition to one of each of the k best features from the LOOCV done in sections 4.5.2 and 4.5.3 (based on whether to predict simultaneous or late xerostomia). This analysis is similar to the analysis done for regression models in section 4.4.3, now performing paired t-tests between squared differences between the estimated class probabilities and binary xerostomia outcome for each observation. The mean of all such squared errors is per definition the brier score (section 2.4.6), such that the p-value from the t-tests corresponds to the probability of the null hypothesis $H_0: BS_{td} \leq BS_{td+ft_i}$ being true. All available data is used for each LOOCV, based on what feature-space is utilized and the prediction mode (simultaneous: Table 4-11, late: Table 4-12).

For prediction of simultaneous xerostomia the 12 combinations of feature-spaces, LR-modes, and classifiers yielded 46 tests in total (with k varying) of which 5 were significant ($p < 0.05$). For prediction of late xerostomia 20 of 86 tests were significant.

Using Bonferroni corrected significance thresholds ($p < 1.08 * 10^{-3}$ for simultaneous prediction, and $p < 5 \times 10^{-4}$ for late prediction) yielded no significant tests for either prediction mode.

While the significant tests indicate a model improvement based on BS (a lower value compared to the td-model) the actual change is seen to be vanishingly small in some cases. A second test criterion was therefore added: in addition to a significant t-test, the negative relative change in

BS should be above 1% relative to the td-model. This additional criterion resulted in 2 feature-models with increased performance for simultaneous predictions seen in Table 4-14, and 9 for late predictions seen in Table 4-15. Each feature-number corresponds to the filter and type seen in the feature-register in Appendix C (with R/ L added for the LR-aggregated features). The AUC and BS from the original from which the k best features are selected are seen in the column denoted AUC / BS MRMR.

Td + $ft_i \downarrow$	Ft-space	LR	Classif	AUC / BS td	AUC / BS MRMR	K	AUC / BS	Sign perc	p-value BS td > BS	Rel diff BS
Ft340R	Delta-p	Agg	Logreg	.564 / .239	.467 / .274	2	.596 / .235	+100	.026	-.018
Ft635R	No-p T1	Agg	Logreg	.686 / .200	.572 / .259	4	.731 / .194	+80	.003	-.029

Table 4-14: Feature improving the prediction of simultaneous xerostomia when added as an explanatory variable to time and dose. The squared difference between estimated class probabilities and binary outcomes across the LOOCV is significantly lower compared to only time and dose ($p<0.05$), in addition to at least -1% relative change in the BS.

Td + $ft_i \downarrow$	Ft-space	LR	Late-mode	Classif	AUC / BS td	AUC / BS MRMR	K	AUC / BS	Sign perc	p-val BS td > BS	Rel diff BS
Ft2	Delta-p	Avg	Baseline	Logreg	.000 / .260	.222 / .548	2	.119 / .254	-100	.014	-.026
Ft498	No-p T1	Avg	Baseline	Logreg	.300 / .215	.804 / .211	5	.300 / .212	-100	.007	-.017
Ft678R	Delta-p	Agg	After irr	Logreg	.708 / .217	.514 / .320	4	.708 / .213	100	.045	-.016
Ft216R	No-p T2	Agg	After irr	RF	.795 / .203	.748 / .204	4	.943 / .098	-	.006	-.518
Ft408	No-p T2	Avg	After irr	RF	.789 / .202	.520 / .304	5	.878 / .146	-	.040	-.278
Ft765	No-p T2	Avg	After irr	RF	.789 / .202	.520 / .304	5	.874 / .144	-	.020	-.290
Ft759R	Delta	Agg	Delta	Logreg	.670 / .207	.590 / .279	1	.668 / .204	-100	.041	-.014
Ft1	Delta	Avg	Delta	Logreg	.670 / .207	.790 / .190	4	.704 / .196	100	.009	-.053
Ft1	Delta	Avg	Delta	RF	.654 / .247	.639 / .258	4	.843 / .157	-	.036	-.363

Table 4-15: Features improving on the late prediction of xerostomia when used as an explanatory variable in addition to time and dose.

5 Discussion

As mentioned in the introduction, the radiomic feature were evaluated on their ability to predict the measured saliva production in individual mice. Both the image features and the measured saliva amounts displays high variabilities in general, which is discussed in sections 5.1 and 5.2. Considerations about, and errors relating to, segmentation and the developed radiomic pipeline is discussed in sections 5.3 and 5.4. The models are evaluated in section 5.5, and the top-performing features are interpreted biologically in section 5.6.

5.1 Major sources of error

The repeatability and reproducibility of features and results is a major issue in radiomic studies. Finding robust features across scanner variations and acquisition protocols, along methods of extraction, selection, and modelling, is critical before any radiomic feature may be incorporated into patient-specific precision oncology [6]. As such, understanding the major error sources in any study using radiomics is highly necessary.

An irreducible error affecting every step of the pipeline, from ROI segmentation to extracted feature values and thus predictions, is artifacts and noise present in the MR-images as discussed in section 2.2.5. The level of noise and blurriness varies much between the images but is generally not very noticeable.

Many of the images contains a horizontal noise pattern in the lower part of the images as seen in both examples in Figure 5-1. Some images contain very periodic parallel horizontal lines (sometimes only a single row), often in the upper parts of the images, as seen in the left example image. The periodicity of the latter artifact implicates that the artifact may be some sort of Gibbs artifact, and the high contrast area close to the lines in the example image supports this.

As seen in the right image in Figure 5-1, the upper part of the head of the mouse disappears almost completely. The vanishing signal may be due to some susceptibility artifact, or an extreme case of the low-frequency bias field. Additionally, the lower part of the mouse is seen to exhibit a very strong signal overall. While the bias field is mostly harder to detect with the naked eye, it is assumed to always be present in some manner – hence the application of the N4 correction to all images. The after-p images were in general more affected by artifacts, since the anesthesia wore off and some mice started to wake up and move during the second round of image acquisitions (section 3.1).

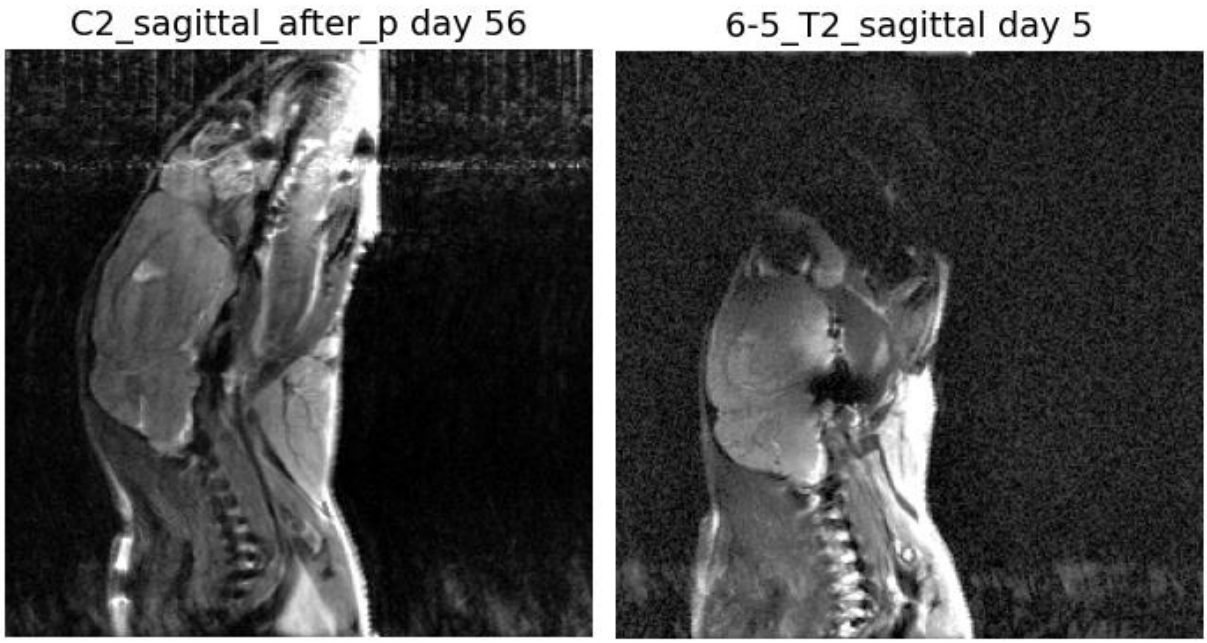


Figure 5-1: Visible artifacts in T2-w images, at a central SMG slice. Left: image showing Gibbs ringing in the upper part. Right: Image showing bias field artifact along generally high noise, and potentially a susceptibility artifact in the centre. Both images also show some artifact in the lower part of the images.

The semi-automatic segmentation procedure necessitates user input and is as such heavily influenced by an observer bias from the author of this thesis. Choosing what slice to consider the central slice for both the left and right SMG unit was in some cases difficult, and largely based on maximizing the area of each left or right unit of the gland with some separation between the two central slices. As the slice distances are much larger than the pixel distances, the true center of the gland units may have ended up between two slices creating a partial volume effect varying between each image acquisition series. Positions of the mice in the scanner might have changed between the acquisitions - slightly changing the orientation or causing some deformation of the glands due to internal pressure. As the SMGs and SLGs are fused in mice, they were hard to differentiate in the images. The segmented ROIs are assumed to contain more of the SMG than the SLG due to its larger size in the mice. The shape-based features are only dependent on the ROI, and as such heavily influenced by errors in the segmentations. If anatomy from outside the SGs are included in the ROIs the segmentation might have affected the first-order or texture-based features as well. A more in-depth discussion regarding the validity of the segmentations is found in section 5.3.

The measured saliva values are seemingly very noisy with a high variance for both control and irradiated individuals. Potential causes for this variability are discussed more in section 5.2. A high error in the saliva measurements have direct consequences for all machine learning based analysis in this thesis, being the outcome for all models. The features are also selected based on their relations to the outcome, as well as the feature-specific preprocessing (FSPS).

Due to the positionings of the mice during irradiation, the left and right unit of the SGs might have received different doses. Assuming a negligible build-up region, for photons in the keV energy range, the left unit received a higher dose than the right due to photon attenuation (as the mice are positioned on their right side during irradiation). While the dose difference was assumed to be negligible, the dose contributions to the two units were not quantified.

As with all radiomic studies, the number of features is much larger than the number of observations (curse of dimensionality). If all features were random values, i.e. just noise, one might still pick up seemingly good predictors due to the random variations corresponding to the outcomes (multiple comparisons problem). This issue further emphasizes the need for external validation. While Bonferroni corrections were used in sections 4.4.3 and 4.5.5, with no significant tests after the correction, a "proper" correction might take into account every feature as they are all evaluated to the outcome in the FSPS –leading to a p-value threshold magnitudes below 0.05.

The choice of method for feature selection has a very large impact on the final models, especially when reducing the feature-space to five explanatory variables as were done for the majority of the models.

Evaluating the models using a single split into training and test data may yield overly optimistic or pessimistic generalization results due to random separation between individuals exhibiting outlier behaviours – either from the measurements itself, or a very high inter-mice variability. This is discussed further when comparing the 5-fold split to the leave-one-out cross-validation (LOOCV) in section 5.5.1. Due to the small number of observations available for the models, especially when predicting late xerostomia, the LOOCV might yield more overall accurate results. However, the LOOCV is known to produce a low bias (overfit) and have a high variance (section 2.4.1) which might overestimate the generalization scores, but the ratio of outlier individuals in training compared to test disappears as all observation are used for testing once. When considering simultaneous regressor or classification, using a LOOCV necessitates using some individuals for both training and test (for each individual having measurements for more than one day) which may lead to overly optimistic generalization results.

Concerning the classification tasks, the method used for thresholding continuous saliva measurements into a binary response variable (section 3.6.2) must be discussed. Due to the high variance among the saliva measurements, the thresholding was based on population estimates from the control individuals rather than on an individual basis taking the baseline measurements from each mouse into account. The thresholding, transforming the machine learning task from a regression to a classification problem, becomes a trade-off between whether signal-containing information is destroyed by such a coarse binary re-binning, or whether the noise is suppressed allowing for more achievable learning tasks given the small sample size.

At certain points in the radiomic workflow presented in this thesis all the available data have been used for some decision or estimation, creating a potential data leakage further downstream the workflow. The first instance was for training the standard scales in the Nyul normalization algorithm (section 3.3.2.3), followed by the choice of optimal bin width for fixed-bin discretization (section 3.3.4) and in the feature-specific preprocessing selection (section 3.4.1).

Data leakage may also have affected the choice of top-performing features used in combined models with time and dose (sections 4.4.3 and 4.5.5) where the top k features was selected from across the LOOCVs for further analysis – effectively using all data in an embedded feature selection.

5.2 Saliva production in control and irradiated mice

Due to the young age of the mice at baseline, one might expect their continued growth to increase the saliva production over time when undisturbed by irradiation. While the salivation measurements are quite noisy and no pattern is obvious, this assumption is supported by the significant correlation between time and saliva for the control mice seen in Figure 4-2 ($\rho = 0.29$ with p = 0.001) and the significant paired t-tests between saliva measurements for control mice at baseline or immediately post-irradiation (denoted acute) to the latest time-points seen in Table 4-3 (section 4.1.1, both p-values < 0.005). The linear regression done on all mice having received no dose for xerostomia thresholding is seen to be significant (p = 0.008) in section 4.1.2 with a positive slope ($0.56 \times \text{day}$) also supports the assumption of increased average saliva production over time for the control mice.

The saliva production for the irradiated mice was seen to be significantly negatively correlated with the total dose delivered ($\rho = -0.34$ with p < 0.000), while the correlation to time was not significant - indicating that the irradiation of the SGs may have halted the growth of functional tissue in, and/ or induced damage to, the SGs. The significant difference in saliva production between baseline and acute time-groups supports earlier studies describing a reduction in saliva following irradiation (section 2.1.4) in both HNC patients treated with RT and rats. However: the saliva values between baseline and late is close to but not significant (p = 0.07). This might be due to some healing of the functional tissue in the SGs or that the SGs continues their growth supplying new healthy tissue. If the latter is the case, the significant difference between measurements from irradiated and control mice in the late time-group indicates that the growth have been at least halted in the irradiated mice. It could of course also be an error due to the high level of variance (assumed to be noise) in the saliva measurements: the median is seen to be at its lowest for the irradiated mice in the late time-group but with a very high CV as seen in Table 4-1 and Figure 4-3.

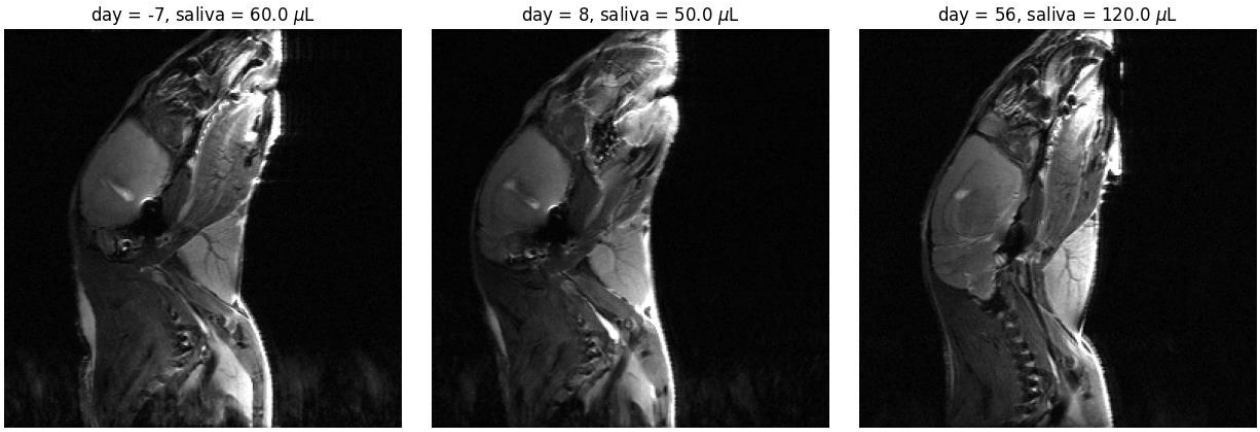


Figure 5-2: Control mouse C2 displaying no visible increase in SG size over 63 days (day -7 to 56).

While the measured saliva values from control mouse C2 increase over time (60 at baseline to 120 at day 56), no increase is seen in the SG size by visual inspection of Figure 5-2. As this is only a single data point no conclusion is inferred by this inspection relating to the correlation between SG size and saliva production. The temporal evolution of the ROIs is discussed in section 5.3.

As a thought experiment the high variation in the saliva measurements may be attributed to three major factors: variations in an individual mouse (intra-mouse variability), variations between the mice (inter-mice variability), and measurement errors. The three sources may be thought to make up the total observed variability. Saliva production on an individual basis is known to be affected by time since last meal, and cyclic changes over a day (circadian dependence) [111]. Different mice might have different baseline production of saliva, growth rates, and radiosensitivity (see section 2.1.2). Thus, one might expect a higher inter-mice variability for saliva values in irradiated mice over time, than control – which is seen as a higher CV for the latest irradiated time-group in Table 4-1. Figure 5-3 attempts to illustrate this point using four hypothetical mice: in the left plot control mice undisturbed by irradiation show cyclic variation across each day (circadian intra-variance), with differing slope for saliva increase over time (growth rates as inter-variance). The total error incorporates both the inter- and intra-variability, with an added buffer to illustrate measurement error. In the right plot different radiosensitivity is illustrated by a dramatic change in slope for mouse 3, with mouse 4 more or less undisturbed (as an exaggerated lower radiosensitivity for this thought experiment) – causing the inter-variance and thus total error to increase more relative the control mice.

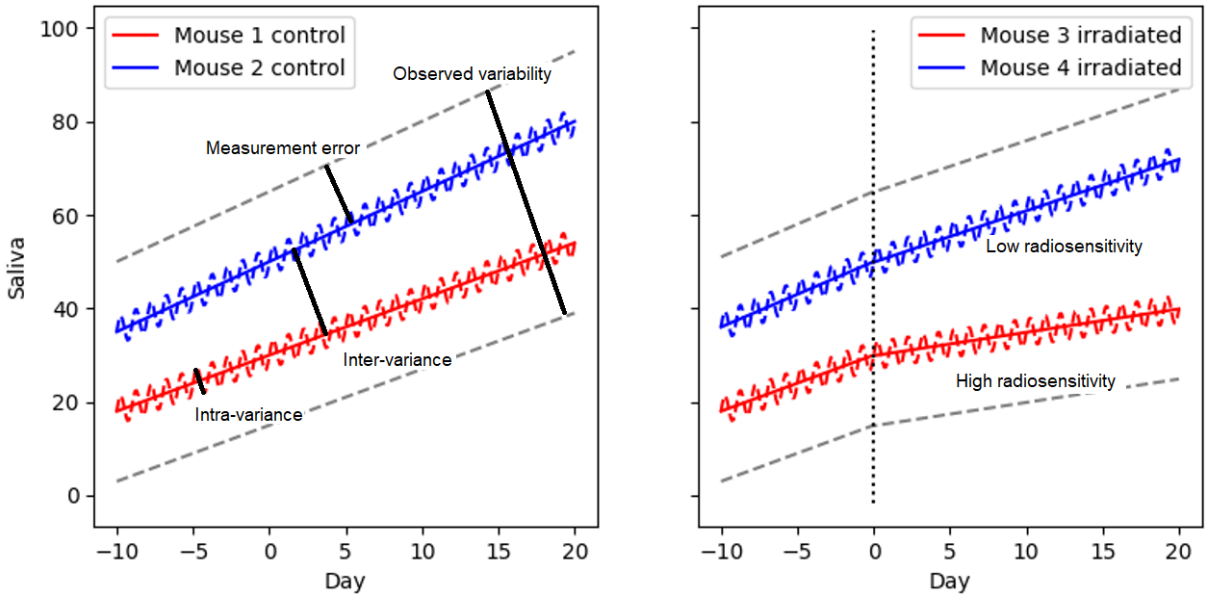


Figure 5-3: Illustration of the three assumed contributors to the high variance (noise) in the measured saliva values. Left: control mice. Right: irradiated mice with different radiosensitivity.

5.3 Segmentation

While state-of-the-art fully automatic segmentation techniques exists, such as the deep convolutional neural net U-Net discussed in [8], a large amount of training data would be needed which was not available. The semi-automatic watershed-based 2-dimensional segmentation algorithm (section 3.2.4) produced seemingly good ROIs overall, but its usage was a time-consuming process as all ROIs had to be manually chosen among the watershed regions. While radiomic studies often utilizes multiple experts for delineation, or works with previously segmented data, in this work the author was responsible for every step of the process. Due to not being an expert in mouse physiology the choice of ROIs representing the SMGs might be assumed affected by human error and observer bias by said author. In an ideal world the segmentation process should have been be repeated and validated by an external observer as discussed in [8].

However: by comparing the ROIs to the area of SG specimen surgically extracted by an actual biologist (PhD student Inga Solgård Juvkam), the validity of the segmented ROIs was checked in section 4.2.1. The SLG size does not necessarily reflect the SMG size, according to the somewhat low correlation of 0.47 between the 20 extracted glands. The SMG size was not significantly correlated to the number of ROI pixels for either the no-p T1 or T2 weighted images, but the SLG was significantly correlated to the T2 images. As the SLG and SMG is fused in mice (section 2.1.3.1) and hard to separate visually in the MR-images, the ROIs might be assumed to contain a bit of both glands. As the sample size for said correlation analysis was

very small (N=9 mice having both tissue samples and late MR images) any further conclusions are not drawn.

Using the ROI sizes between images of various types, before or after pilocarpine (before-p, after-p) and MR-weightings T1 or T2, the variability in image ROIs from the same days was quantified by correlation in section 4.2.2. As the segmentations attempts to capture the exact same anatomic ROI at the same day, the difference in ROI sizes may capture some observer- or method-bias in the segmentation process. While the correlations between before- and after-p images for each weighting was high ($\rho > 0.6$) the correlations between before-p T1 to after-p T2 and after-p T1 to before-p T2 was lower ($\rho = 0.32$ and 0.41, respectively). This may indicate that some partial-volume effect between the modalities more strongly affects the size of the central ROIs (for both the left and right unit), having a larger effect on ROI variability than positional changes before and after pilocarpine injections.

As previously discussed in section 5.1 the control mice are expected to continue growing during the experiments and increases in saliva over time have been established. The expectation for the ROIs is therefore an increase in the ROI sizes for the control mice over time, and potentially shrinkage of the irradiated ROIs as observed in earlier studies (section 2.1.4). The box plot in Figure 4-10 (section 4.2.3) show a lower ROI size for control than irradiated mice in the latest time-group, going against the aforementioned hypothesis. While the difference in means (151 pixels) is significant under an unpaired t-test, the difference is very small compared to the IQRs (which is above 1000 pixels for the irradiated ROIs). The difference in ROI sizes between the left and right SG subunit was also evaluated over time, as the left unit received a higher dose than the right due to being closer to the skin surface where the irradiation beam entered (section 3.1), with no observed differences.

The discordant results in image-type-variability and the temporal evolution of the ROI sizes might be explained by two factors: the MR slice thickness for the images is not negligible compared to the size of the mouse SGs, and the positioning of the mouse might have changed between scans. As such different anatomy from each mouse might be contained in the slices selected for analysis between scans of the mice (partial volume effect). Regarding the temporal analysis of ROI sizes, the assumption that increased saliva production and overall growth implies bigger SGs might be wrong. This is because the glands may exhibit increased efficiency (e.g. as a higher fraction of saliva-producing acini relative adipose tissue) while maintaining their original size.

Qualitatively, the segmentation of the T1 images was harder than the T2 images due to the lower variability in pixel intensities (see Figure 4-12) and the SG edges was therefore not picked as easily by the watershed algorithm.

5.4 Preprocessing and feature extraction

Preprocessing is a major component of the radiomics pipeline with direct impact on the quantitative nature of the features, as discussed in section 2.3. To improve feature reproducibility, and overall repeatability of radiomic studies, all preprocessing steps was reported in detail as emphasized in IBSI [7] and multiple review studies on radiomics [8], [45], [112].

As each MR-acquisition matrix was assumed to contain central slices for the left and right subunit of the SMG, one may ask why 3D radiomics was not utilized instead of having two 2D images per instance with an added challenge on how to deal with this – referred to as LR-modes. 3D radiomics would have utilized a VOI containing the whole SMG organ, i.e. containing more information as a whole than 2D radiomics, but requires interpolation to an isotropic voxel space to make extraction matrices (e.g. the GLCM) rotationally invariant. Interpolation has been shown to increase the reproducibility of features between data sets [8]. While this sounds like an obviously better option than 2D radiomics in theory, the MRI acquisition protocol used in this work provides a challenge regarding interpolation: the large discrepancy between pixel spacing (0.12 mm) and slice thickness (0.70 mm) described in section 3.1. Interpolation may be done by either up-sampling to a higher resolution than the original image matrix, or by down-sampling to a lower resolution. Due to the pixel spacing / slice distance discrepancy up-sampling would imply a higher proportion of artificially generated voxel intensities relative the original data, while down-sampling would imply a significant loss of information. To preserve the raw information as much as possible this was discarded in favour of using 2D radiomics.

A general question regarding the radiomic feature extraction used is why hand-crafted radiomics was used in favour of deep radiomics as mentioned in section 2.3. While deep neural networks are able to find more arbitrary pixel-relationships than the mathematically defined hand-crafted features, a large sample size of data is required for training the deep network for discovery of deep features – which would not be feasible with the small data set utilized in this work. Additionally, the hand-crafted features are more easily explained in a biological context – necessary for potential future clinical implementation as imaging biomarkers.

5.4.1 Post-acquisition processing

As discussed in section 3.3 IBSI does not cover post-acquisition processing, so the choice of what methods to include and their order was largely based on a 2011 study [98]. Three image processing methods were considered in the paper: bias field correction, a landmark-based intensity standardization for normalization, and noise filtering. The paper made no mention of Nyul normalization (section 3.3.2.3) but the landmark-based method is conceptually the same. As suggested, bias field correction (using the N4 algorithm described in section 3.3.1) was used before normalization (section 3.3.2), but no noise filter was applied. By visual inspection of the ROIs noise was determined to generally not be a big issue, such that a choice was made to omit

noise suppression to keep as much raw information as possible by limiting unnecessary preprocessing.

Two methods of intensity normalization were used in the radiomics pipeline (section 3.3.2), Nyul normalization and shifted standardization in addition to no normalization, of which one was decided upon and used on a feature-specific basis (FSPS, section 3.4.1). Using some intensity normalization have been shown to improve the repeatability of features and make MR-images from differing acquisitions more suitable for comparison [113]. The two normalization methods used were recommended specifically for T2-weighted radiomics in a 2020 study [47]. While not applying any intensity normalization is denoted as no normalization in FSPS the images are still normalized in a small manner by the N4 correction (affecting all radiomic features except shape-based) and the discretization (affecting texture-based features).

Looking at the pixel intensity distributions within the ROIs after Nyul normalization in Figure 4-11 (section 4.3.1) the normalization is seen to produce more equal distributions relative no normalization. The intensity distributions after shifted standardization are hard to interpret as they appear as thin lines scattered across the plot, due to the fact that the shift constant is calculated on a per-image basis.

The relative variation of the means cv_{μ} is seen to decrease after both Nyul normalization and shifted standardization, relative no normalization, for both the T1 and T2 images (Figure 4-12). The CV of the means in the ROIs could be interpreted as a descriptor of between-images intensity harmonization within the ROIs, where a lower value means more equal intensity distributions. The average CV is also lowered when applying the shifted standardization compared to no normalization, as well as the N4 correction compared to the raw images.

The T1-weighted MR-images are seen to exhibit lower relative variation than the T2 images as is expected due to the T2 weightings increased contrast for water relative the T1 images (section 2.2.3), given the high water-content in saliva. However, the differences in relative variation between before- and after-p injections are small across the normalizations and follow no obvious pattern. As the saliva, and thus water, content is lower in the SGs when the after-p images were taken (following the saliva extraction for measuring) the CV might reflect the signal from fatty tissue more strongly than water. None of the MR-acquisition protocols utilized fat suppression (section 2.2.3).

While the CV is susceptible to outliers and unstable for distributions centered around zero (as the CV would go towards infinity), this is assumed to not be any issue as very few pixel intensities are below zero after any normalization.

5.4.2 Discretization

As the MR-image intensities are continuous, discretization is required before calculation of the radiomic texture matrices – to avoid arbitrarily large matrices with a single entry for each separate pixel-value in the ROI. The choice between using a fixed bin width (FBW, or absolute

discretization) and fixed bin count (FBC, or relative discretization) is a topic of debate within the field of radiomics. IBSI recommends using FBC for MRI-based radiomics due to the arbitrary nature of pixel intensities [7], but other studies have shown that using a FBW increases the inter-observer reproducibility of the features [105]. Additionally, as the Nyul normalization attempts to establish a physiologic relationship between the images and intensities within the set of images (section 3.3.2.3) the argument for FBC by IBSI is weakened. This assumption is supported by Scalco et al. (2020, [47]) which evaluates the interplay between normalization methods and using either FBC or FBW for feature reproducibility, in combination with T2-w MRI, where using FBW obtained the most reproducible features both considering image information content and with respect to inter-observer variability.

The choice to use separate discretization bin widths between normalization methods, described in section 3.3.4, is trivial as the intensities after normalization are on completely different scales than the raw values – which requires adaptation by the FBW procedure. As the Nyul normalization utilizes separately trained standard scales between the T1 and T2 images, it is a logical continuation to split the optimal bin width between T1 and T2 within each normalization. However, the before- and after-p images are not considered separately as this is exactly the type of physiological changes (i.e., the change in saliva content) assumed to be picked up by the features when working with radiomics.

5.4.3 Feature-specific preprocessing selection

In the feature-specific preprocessing selection (FSPS, section 3.4.1) the intensity normalization is decided upon on a per-feature basis: either no normalization, shifted standardization, or Nyul normalization. While the paper by Fave et al. (2016, [44]), inspiring the FSPS, considered different preprocessing steps for selection (smoothing filter and resampling) the method have no reason not to work for choosing between intensity normalizations instead.

The selection is based on two subsequent criteria of which the first incorporates a univariate filtering by discarding features which are not significantly correlated to the saliva production for any of the three normalization modes. The FSPS procedure therefore effectively becomes the first step in the feature selection process, as a filtering method (section 2.3.5), reducing the size of the feature-spaces available for later selection and modelling. FSPS is considered separately for the T1- and T2-weighted images, as well as the LR-modes.

The second criterion, relevant only if multiple normalizations pass the first criterion for a given feature, selects the normalization which obtains the lowest correlation with ROI area. As ROI area is a radiomic feature easy to interpret (assuming the ROIs actually corresponds to the SMGs / SLGs) one may describe the criterion as maximizing relevant information to the outcome while minimizing the overlap between shape-based information. The shape-based features are unaffected by normalization and being only 9 features they are all excluded from FSPS, and simply included in the post-FSPS feature-spaces.

Due to the arbitrary intensity units in MR-images (section 2.2) one might expect that applying some normalization is preferrable for most features, yielding higher-performing features compared to none. Especially is Nyul normalization expected to work better with FBW discretization as discussed in section 5.4.2. Looking at the percentage of features selected with each normalization seen in Table 4-5 (section 4.3.3) the Nyul normalization is seen to be selected the most across both LR-average and LR-aggregated features (43% - 72%), except for T1 LR-average where shifted standardization is selected for 73% of the features, partly confirming this expectation.

The T2-features is seen to have a much higher chance of surviving FSPS compared to the T1-features, even with a more strict selection threshold ($p = 0.05$ for T2 and 0.15 for T1). It should be noted that the baseline sample size, used for FSPS, is more than twice the size for T2-images (N = 55) compared to T1 (N = 24). LR-average features are also seen to have lower feature-space size after FSPS than LR-aggregated features.

The first iteration on how to deal with image features from both the left and right subunit of the SMG, was to simply treat features from the two units as separate observations with respect to the measured saliva production – referred to as LR-split (since images from the same instance are split into two observations). The first iteration of delta-p and delta-features were therefore calculated from the pool of surviving LR-split features following FSPS. The LR-split method was quickly deprecated, as observations from the same mouse and day would be treated independently and as such easily end up in both train and test subsets of the data (increasing the chance of overfitting the models). FSPS was then performed for LR-aggregated and LR-average features, which treats each observation separately. However, the FSPS was not re-done for the LR-average delta-p and delta-features such that the LR-split iteration of the FSPS became the initial selection for those feature-spaces. This human error might have caused important features to be discarded, while keeping less important, for the LR-average delta-p and delta-features. For all other combinations of feature-spaces and LR-modes the re-run FSPS results was used correctly.

5.5 Selection and modelling saliva as outcome

Assuming the high variance in the measured saliva production to be related to intra-mouse and inter-mice variability (as discussed in section 5.2), more than noise due to a high measurement error, an optimistic viewpoint on radiomics may assume that the image features should reflect the physiological state of the SGs. Thus, the features should be expected to contain information related to the measured saliva values, and as such be suitable for prediction-based machine learning models.

The high water-content in saliva (section 2.1.3) is expected to reflect the signal more strongly in the T2-images over the T1-images (section 2.2.3). As the original first-order radiomic features

are closely related to said signal (section 2.3.1), one might expect first-order features to be good predictors of simultaneously measured saliva – especially using the T2-weighted delta-p features describing the relative change in each feature before and after pilocarpine injections for measuring saliva (section 3.5.1).

Random forest (RF) models are used both for classification tasks relating to the binary xerostomia outcome, and for regression to the continuous saliva production outcome. RF have shown previous success within radiomic studies ([74]) and may capture complex non-linear relationships between the features due to the non-parametric nature of decision trees in combination with the probabilistic nature of the ensemble forest made up of said decision trees using bootstrapped data (section 2.4.3.3). In addition to RF-models multiple linear regression is used for the regression tasks, and logistic regression for estimating class probabilities in the binary classification tasks. As they are parametric methods, the influence from features onto the model's structure is more easily explained relative the non-parametric RF models - as both linear and logistic regression estimates a coefficient for each feature used as an explanatory variable in the model (sections 2.4.4 and 2.4.5).

All models using radiomic features are evaluated in their relation to the simplified NTCP-model using only time and dose as explanatory variables, denoted td-models. As the subset of data available between analyses differs the td-models are seen in relation to the td-analysis using all 347 available saliva measurements, which may provide information of the optimism of the model performances based on random variations in the subset of data used for each analysis.

5.5.1 Predicting simultaneously measured saliva and xerostomia using only time and dose

In sections 4.4.1 and 4.5.1 all available saliva measurements (N = 347) were used to predict the continuous saliva production measurements and the binary xerostomia after thresholding, independently of the radiomic features. Both the regression and classification tasks were based on the same data, and the machine-learning models used the exact same explanatory variables: day of measurement relative start of irradiation (day 0), and the dose delivered at said day. This attempted to establish a simplified NTCP-model (see section 2.3.6), without individual-specific data, in order to establish the predictivity of time and dose as explanatory variables in a machine learning context (referred to as a td-model). Only subsets of the saliva measurement data were utilized when creating models using radiomic features, where td-models were used for comparison. As such, the td-only modelling gave the most accurate picture of the predictive abilities of time and dose for both regression and classification.

In the td-models for both regression and classification, time is seen to be a less stable predictor than dose: the estimated coefficients vary more for time than dose in the regression models and are non-significant in the classification models. This corresponds to the non-significant correlation between time and saliva measurements from irradiated individuals as discussed in section 5.2. The time-values take on a larger span of values than the dose-values, where the latter

is either zero or the total delivered dose after irradiation (e.g. 44 Gy) as no measurements were taken during the fractionated irradiation days – and is as such a binary explanatory variable on a per-individual basis. In the logistic classification model with second-order interactions between time and dose however, the interactions term is significant when fit to all data.

The RF-models are seen to perform better (higher R2) than the linear regressor models when evaluated on the training data in regression tasks, but scores worse (lower R2 and higher MSE) than linear regression when averaged across the 5-fold test data and in the LOOCV. Interestingly, when comparing RF to logistic regression for binary xerostomia classification the RF-models performs better both when trained and evaluated on all data (higher accuracy, AUC, and lower BS) *and* on unseen data in the LOOCV (higher AUC). The classification performance on unseen data is however unchanged when evaluating the models probabilistically, i.e. unchanged BS, meaning the RF-models might have overfit to noise in the data (low bias but high variance reflected when predicting unseen data). However, if some estimated class probabilities are way off the binary ground truth, such outliers might affect the BS more than the AUC.

As the logistic regression model for xerostomia classification with interactions performs better across the LOOCV than without interactions, some nonlinear effects between time and dose seems to be captured. Seen in combination with the assumed overfit by the RF-models, which could capture more non-linear effects due to its non-parametric modelling nature, the biggest explanatory impact on saliva production from time and dose may be assumed to be mostly independent of each other.

Even when having about 70 observations in each test set across the 5-fold cross-validation (CV) evaluating regression performance in Table 4-6 the test scores vary for both regressors across the CV. This implies a heavy split-dependence on the td-model's ability to generalize to unseen data. The hypothesized intra- and inter-mice variability as discussed in section 5.2, would not be captured by the td-models as time and dose are non-specific features between individuals.

The RF hyperparameter (HP) tuning for the td-models seems to have little impact on the model's performance, as illustrated by Figure 4-16 evaluating the averaged R2 across a 5-repeated 2-fold CV against the number of decision trees (estimators) in the RF. There is no observed pattern across the 5-fold CV, with peak performance (highest R2) at some number of estimators varying between the folds. This trend is observed across all HPs for the RF regressor, except certain HPs preferring lower values such as the minimum samples considered per new created split in the decision trees (section 2.4.3). While the 5-repeated 2-fold CV on the training data attempts to achieve higher generalization for the HP optimum, the optimum for the test data (unseen in the HP tuning) is not necessarily achieved. As illustrated in Figure 5-4, a much higher HP value would achieve better results on some of the unseen test data (upper plot) but would be chosen at a very low value in the hp-tuning using training data (lower plot).

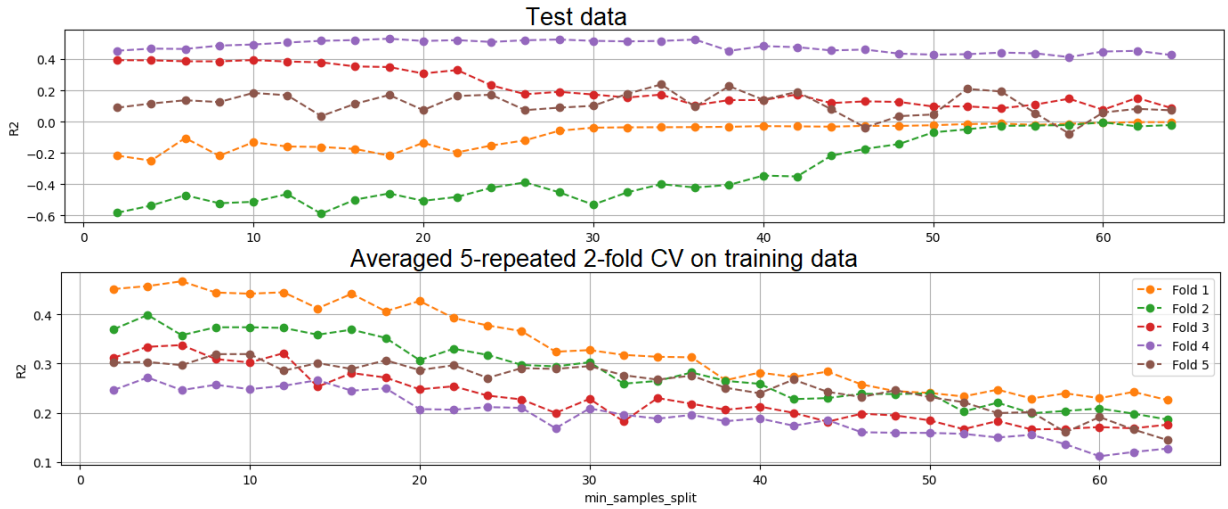


Figure 5-4: Illustration of the discrepancy between optimal HP from tuning (high point bottom), with the corresponding effect on test data (upper).

5.5.2 Predicting simultaneous saliva and xerostomia using only radiomic features

A natural starting point for establishing any potential relationships between the saliva measurements and the radiomic features in a machine learning context, is to use the image features to predict saliva values from the same time-points (days) as the images were taken. Regression models (section 4.4.2) were used to predict the continuous saliva amount as outcome, and classification models (section 4.5.2) were used to predict the class probability for binary grouped xerostomia following the thresholding method (section 4.1.2).

Three feature-spaces were considered for the analysis: using radiomic features extracted from T1- or T2-weighted MR-images before pilocarpine extraction for saliva measurement (no-p T1 or T2) and the delta-p features calculated as the relative difference between each T2-feature from before and after pilocarpine injections (section 3.5.1). The number of available observations, i.e. the overlap in time and mouse-id between the image features and the saliva measurements, was 140 for the no-p T2 features and 69 for no-p T1 and delta-p. The percentage of observations being from control mice were between 41% and 48% (Table 4-8) for both regression and classification analysis. For the classification analysis the percentages of mice having saliva measurements determined to be xerostomic was between 38% and 45% (Table 4-11).

The same split into training and test data was used for both regression and classification, separate for each feature-space with as much overlap as possible (section 3.6.1). In the regression tasks, only using time and dose as predictors (td-models) performed better on the test set than the image features which had all R2-values below zero (Figure 4-17). Some feature-based linear regression models achieved a perfect fit to the training data (R2 = 1.00), not unexpected when

using all available features without any regularization parameter penalizing multiple predictors, but had the worst results on the test data (as the models are extremely overfit to the training data). The RF-models utilizing the whole feature-spaces as predictors, were less overfit compared to the similar linear regression models, but the inherent RF feature selection (by importance, see section 2.4.3.3) performed worse than primary selection by MRMR.

In the classification tasks (Figure 4-20) the td-models also performed better than the feature-models, except for the no-p T1 model using the top 5 LR-average features with a RF regressor having AUC = 0.78. The no-p T2 and delta-p models had AUCs below, or barely above, 0.50 – making the models slightly better than random guesses. Between both regression and classification the td-models using the no-p T1 subset is seen to perform worse than the T2 and delta-p based td-models in the single split.

In order to mitigate potential imbalance in the randomly split train and test sets, LOOCV were performed in addition to the single split for both regression and classification. In the regression tasks only two feature-based models had R2-scores above zero (Figure 4-18): a no-p T2 LR-aggregated model using linear regression (R2 = 0.09) and delta-p LR-average using random forest (0.06). The no-p T1 models performed the worst with a maximum score of R2 = 0.00. In the classification tasks (Figure 4-21) the no-p T1 LR-aggregated feature models were seen to have a maximum AUC at 0.65. The no-p T2 models had some models with AUC barely above 0.50, and the delta-p models were all below. No pattern is seen between the optimal td-models from the regression and classification tasks, except some small improvements when using a RF-based model compared to linear or logistic regression.

Looking at the calibration curves (Figure 5-5) for the three feature-spaces predicting simultaneous xerostomia, evaluated by bootstrapped LOOCV, no-p T1 is seen to stay closest to the central line (red stippled line) with a sigmoidal shape while delta-p strays the farthest. This reflects the results in Figure 4-21, as the reliability curves describes how well the predicted probabilities relates to class prevalence in the ground truth. In general, the RF classifiers tends to create some prediction bias for estimated probabilities away from 0 or 1. All decision trees in the forest would have to agree on 0 or 1 as the estimated outcome in such a case, which is unlikely when building each tree using bootstrapped data in the bagging process [114]. This trend is more visible in the no-p T2 and delta-p distributions of estimated probabilities, which is seen to go towards zero as the probability approaches 1 - leading to a large deviation from the central axis in the right part of the calibration diagram.

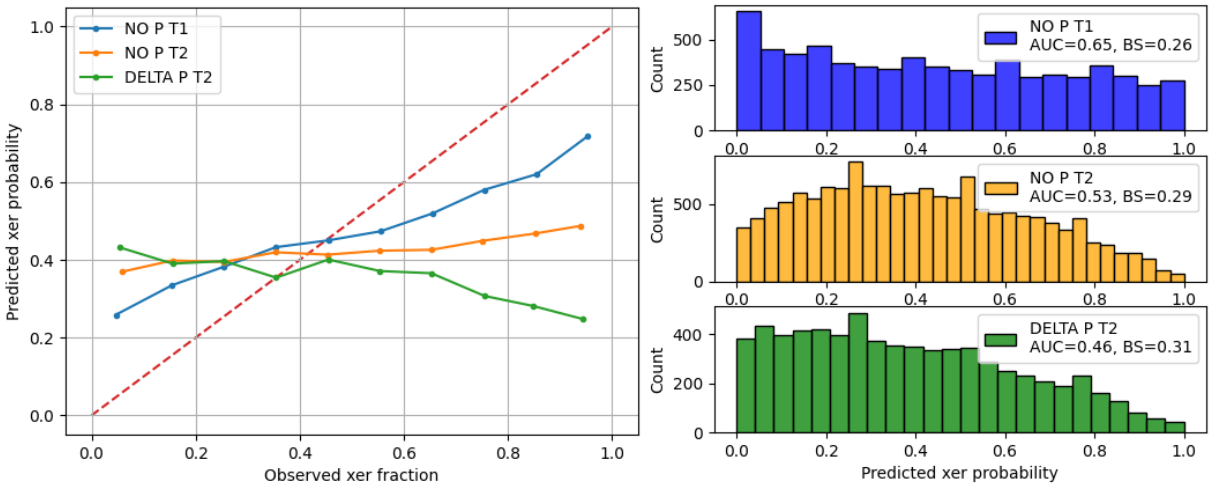


Figure 5-5: Left: calibration curves for three random forest models using LR-aggregated features to predict simultaneous xerostomia, evaluated across a LOOCV with 100 bootstrapped training sets. Right: distributions of estimated class probabilities, across the 100 bootstraps and LOOCVs.

Across all the regression and classification models the td-models generally performed better than the feature-based models, with some exceptions (no-p T1 in the single split classification, and delta-p in the LOOCV regression). All feature-based models performed better in a LOOCV than the single train / test split, except for some higher AUCs when using no-p T1 features for classification. Generally one might expect better model scores from a LOOCV compared to a single split as almost all $(N - 1)$ observations are used for training, creating models with higher bias (section 2.4.1). This effect might have been strengthened by the small sample sizes for the available data.

The regression td-models had some R2 scores on the single split test data which were unrealistically high compared to the regression models utilizing all salivation data (discussed in section 5.5.1). Assuming an optimistic viewpoint on radiomics in general one may assume that the radiomic features captures some of the intra-mouse and inter-mice variance hypothesized in section 5.2, while the td-models does not (as no information in the time of sampling or dose given is individual-specific). Under this assumption might the td-models be more susceptible to such variations and potential outliers in a single train / test split, scoring models much higher or lower than the td-models utilizing all saliva measurements. The td-models evaluated by LOOCV have R2's more closely resembling the td-models using all saliva measurements, supporting this hypothesis as they are unaffected by random variations in a train/ test split by using all data for testing once.

Using delta-p features had higher performance on regression tasks than classification. One may hypothesize that the delta-p features manage to capture more of the intra-variability in the mice, due to being the relative difference in features from before and after saliva extraction. As some of this variability information may be lost when reducing the outcome to a binary xerostomia variable, the decreased performance on the classification tasks might be reasonable.

5.5.3 Predicting late xerostomia using radiomic features

Predicting xerostomia forward in time using the radiomic features is a more interesting problem from a clinical perspective, as successful features potentially could help decision-making for preventive measures before or during RT of HNC. The majority of radiomic studies therefore considers predictions to endpoints forward in time relative the image acquisitions. Classification tasks were prioritized over regression for this analysis, to allow for comparison with a previous study on radiomics and xerostomia [10]. This section considers the results in section 4.5.3.

Based on the time of image acquisition, the features are divided into three temporal groups: using baseline features, features from after irradiation (after-irr), or as a combination of the two (delta-features, section 3.5). For this analysis each individual corresponds to a single observation in each data set, compared to some individuals having multiple observations in the analysis of simultaneously measured saliva.

Some intuitive expectations regarding the temporal feature groups are an increased performance for after-irr features relative baseline, having information closer in time to the outcome for prediction, and that the delta-features performs the best due to effectively incorporating temporally dependent information from two observations. The prediction of xerostomia forward in time is expected to a more difficult ML-task than simultaneous predictions. Additionally, the sample sizes are smaller while the number of features is unchanged (e.g. 42 observations for no-p T2 after-irr, compared to the 140 observations for simultaneous predictions using no-p T2 features).

A potential major error source from the implementation of this analysis, is the large temporal variation in, and overlap between, the after-irr and delta-features and the saliva measurements as discussed in section 3.5. While there is no overlap on a per-individual basis, some individuals have after-irr features from images taken at day 35 while others have the “late” saliva measurements from the same day. While this is obviously not optimal, allowing this overlap was necessary to have sample sizes big enough for any analysis at all.

Similarly to the analysis of simultaneous predictions, was the classification tasks evaluated by both a single split into training and test along a LOOCV. Looking at the results from the single split in Figure 4-22 and Figure 4-23, the after-irr features is seen to perform better than the baseline features as expected. The delta-features, however, is seen to be the overall worst predictor with no models obtaining AUC above 0.50. Across the feature-spaces using after-irr features, are some models using no-p T1 and delta-p features seen to obtain AUCs above 0.80. The no-p T2 features maxes out at AUC = 0.60, being lower than the exact same model (LR-aggregated RF) using baseline features with AUC = 0.63. Some variation is seen between the AUCs for the td-models, which overall performs well when using after-irr values for time and dose. The td-model on the no-p T1 data performs almost perfectly with AUC = 0.95.

Concerning the LOOCV evaluated models, seen in Figure 4-24 and Figure 4-25, the after-irr features are seen to overall perform better than the baseline features. An obvious exception

standing out among the baseline models, is the no-p T1 models using LR-average features having AUC = 0.80. Using LR-average delta-features with a logistic regression model is seen to be a good model with AUC = 0.80, higher than the maximum AUC attained between after-irr feature models at 0.75 using no-p T2 LR-aggregated features with a RF classifier. Using after-irr no-p T1 features have some lowered performance in the LOOCV with max AUC = 0.72, and the delta-p features perform equally bad compared to the single split. All baseline td-models are bad predictors with AUC < 0.50, where the delta-p td-models are perfectly wrong for all predictions (AUC = 0.00). This is expected, as all dose values are zero while the only difference in the time variable is being either day -3 or day -7 – i.e. no information relating to xerostomia. However, one might have expected the AUCs for such models to approach 0.50 rather than 0.00, indicating the models attempts to use some relationship between xerostomia prevalence between the two baseline days (unsuccessfully). Similarly, the baseline td-model having AUC = 0.67 in the single split must therefore be due to a random imbalance in xerostomia between the baseline times in the test set, which does not affect the LOOCV.

An improvement in both delta-feature and no-p T2 based models are seen in the LOOCV relative the single split. In combination with the observed decreased performance for the no-p T1 features, one may assume the LOOCV results to be more realistic.

5.5.4 Comparing T1- and T2-based feature models on the same subset of data

To mitigate the potentially different random imbalance in variability for the data used in T1- and T2-based models, due to predicting different subsets of the total outcome space, a separate analysis was performed using features from the two MRI-weightings for predicting the exact same outcome. In practice, this meant discarding observations from the T2 data until the T1 and T2 were left with the same individuals and times. The models were scored on simultaneous predictions, or late predictions using baseline or after-irr features. A 3-fold train test split was performed in addition to a LOOCV. In addition to creating models using only features from each sequence, a combined model was created by aggregating the no-p T1 and T2 feature-spaces column-wise allowing for selection of features from either MRI-weight.

While the td-models performed the best across the 3-fold split (Figure 4-26), the baseline td-model performed the worst in the LOOCV (Figure 4-28). Looking at the distributions of AUC-values across the bootstrapped 3-fold split (Figure 4-27) the td-model is seen to have a peak at AUC = 0.50, which the T1 model does not. The T1-model was therefore a more stable performer on baseline features, in addition to being the best baseline model in the LOOCV. However, the T1-model was not significantly better (lower BS) than the T2-model using baseline data, when comparing the squared probability errors across the LOOCV, and the AUC was not particularly high (AUC = 0.57).

The T1-models outperformed the T2-models for prediction of simultaneous xerostomia for both CV methods and had significantly higher average AUC across the 3-fold split ($p < .000$) and lower BS across the LOOCV ($p = 0.039$). The combined model performed similarly to the T1 model, with a slight increase (3-fold split) or decrease (LOOCV) in AUC.

The combined model performed significantly better (higher AUC with $p < .000$) than both the T1- and T2-model for late prediction using after-irr features in the 3-fold split. In the LOOCV the T2- and combined models performed better than the T1 model, but the lower BS was not significantly different for any combination.

As the paired t-test in the 3-fold split included all AUC-scores from bootstrapped validation 1000 times, effectively creating 1000 "pseudo-observations" for each real observation, the estimated p-values became extremely low. Thus, the t-test between BS in the LOOCV only considering each observation once, produced more sober results.

5.5.5 Comparing the added predictive ability of time and dose with best radiomic features

While models using only time and dose as predictors generally outperformed the models only using radiomic features discussed in the previous sections, the radiomic features may have contained additional information on a more individual basis than time and dose. Thus, using both time, dose, and some relevant radiomic feature was assumed to increase the model performance compared to only using time and dose. Based on the LOOCVs for both regression and classification, a feature was considered for this analysis if it were chosen more than 50% across the left-out observations. In another words: all data have been used for selection of the presumably top-performing features, across all feature-spaces and LR-modes, making this effectively embedded feature selection method (section 2.3.5) a potential source of data leakage.

For each LOOCV-feature a combined model using time and dose, in addition to the feature, was created for univariate evaluation. T-tests between the squared error for the combined model and the td-only model, across a LOOCV, was performed to test whether the combined model had significantly lower MSE (when regressing to continuous saliva values) or BS (when classifying xerostomia). The BS is effectively the probabilistic version of the MSE given a binary outcome, making the t-tests performed for both regression and classification comparable. Due to the grand number of tests performed, the tests were also evaluated under Bonferroni corrected significance thresholds in addition to the standard 0.05.

While no models had significantly improved performance under the Bonferroni corrections, 15 models using 14 separate features were significant with $p < 0.05$. Of the 14 features 4 were from regression to simultaneous saliva, 2 from classification of simultaneous xerostomia, and 8 from prediction of late xerostomia. Of the 8 features being LR-aggregated, 7 were the version from the right SG subunit.

As seen in Figure 5-6 the 9 texture-based individual features had the highest prevalence, followed by three first order features and two shape-based. Excluding the shape features (which are denoted as original by the pyRadiomics python package), the gradient filter was the most prominent with 3 features, followed by 2 features without any filter, logarithm filtered, or wavelet filtered (one H and one L), each. The indexing between feature number and filter and type is seen in Appendix C.

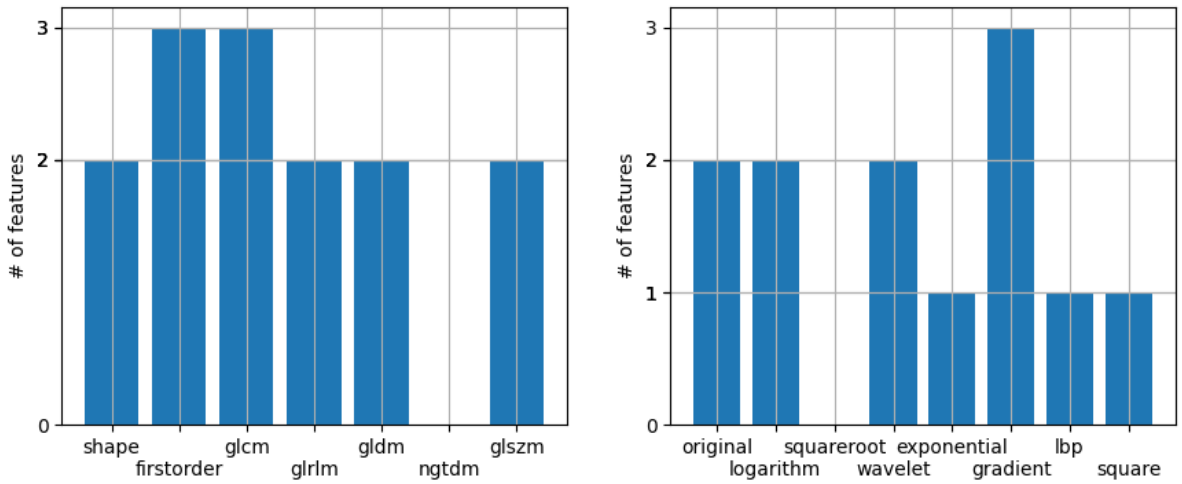


Figure 5-6: Histograms of feature types (left) and image filters (right) among the 14 features significantly improving upon the td-models. Shape-based features were excluded before counting filter types.

Concerning the regression features, an interesting observation is that a lower MSE for the combined model relative the td-model does not mean a significant difference in the squared errors: while adding Ft192 to the td-model (using no-p T1 LR-aggregated data with a RF regressor) lowered the MSE from 1970 to 1570, the combined model was not significant ($p = 0.110$) and may therefore not be considered a stable predictor. On a similar note: low-performing models using the top 5 MRMR selected features across the LOOCV, such as delta-p LR-aggregated with a linear regressor (see Table 4-9) having originally a MSE of 3707, contrarily shows to have stable features which significantly improves the performance. Using Ft297R from the aforementioned delta-p model in addition to time and dose significantly reduced the MSE by 15% relative the td-model (from 2888 to 2452). Oppositely, models performing well in the MRMR-selected LOOCV does not imply any significant features here: the delta-p LR-average RF model had the second highest R2 in Figure 4-18 (among the feature-based models, with R2 = 0.06) but none significantly univariate features in the combined models. The top-performing feature-model in Figure 4-18 however, no-p T2 LR-aggregated with linear regression, had two univariate significant features in the combined models: Ft522R and ft240L, reducing the MSE with 9% and 14% respectively.

Much more features created combined models with significant improvements from the td-models in the classification task, relative the regression tasks. Even more features reported significantly lowered BS, the change was negligible and thus a second criterion demanding a minimum 1% relative negative difference in BS between the combined and td-models. This issue was not present when evaluating the regression models, which only had the four significant features overall. One may argue that this difference substantiates the binary grouping by xerostomia thresholding, and that the classification ML-task was more feasible when having such a noisy outcome than regression.

Regarding prediction of simultaneous xerostomia the no-p T1 LR-aggregated feature 635R with logistic regression improved the td-model the most, increasing the AUC (from 0.686 to 0.731) and lowering the BS (by 2.9%) as seen in Table 4-14. This reflects the results from LOOCV evaluated simultaneous xerostomia prediction (Figure 4-21) where no-p T1 LR-aggregated features performed the best among the feature-only models.

Among the features improving late predictions (Table 4-15), two features are from baseline. While the no-p T1 LR-average models scored a high AUC using only baseline features (AUC = 0.80, Figure 4-24) only one of the 5 features improved upon the td-model significantly (Ft498, see Table 4-15). However, while the combined model decreased the BS (by 1.7%), the high performance observed using only image features is not present with a low AUC (0.300). Similarly, Ft2 (from delta-p LR-avg baseline) increases the AUC from 0.000 to 0.119 and is therefore still a very bad model.

Among the four after-irr features improving upon the BS, the delta-p feature (Ft678R) did not improve the AUC. The other three features were all no-p T2 features in combinations with a RF classifier and had a bigger impact on both AUC and BS resulting in very good models (all AUCs above 0.870). Especially Ft216R is seen to be improve upon the td-model massively, resulting in a combined model being the highest performer across all models in this work (with AUC = 0.943 and BS = 0.098, Figure 5-7). However, the td-model is seen to already have a high AUC (at 0.795) above the td-model utilizing all data in section 4.5.1 (having maximum AUC = 0.67, see Table 4-10), and might therefore be assumed to produce somewhat optimistic results. Still, the improvement in BS is large and such a high-performing feature is a great candidate for further investigation.

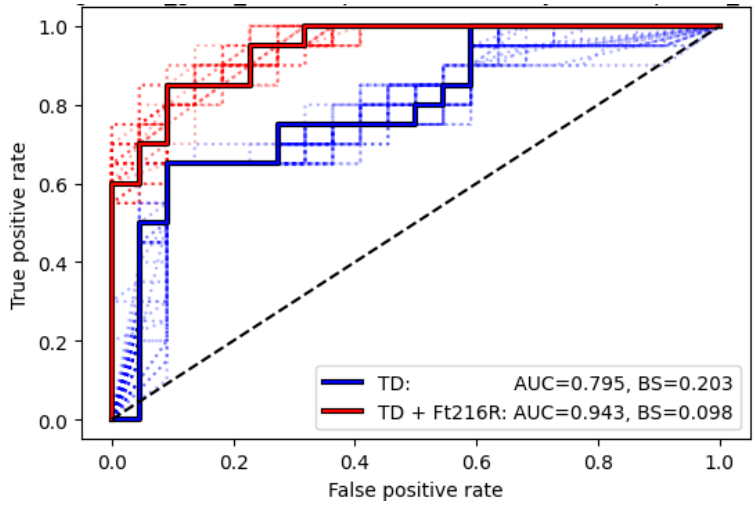


Figure 5-7: ROC curve for highest performing feature (216R) in a combined model with time and dose, using no-p T2 LR-aggregated data with a random forest classifier. The classifier was fitted 100 times, seen as dotted curves surrounding the solid curve created using the averaged probability estimates.

Among the three delta-features identified, two models (RF and logistic regression) both used the same shape-based LR-averaged feature (Ft2) to achieve high-performing combined models. Having a shape-based delta-feature as a good predictor of late xerostomia corresponds with the study by van Dijk et al. (2018, [10]), which identified the relative (delta) surface change between baseline and week 3 to be the strongest univariate predictor (as discussed in section 2.3.6).

5.6 Evaluation of top performing features with biological interpretations

The Quantitative Image Biomarker Association (QIBA) attempts to standardize the clinical implementation of imaging biomarkers, such as radiomic features, using a five-step process [115]. The first step relates to a public claim regarding some biomarker identification with an explanation of the clinical relevance.

As this work is a part of a pre-clinical study evaluating short- and long-term effects after RT to the HN region (PROCCA), biological interpretations of the top performing radiomic features are of interest as they might contain information regarding the biological mechanism post-irradiation on a macroscopic level.

5.6.1 Shape-based features

The shape-based radiomic features might be expected to capture macroscopic changes in the SGs, unrelated to pixel intensities, such as loss of mass or inflammation-induced swelling (see

section 2.1.4). Inter-mice variabilities such as growth rate or radiosensitivity might have affected the shape features, but one may assume intra-mice variability over shorter periods (circadian dependence) to be unrelated. Seen in relation to the questionable validity of segmentations performed in this work discussed in section 5.3, any claims regarding the shape features should be done with caution.

The delta-p LR-averaged feature Ft2, having improved upon the td-model as described in section 5.5.5, is a measure of the major axis length in the 2-dimensional ROI (the maximum length across the ROI). Being a delta-p feature, the feature describes some ROI shape change before and after saliva extractions by pilocarpine injections on the same day. The relative delta-p change is both positive and negative at baseline for both control and irradiated individuals. While the baseline feature improved upon the td-model the predictions were still bad in the combined model with AUC below 0.50 and BS equal to 0.25, as well as when being used as a univariate predictor. The feature might therefore be significant only due to random variations in the segmentations, or positional changes before and after pilocarpine injections, rather than a change in shape on the same day induced by extraction of saliva from the glands.

The relative difference over time of the average elongation between the left and right segmented units (delta LR-average Ft1) improved upon the td-model both using a RF and logistic classifier. The elongation was calculated as the squared ratio between the lengths of the major and minor axis following IBSI [49]. From an optimistic standpoint, one may assume the averaging between two segmented ROIs from the same individual to cancel out segmentation errors – thus improving the feature accuracy. Looking at the scatter plot of the delta-feature in Figure 5-8 the xerostomic individuals seems to have a larger increase in elongation between baseline and after irradiation, when evaluating the control and irradiated individuals separately. Being a shape-based delta-feature a natural comparison is to the delta-radiomics analysis by van Dijk (2018, [10]), where the relative change in PG surface between baseline and week 3 was the best predictor of late xerostomia. As the elongation feature have a low insignificant correlation to ROI surface ($\rho = 0.09$, $p = 0.27$) the delta-version of Ft1 may be assumed to either describe some different biological phenomenon than the previously discovered shrinkage in PG-surface, or be due to variations in mouse positionings during the MRI-acquisitions across the experiments – changing the shape of the organ due to internal pressures affecting elongation with a random relationship to saliva changes in the small sample size for delta-analysis (N=39).

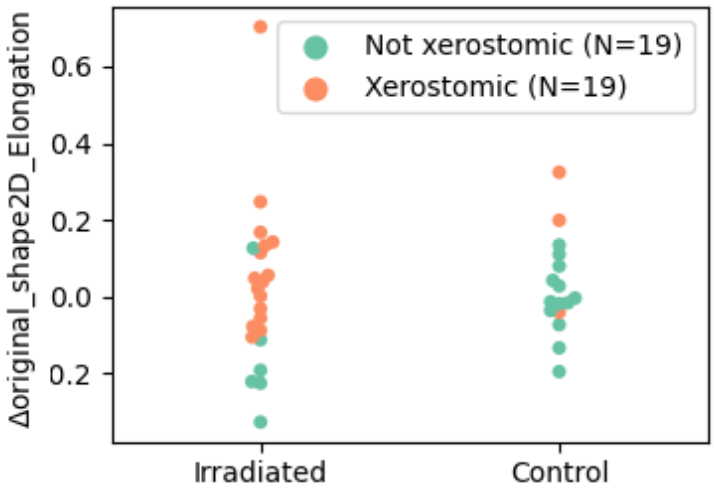


Figure 5-8: Relative change in ROI elongation over time (delta-features), for irradiated and control individuals.

5.6.2 First-order features

As discussed in section 5.5, the signal in T2-w MR-images is expected to be more directly related the water content in saliva relative the T1-w images. Among the three top-performing first-order features two are T2-weighted delta-p features, and all three are from predictions to saliva measurements from the same day as the images (simultaneous). Being the relative difference in feature values before and after saliva measurements, the delta-p features may be expected to more strongly reflect both intra- and inter-variabilities in the saliva productions for the mice than the before-p (no-p) T2 features. An earlier study found the maximum intensity in CT-images to improve predictions of late xerostomia [62].

The combined model using the LR-average delta-p feature Ft12 had the biggest relative reduction in MSE compared to the td-only regression model in section 4.4.3. Ft12 is a measure of the image *energy*, without any image filtering, calculated as the sum of all squared intensities within the ROI. As the squaring amplifies high-valued intensities, the feature becomes a measure of the magnitude of the intensities. Most delta-p Ft12 values are negative, indicating a lower magnitude among the pixel intensities overall after saliva measurements. The delta-p feature has a Pearson correlation of 0.329 ($p = 0.006$) to the saliva values, increasing to 0.579 ($p < .000$) when only considering irradiated individuals while being very low and insignificant when only considering control individuals (correlation at 0.031 with $p = 0.874$). This may indicate the first-order delta-p feature captures variability factors such as radiosensitivity, affecting the SG destruction and complementary loss of saliva production, more than factors relating to control individuals such as variations in growth rate.

5.6.3 Texture-based features

The texture-based radiomic features might be expected to capture changes in the microscopic environment, visible as macroscopic relationships between neighbouring pixels in the images. Earlier studies have shown an increase in heterogeneity in the salivary glands post-irradiation, and related texture-features to observed changes in the epithelial architecture, increase in adipose tissue, or a decrease in vascularization as discussed in sections 2.1.4 and 2.3.6. As the abovementioned biological changes may be highly individual, the texture-features might be expected to capture inter-mice variations.

Among the top performing texture features, features based on the gray level co-occurrence matrix (GLCM) was represented the most (Figure 5-6). Of the three GLCM-features, two were calculated after applying a gradient filter. The relationship between the GLCM and the gradient filter is illustrated in Figure 5-9. As the gradient filter enhances edges in the image (sections 2.3.4 and 3.2.2) the branching structure central in the ROI becomes more visible after applying the filter. Comparing the structure to the mouse anatomy in Figure 2-5, it may be either a salivary duct (for saliva transportation) or vascular such as the carotid artery.

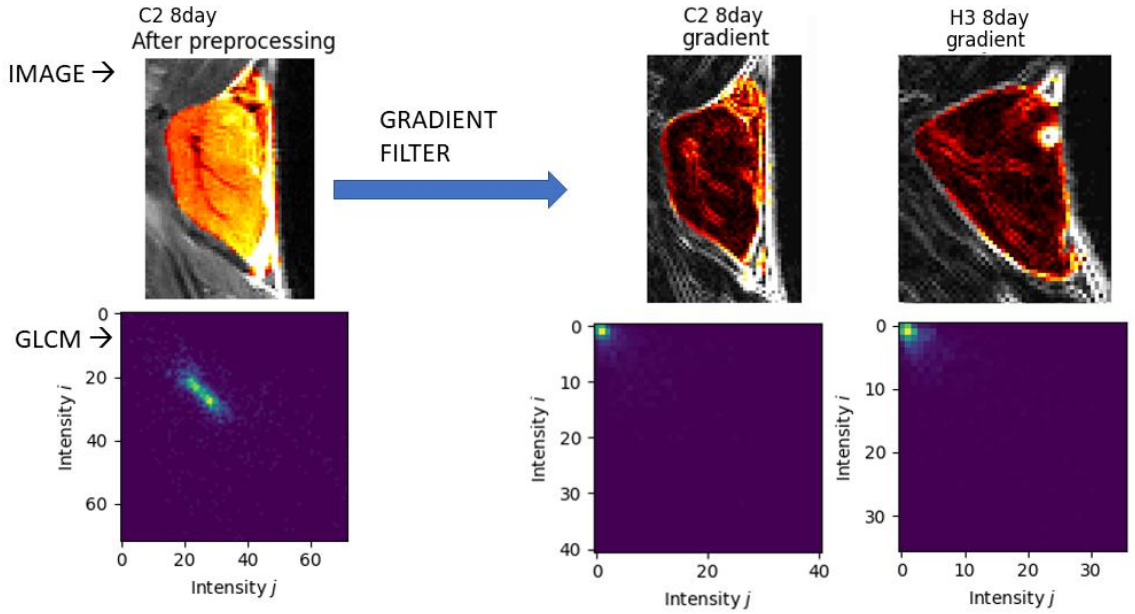


Figure 5-9: Illustrating the relationship between the gradient image filter and the gray level co-occurrence matrix (GLCM). Right: gradient filtered image and GLCM for control (C2) and irradiated individual (H3) after irradiation (day 8).

As mentioned in section 2.3.6 the GLCM correlation was shown to be a significant predictor of xerostomia in a 2018 texture-analysis study using CT images [65]. As a delta-feature the right version of the LR-aggregated GLCM correlation after applying the gradient filter (Ft759R) was

among the top predictors for improving predictions of late xerostomia relative the td-model. Looking at the temporal evolution of the feature in Figure 5-10 (left plot), a separation between control and irradiated individuals is present post-irradiation (from day 8) with an increased difference in the latest time-group. The GLCM correlation describes the linear dependency between discretized gray level values in the ROI and their respective entries in the GLCM [49].

For prediction of late xerostomia, the no-p T2 LR-average feature 765 was among the top models in combination with time and dose. Using the GLCM from after gradient filtering, Ft765 describes the informational measure of correlation (IMC) relating the joint distributions between i and j in the GLCM using mutual information [49]. IMC is a descriptor of the texture complexity which, in combination with the gradient filter, may be assumed to be related to the visible branching structure. Ft765 is seen to separate the controls and irradiated individuals post-irradiation in Figure 5-10 (right plot).

The GLCM correlation in the 2018 texture-analysis was assumed, along a GLRLM feature, to possibly relate to a lower vascularity in xerostomic patients post-irradiation or an increase in adipose tissue [65]. This corresponds to the reported increased vascular resistance found by ultrasonic found by earlier studies, as described in section 2.1.4.

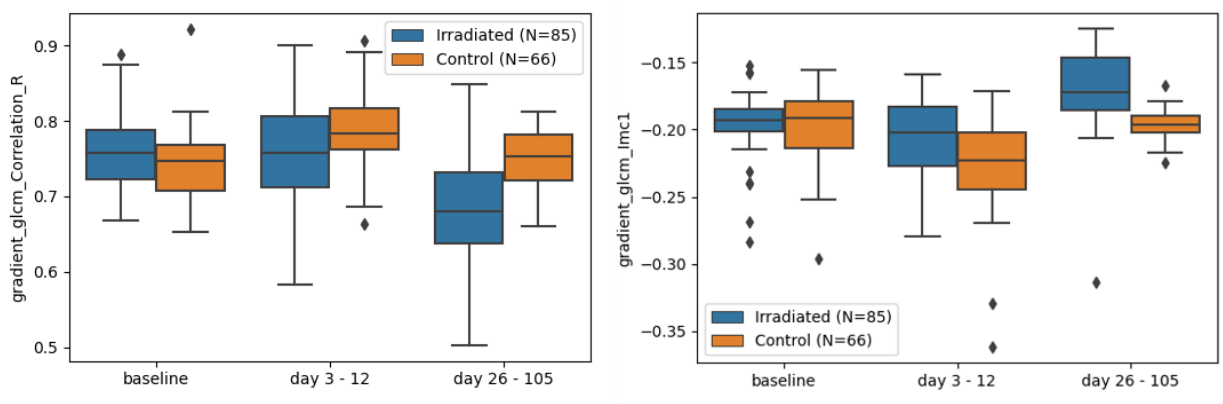


Figure 5-10: Left: Right version of LR-aggregated feature 759, right: LR-average of feature 765. Both are GLCM-based features calculated after applying a gradient filter.

Only two of the top-performing features scored with AUC above 0.50 when evaluated *univariately* for xerostomia predictions, both being textural features calculated from the gray level run length matrix (GLRLM).

Using no-p T2 LR-average feature 408 from after irradiation, without any image filtering, the long-run high gray-level emphasis predicts late xerostomia with AUC = 0.60 univariately and with AUC = 0.878 when combined with time and dose. The feature increases with longer runs in the GLRLM among the higher-intensity pixels, indicating a coarser textural structure among the high-intensity regions in the ROI.

The second GLRLM feature is the no-p T1 LR-aggregated feature 635R, for prediction of simultaneous xerostomia. Ft635 is the run variance (RV) in the GLRLM calculating after applying a square filter. The square filter squares all pixel intensities in the image, increasing the separation between high and low intensities in the ROI. The RV measures the variance in the runs, across the run lengths in the GLRLM. As seen in the right plot in Figure 5-11 xerostomic individuals have a higher RV than the non-xerostomic individuals, with little separation between irradiated and control.

In the 2016 study by van Dijk evaluating texture-features from CT images in relation to xerostomia, the GLRLM short run emphasis from the parotid gland was among the top predictors (section 2.3.6 and [61]). By visual inspection the feature was linked to an increase in fat saturation, as was also hypothesized by the 2018 CT study from IMRT-treated patients finding both a GLCM and GLRLM feature to be of significance [65] (as mentioned above in this chapter). As T1 images have bright signal from fat, but not from water, one might suspect the T1-version of feature 635R to capture some inter-mouse variance between the increase in adipose tissue following irradiation, which might also be present in a small subset of the control individuals, hindering normal function of the salivary glands making the mice xerostomic.

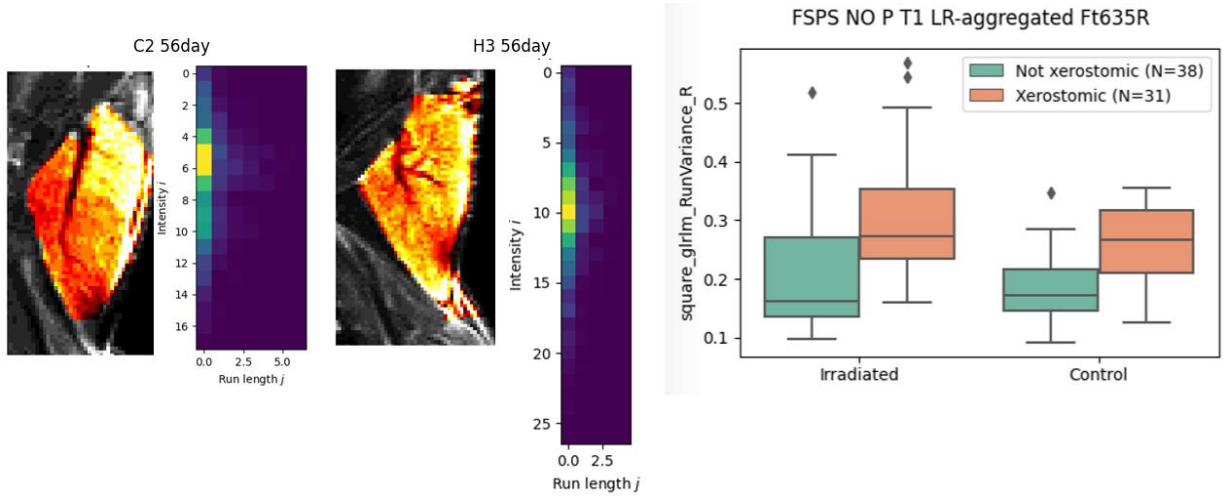


Figure 5-11: Left: images from control (C2) and irradiated (H3) individuals after applying a square filter, with corresponding GLRLMs. Right: the square GLRLM run variance for all control and irradiated individuals, separated by whether they are xerostomic.

The absolute top-performing feature in combination with a td-model was feature 216R from the no-p T2 LR-aggregated after-irr features (AUC = 0.941, BS = 0.098). The feature more than halved the BS relative the td-only model (relative change -0.514). Feature 216R measures the joint distribution of low dependence entries with emphasis on lower gray levels calculated from the GLDM [49], following a logarithm filter. As seen in Figure 5-12 the logarithm filter, in combination with the FBW discretization, compresses the number of gray levels relative the unfiltered image (8 intensities levels in the y-direction of the GLDM). However, this seemed to

increase the range of the dependencies (x-direction of the GLDM). One may therefore wonder if using a lower amount of discretization levels (bins) increased the differentiation between controls and irradiated individuals in the low-intensity dependency counts, measured by the feature.

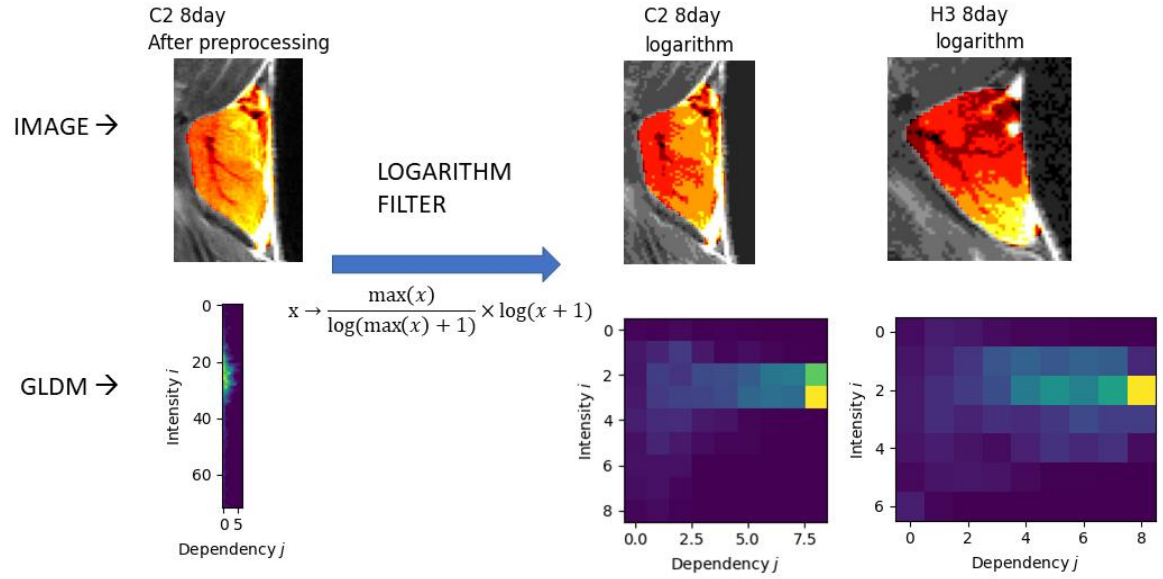


Figure 5-12: Illustration of resulting GLDMs before and after applying a logarithm filter. The GLDM for a control mouse (C2) and irradiated (H3) after logarithm filtering seen to the right, from the fist imaging day after irradiation (day 8). All images are from the right SMG unit.

5.7 Improvements and further studies

As with all radiomics-based studies, the repeatability of identified features with their predictive performance should optimally be evaluated on new external data. The reproducibility of radiomic features, e.g. stability across differently acquired data, could also be assessed to evaluate whether the acclaimed signal is just that – or noise with random correlations to the outcome. Evaluation of the feature robustness across different MRI scanners, protocols, and B-field variations is of interest when using MR-images for radiomics.

Assuming the results to have some relation to the true biological phenomenon as hypothesized, the best radiomic features from this work could be compared to features identified using a similar workflow in the next step of PROCCA: irradiated mice using protons. This may identify imaging biomarkers which are stable across external irradiation using both x-rays and protons, which have different ways of delivering dose to the target medium.

To aid the interpretation of top-performing features in future work, especially after image filtering and texture-based features, comparisons to other types of analysis of irradiated mice

relative control could be done. Examples include visual assessments of SG biopsies, histology analysis, and cytokine data.

As the segmentation method developed and used in this study have a high level of observer bias, the segmentation would optimally be repeated by an external observer (with expert knowledge about mouse physiology) to assess inter-observer variations in the delineation, such as what watershed regions to include and the choice of central L/ R slices. Future work may also use MRI-acquisitions with a lower difference between pixel distance and slice distance allowing for less loss of, or creation of artificial, information through interpolation into an isotropic voxel space for 3D radiomics. Capturing the SGs as single whole organs in 3D-radiomic features would allow for a less confusing analysis without considering LR-modes.

While the dose differences between the left and right SGs were assumed to be negligible, a Monte Carlo simulation may be used for accurate voxel-based dosimetry in a future work. This would allow for a more proper NTCP-model for comparison with the radiomic features, by incorporating the dose-volume histogram parameters such as the volume having received 90% of max dose (D90). In both future work, and as an improvement on this work, baseline saliva measurements could have been implemented into a separate NTCP comparison analysis – which would in this work have discarded some data, but allowed for a NTCP model more closely resembling what is in use clinically.

Regarding the preprocessing choices made in this work, some improvements may be suggested. When normalizing image intensities with a shifted standardization, the shift could have been made global based on the image acquisition (T1 or T2) in a similar fashion to the Nyul standard scales – making the image intensities at roughly the same values. Concerning the choice of using a FBW discretization, a FBC discretization may be more feasible for calculation of features after image filtering – which largely changes the scale and range of intensities in the images for some filters.

To mitigate potential data leakage the data might have been split into training and test data as early as normalization, using only training data to create the standard scales for Nyul normalization and in the further steps with potential data leakage. However, this could be hard to implement in combination with using LOOCV for evaluation of the models.

In a future work having more accurate saliva measurements, the xerostomia thresholding could be implemented on a per-individual basis: using the relative change to relative baseline saliva with some ratio-based threshold – allowing for inter-mice variabilities when assuming an individual to be xerostomic.

Future studies may also use more advanced MRI techniques, if possible with the small scanners used for mice, such as diffusion weighted MRI (dMRI) or sialography. Using dMRI may yield

information regarding changes in vascularity post-irradiation or over time. MR sialography utilized a heavily T2-weighted sequence to map the ductal architectures of the SGs [25], which may contain texture-related information obtained by radiomic analysis.

6 Conclusions

Imaging biomarkers from radiomic studies have the potential for incorporation into clinical oncology, supplying patient-specific information for baseline risk assessments and changes post-irradiation. Such information may be used as a decision-making tool for the oncologist, helping the shift towards precision oncology improving patient care in radiotherapy of head and neck carcinoma. Specifically, radiomic features from both T2- and T1-weighted MR-images showed potential for predicting radiation-induced changes in saliva production in vivo, using data from a pre-clinical study on mice. The features were calculated from 2D ROIs in the images, from both the left and right subunit of the fused sublingual and submandibular glands. Delta-features from before and after irradiation, and delta-p feature from before and after pilocarpine injections for measuring saliva production at the same day, were calculated and compared to the standard radiomic features (denoted no-p, or before-p, in relation to the pilocarpine injections).

Segmenting the salivary glands to create a region of interest was successfully performed using a watershed-based semi-automatic algorithm on each 2D MR-image separately. While segmented ROIs in both T1 and T2-images had a higher correlation to measured SLG area than SMG area in surgical specimens from 9 imaged individuals, the ROIs were assumed to contain most or all of the SMG with some SLG as the SMG is bigger than the SLG and the two glands are fused in mice. A pipeline for extraction of radiomic features considering preprocessing on a feature-specific level was successfully developed in accordance with the IBSI guidelines.

The saliva production from control individuals was found to be significantly different to irradiated individuals for measurements taken at day 26 or later, confirming earlier studies showing a post-irradiation reduction in saliva for both humans and rats. A regression analysis was performed to evaluate the predictive performance of time and dose on all saliva measurements (N=347), having a maximum R2 of 0.09 across LOOCV evaluated linear regression models. A xerostomia thresholding was performed to shift the prediction problem from regression to classification, using measured saliva from non-irradiated individuals, but due to high variance in the outcome the regressed thresholding line had a low R2 at 0.037. Nonetheless, the td-model using all data scored a maximum AUC of 0.67 when predicting xerostomia.

Using only image features, the no-p T2 LR-aggregated features had the highest R2 in a LOOCV evaluated regression analysis to simultaneously measured saliva (R2 = 0.09). However, the no-p T1 features had the highest AUC across LOOCV evaluated classifications to simultaneous saliva (AUC = 0.62) and were better predictors of late xerostomia (AUC = 0.72) along the T2 no-p features (AUC = 0.75). The no-p T1 features also showed a high predictive ability using only

baseline features to predict late xerostomia (AUC = 0.80). Only using delta-features for predicting late xerostomia was unsuccessful with no models having AUC above 0.50. Comparing no-p T1 to T2 features on the same subset of data the T1-models significantly outperformed the T2-models on predicting simultaneous xerostomia, while the T2-features had some non-significant improvement over the T1 models for predicting late xerostomia.

Using td-models in addition to one of each top-selected radiomic features from the previous LOOCV evaluated models, 14 features improved upon the td-models significantly. Among the top features almost all LR-aggregated features were from the right subunit, indicating some possible differences in post-irradiation responses between the two SG subunits.

The shape-based delta-feature elongation improved upon prediction of late xerostomia, possibly related to radiation-induced changes in SG shape. The delta-p first-order feature *energy* improved the td-model for predicting simultaneous xerostomia, assumed to be related to the difference in saliva content before and after saliva extraction. Most of the radiomic features improving upon the td-models were texture-based. A delta-feature related to the GLCM, along a no-p T2 feature, possibly captured some irradiation-induced changes related to vascularity or the salivatory ducts with clear separation between control and irradiated individuals in the latest days. Both a T1 and T2 weighted feature from the GLRLM was hypothesized to capture changes in fat-content, known to increase post-irradiation in the SGs.

Overall, the radiomic features seem to capture inter-mice variabilities relating to saliva production at the same day as the MR-images were acquired, and forward in time. However, due to small sample sizes and a lot of potential error sources, the results are in need of validation studies on external data to evaluate robustness and repeatability of the best prediction models and top-performing features.

Bibliography

[1] B. Roald, T. Sauer, and O. Klepp, "kreft," *Store medisinske leksikon*. May 29, 2022. Accessed: Sep. 29, 2022. [Online]. Available: http://sml.snl.no/kreft

[2] "Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av hode-/halskreft, 06.05.2020." https://www.helsebiblioteket.no/retningslinjer/hode-og-halskreft/forord (accessed May 19, 2022).

[3] "Strålebehandling, hode og hals," *NHI.no*. https://nhi.no/sykdommer/kreft/behandlingsmetoder/stralebehandling-hode-og-hals/ (accessed Sep. 29, 2022).

[4] P. Dirix, S. Nuyts, and W. Van den Bogaert, "Radiation-induced xerostomia in patients with head and neck cancer," *Cancer*, vol. 107, no. 11, pp. 2525–2534, 2006, doi: 10.1002/cncr.22302.

[5] T. Gupta *et al.*, "Prospective longitudinal assessment of parotid gland function using dynamic quantitative pertechnate scintigraphy and estimation of dose–response relationship of parotid-sparing radiotherapy in head-neck cancers," *Radiat. Oncol.*, vol. 10, no. 1, p. 67, Dec. 2015, doi: 10.1186/s13014-015-0371-2.

[6] A. K. Jha *et al.*, "Radiomics: a quantitative imaging biomarker in precision oncology," *Nucl. Med. Commun.*, vol. 43, no. 5, pp. 483–493, May 2022, doi: 10.1097/MNM.0000000000001543.

[7] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[8] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—'how-to' guide and critical reflection," *Insights Imaging*, vol. 11, no. 1, p. 91, Aug. 2020, doi: 10.1186/s13244-020-00887-2.

[9] P. Lambin *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, doi: 10.1016/j.ejca.2011.11.036.

[10] L. V. van Dijk *et al.*, "Delta-radiomics features during radiotherapy improve the prediction of late xerostomia," *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Aug. 2019, doi: 10.1038/s41598-019-48184-3.

[11] F. H. Attix, *Introduction to radiological physics and radiation dosimetry*. New York: Wiley, 1986.

[12] E. J. Hall and A. J. Giaccia, *Radiobiology for the radiologist*, 8th edition. Philadelphia Baltimore New York London Buenos Aires: Wolters Kluwer, 2019.

[13] J. B. Little, "Principal Cellular and Tissue Effects of Radiation," *Holl.-Frei Cancer Med. 6th Ed.*, 2003, Accessed: Sep. 26, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK12344/

[14] S. J. McMahon, "The linear quadratic model: usage, interpretation and challenges," *Phys. Med. Biol.*, vol. 64, no. 1, p. 01TR01, Dec. 2018, doi: 10.1088/1361-6560/aaf26a.

[15] P. Mayles, A. E. Nahum, and J.-C. Rosenwald, Eds., *Handbook of radiotherapy physics: theory and practice*. New York: Taylor & Francis, 2007.

[16] S. Cilla *et al.*, "Personalized automation of treatment planning in head-neck cancer: A step forward for quality in radiation therapy?," *Phys. Medica PM Int. J. Devoted Appl. Phys.*

Med. Biol. Off. J. Ital. Assoc. Biomed. Phys. AIFB, vol. 82, pp. 7–16, Feb. 2021, doi: 10.1016/j.ejmp.2020.12.015.
[17] S. M. Bentzen and J. Overgaard, "Patient-to-patient variability in the expression of radiation-induced normal tissue injury," *Semin. Radiat. Oncol.*, vol. 4, no. 2, pp. 68–80, Apr. 1994, doi: 10.1016/S1053-4296(05)80034-7.
[18] "saliva | biochemistry | Britannica." https://www.britannica.com/science/saliva (accessed May 19, 2022).
[19] "salivary gland | anatomy | Britannica." https://www.britannica.com/science/salivary-gland (accessed May 19, 2022).
[20] C. Maruyama, M. Monroe, J. Hunt, L. Buchmann, and O. Baker, "Comparing human and mouse salivary glands: A practice guide for salivary researchers," *Oral Dis.*, vol. 25, no. 2, pp. 403–415, 2019, doi: 10.1111/odi.12840.
[21] C. Rocchi, L. Barazzuol, and R. P. Coppes, "The evolving definition of salivary gland stem cells," *Npj Regen. Med.*, vol. 6, no. 1, p. 4, Dec. 2021, doi: 10.1038/s41536-020-00115-x.
[22] K. K. Skjørland, "xerostomi," *Store medisinske leksikon*. Aug. 12, 2019. Accessed: Jul. 17, 2022. [Online]. Available: http://sml.snl.no/xerostomi
[23] L. Franzén, U. Funegård, T. Ericson, and R. Henriksson, "Parotid gland function during and following radiotherapy of malignancies in the head and neck: A consecutive study of salivary flow and patient discomfort," *Eur. J. Cancer*, vol. 28, no. 2, pp. 457–462, Feb. 1992, doi: 10.1016/S0959-8049(05)80076-0.
[24] C. De la Cal *et al.*, "Radiation produces irreversible chronic dysfunction in the submandibular glands of the rat," *Open Dent. J.*, vol. 6, pp. 8–13, 2012, doi: 10.2174/1874210601206010008.
[25] S. C. H. Cheng, V. W. C. Wu, D. L. W. Kwong, and M. T. C. Ying, "Assessment of post-radiotherapy salivary glands," *Br. J. Radiol.*, vol. 84, no. 1001, pp. 393–402, May 2011, doi: 10.1259/bjr/66754762.
[26] A. Bjørnerud, *Compendium FYS4740/9740: The Physics of Magnetic Resonance Imaging*. Department of Physics, University of Oslo, 2020.
[27] M. H. Levitt, *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, 2008.
[28] F. Bloch, "Nuclear induction," *Physica*, vol. 17, no. 3, pp. 272–281, Mar. 1951, doi: 10.1016/0031-8914(51)90068-7.
[29] E. W. Weisstein, "Fourier Transform." https://mathworld.wolfram.com/ (accessed May 10, 2022).
[30] "k-space parts," *Questions and Answers in MRI*. http://mriquestions.com/parts-of-k-space.html (accessed May 13, 2022).
[31] "PE gradient," *Questions and Answers in MRI*. http://mriquestions.com/phase-encoding-gradient.html (accessed May 13, 2022).
[32] "Frequency Encoding," *Questions and Answers in MRI*. http://mriquestions.com/frequency-encoding.html (accessed May 13, 2022).
[33] "MRI interpretation - T1 v T2 images." https://www.radiologymasterclass.co.uk/tutorials/mri/t1_and_t2_images (accessed Sep. 18, 2022).
[34] M. T. Niknejad, "Short tau inversion recovery | Radiology Reference Article | Radiopaedia.org," *Radiopaedia*. https://radiopaedia.org/articles/short-tau-inversion-recovery?lang=us (accessed Oct. 07, 2022).

[35] J. Hennig, A. Nauerth, and H. Friedburg, “RARE imaging: A fast imaging method for clinical MR,” *Magn. Reson. Med.*, vol. 3, no. 6, pp. 823–833, 1986, doi: 10.1002/mrm.1910030602.
[36] E. Bellon, E. Haacke, P. Coleman, D. Sacco, D. Steiger, and R. Gangarosa, “MR artifacts: a review,” *Am. J. Roentgenol.*, vol. 147, no. 6, pp. 1271–1281, Dec. 1986, doi: 10.2214/ajr.147.6.1271.
[37] Z. Hou, “A Review on MR Image Intensity Inhomogeneity Correction,” *Int. J. Biomed. Imaging*, vol. 2006, p. 49515, 2006, doi: 10.1155/IJBI/2006/49515.
[38] “Gibbs (truncation) artifact,” *Questions and Answers in MRI*. http://mriquestions.com/gibbs-artifact.html (accessed Aug. 02, 2022).
[39] U. Bashir, “Magnetic susceptibility artifact | Radiology Reference Article | Radiopaedia.org,” *Radiopaedia*. https://radiopaedia.org/articles/magnetic-susceptibility-artifact (accessed Oct. 06, 2022).
[40] S. S. F. Yip *et al.*, “Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer,” *Sci. Rep.*, vol. 7, p. 3519, Jun. 2017, doi: 10.1038/s41598-017-02425-5.
[41] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. W. L. Aerts, “Machine Learning methods for Quantitative Radiomic Biomarkers,” *Sci. Rep.*, vol. 5, p. 13087, Aug. 2015, doi: 10.1038/srep13087.
[42] E.-R. Choi *et al.*, “Quantitative image variables reflect the intratumoral pathologic heterogeneity of lung adenocarcinoma,” *Oncotarget*, vol. 7, no. 41, pp. 67302–67313, Oct. 2016, doi: 10.18632/oncotarget.11693.
[43] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, “The causes and consequences of genetic heterogeneity in cancer evolution,” *Nature*, vol. 501, no. 7467, Art. no. 7467, Sep. 2013, doi: 10.1038/nature12625.
[44] X. Fave *et al.*, “Delta-radiomics features for the prediction of patient outcomes in non–small cell lung cancer,” *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, Apr. 2017, doi: 10.1038/s41598-017-00665-z.
[45] A. Traverso, L. Wee, A. Dekker, and R. Gillies, “Repeatability and Reproducibility of Radiomic Features: A Systematic Review,” *Int. J. Radiat. Oncol.*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018, doi: 10.1016/j.ijrobp.2018.05.053.
[46] A. Carré *et al.*, “Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Jul. 2020, doi: 10.1038/s41598-020-69298-z.
[47] E. Scalco *et al.*, “T2w-MRI signal normalization affects radiomics features reproducibility,” *Med. Phys.*, vol. 47, no. 4, pp. 1680–1691, 2020, doi: 10.1002/mp.14038.
[48] IBSI, “IBSI 2.” https://theibsi.github.io/ibsi2/ (accessed Sep. 22, 2022).
[49] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, and for Initiative, “Image biomarker standardisation initiative - feature definitions,” Dec. 2016.
[50] S. Rizzo *et al.*, “Radiomics: the facts and the challenges of image analysis,” *Eur. Radiol. Exp.*, vol. 2, no. 1, p. 36, Nov. 2018, doi: 10.1186/s41747-018-0068-z.
[51] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[52] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Oct. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[53] M. M. Galloway, “Texture analysis using gray level run lengths,” *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, Jun. 1975, doi: 10.1016/S0146-664X(75)80008-6.
[54] G. Thibault *et al.*, “Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification,” presented at the 10th International Conference on Pattern Recognition and Information Processing, Nov. 2009.
[55] C. Sun and W. G. Wee, “Neighboring gray level dependence matrix for texture classification,” *Comput. Vis. Graph. Image Process.*, vol. 23, no. 3, pp. 341–352, Sep. 1983, doi: 10.1016/0734-189X(83)90032-4.
[56] M. Amadasun and R. King, “Textural features corresponding to textural properties,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 5, pp. 1264–1274, Sep. 1989, doi: 10.1109/21.44046.
[57] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
[58] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O’Leary, “PyWavelets: A Python package for wavelet analysis,” *J. Open Source Softw.*, vol. 4, no. 36, p. 1237, Apr. 2019, doi: 10.21105/joss.01237.
[59] K. Pearson, “LIII. *On lines and planes of closest fit to systems of points in space*,” *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.
[60] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.
[61] I. Beetz *et al.*, “NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: The role of dosimetric and clinical factors,” *Radiother. Oncol.*, vol. 105, no. 1, pp. 101–106, Oct. 2012, doi: 10.1016/j.radonc.2012.03.004.
[62] T. Berger *et al.*, “MO-0876 Assessing the generalisability of radiomics features for predicting sticky saliva and xerostomia,” *Radiother. Oncol.*, vol. 170, pp. S762–S764, May 2022, doi: 10.1016/S0167-8140(22)02442-2.
[63] H. Elhalawani *et al.*, “EP-2121: Serial Parotid Gland Radiomic-based Model Predicts Post-Radiation Xerostomia in Oropharyngeal Cancer,” *Radiother. Oncol.*, vol. 127, pp. S1167–S1168, Apr. 2018, doi: 10.1016/S0167-8140(18)32430-7.
[64] H. S. Gabryś, F. Buettner, F. Sterzing, H. Hauswald, and M. Bangert, “Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia,” *Front. Oncol.*, vol. 8, p. 35, Mar. 2018, doi: 10.3389/fonc.2018.00035.
[65] V. Nardone *et al.*, “Texture analysis as a predictor of radiation-induced xerostomia in head and neck patients undergoing IMRT,” *Radiol. Med. (Torino)*, vol. 123, no. 6, pp. 415–423, Jun. 2018, doi: 10.1007/s11547-017-0850-7.
[66] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. 2013. [Online]. Available: http://link.springer.com/book/10.1007/978-1-4614-7138-7
[67] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, Second edition. Beijing [China] ; Sebastopol, CA: O’Reilly Media, Inc, 2019.

[68] S. Mottaghinejad, “Bias and variance, but what are they really?,” *Medium*, Mar. 18, 2021. https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171 (accessed Jun. 14, 2022).
[69] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer Science & Business Media, 2012.
[70] L. Breiman, “Bagging Predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1023/A:1018054314350.
[71] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Aug. 1995, vol. 1, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.
[72] P. Bühlmann and S. A. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*. Heidelberg ; New York: Springer, 2011.
[73] “sklearn.metrics.r2_score,” *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.metrics.r2_score.html (accessed Aug. 16, 2022).
[74] S. Shayesteh *et al.*, “Treatment response prediction using MRI-based pre-, post-, and delta-radiomic features and machine learning algorithms in colorectal cancer,” *Med. Phys.*, vol. 48, no. 7, pp. 3691–3701, 2021, doi: 10.1002/mp.14896.
[75] Dr James Newcombe, “Sensitivity and specificity - NSW Health Pathology - Website.” https://www.pathology.health.nsw.gov.au/covid-19-testing/sensitivity-and-specificity (accessed Aug. 26, 2022).
[76] D. J. Hand, “Measuring classifier performance: a coherent alternative to the area under the ROC curve,” *Mach. Learn.*, vol. 77, no. 1, pp. 103–123, Oct. 2009, doi: 10.1007/s10994-009-5119-5.
[77] Jacob Goldstein-Greenwood, “A Brief on Brier Scores | University of Virginia Library Research Data Services + Sciences.” https://data.library.virginia.edu/a-brief-on-brier-scores/ (accessed Sep. 01, 2022).
[78] A. Bani-Sadr *et al.*, “Conventional MRI radiomics in patients with suspected early- or pseudo-progression,” *Neuro-Oncol. Adv.*, vol. 1, no. 1, p. vdz019, May 2019, doi: 10.1093/noajnl/vdz019.
[79] Eirik Malinen, Hilde Kanli Galtung, Åslaug Helland, and Silje Endresen Reme, “Protons contra cancer (PROCCA) - UiO:Life Science.” https://www.uio.no/english/research/strategic-research-areas/life-science/research/convergence-environments/procca/index.html (accessed Sep. 23, 2022).
[80] I. S. Juvkam *et al.*, “A preclinical model to investigate normal tissue damage following fractionated radiotherapy to the head and neck.” bioRxiv, p. 2022.05.19.492439, May 19, 2022. doi: 10.1101/2022.05.19.492439.
[81] *Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes Text with EEA relevance*, vol. 276. 2010. Accessed: Jun. 27, 2022. [Online]. Available: http://data.europa.eu/eli/dir/2010/63/oj/eng
[82] B. Baeßler, K. Weiss, and D. Pinto dos Santos, “Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study,” *Invest. Radiol.*, vol. 54, no. 4, pp. 221–228, Apr. 2019, doi: 10.1097/RLI.0000000000000530.
[83] I. N. Bankman, Ed., *Handbook of medical imaging: processing and analysis*. San Diego, Calif.: Academic Press, 2000.

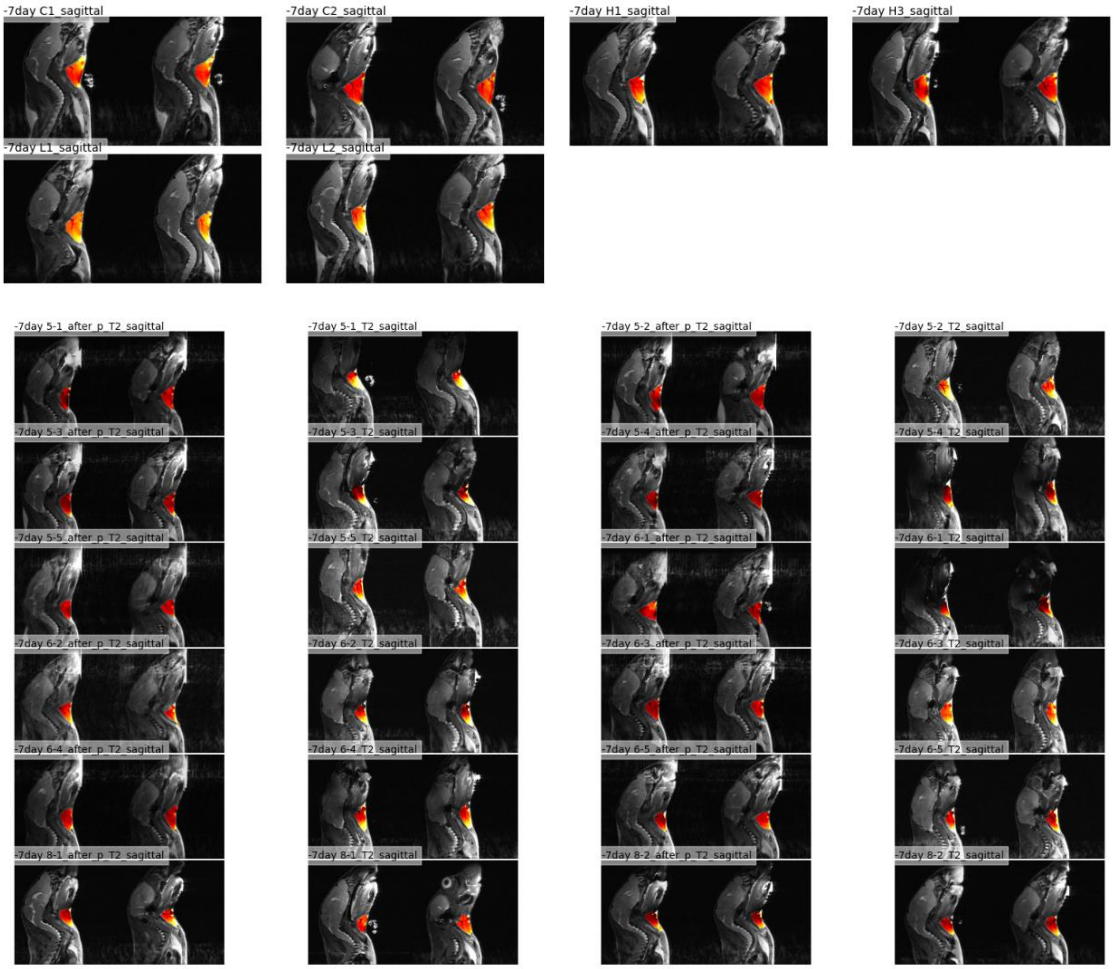
[84] S. M. Pizer *et al.*, “Adaptive histogram equalization and its variations,” *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.
[85] G. van Rossum and F. L. Drake, *The Python language reference*, Release 3.0.1 [Repr.]. Hampton, NH: Python Software Foundation, 2010.
[86] “opencv/opencv-python.” OpenCV, Jun. 29, 2022. Accessed: Jun. 29, 2022. [Online]. Available: https://github.com/opencv/opencv-python
[87] P. Soille, “On morphological operators based on rank filters,” *Pattern Recognit.*, vol. 35, no. 2, pp. 527–535, Feb. 2002, doi: 10.1016/S0031-3203(01)00047-4.
[88] S. van der Walt *et al.*, “scikit-image: image processing in Python,” *PeerJ*, vol. 2, p. e453, Jun. 2014, doi: 10.7717/peerj.453.
[89] L. G. Shapiro and G. C. Stockman, *Computer vision*. Upper Saddle River, NJ: Prentice Hall, 2001.
[90] R. C. Gonzalez, R. E. Woods, and B. R. Masters, “Digital Image Processing, Third Edition,” *J. Biomed. Opt.*, vol. 14, no. 2, p. 029901, 2009, doi: 10.1117/1.3115362.
[91] L. Vincent and P. Soille, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, Jun. 1991, doi: 10.1109/34.87344.
[92] W. E. Higgins and E. J. Ojard, “Interactive morphological watershed analysis for 3D medical images,” *Comput. Med. Imaging Graph.*, vol. 17, no. 4, pp. 387–395, Jul. 1993, doi: 10.1016/0895-6111(93)90033-J.
[93] “pydicom: An open source DICOM library.” Apr. 07, 2022. Accessed: Apr. 07, 2022. [Online]. Available: https://github.com/pydicom/pydicom
[94] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, Art. no. 7825, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[95] “Markers for watershed transform — skimage v0.19.2 docs.” https://scikit-image.org/docs/stable/auto_examples/segmentation/plot_marked_watershed.html?highlight=median (accessed Jul. 04, 2022).
[96] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: 10.1109/TSMC.1979.4310076.
[97] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, “SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research,” *J. Digit. Imaging*, vol. 31, no. 3, pp. 290–303, Jun. 2018, doi: 10.1007/s10278-017-0037-8.
[98] D. Palumbo, B. Yee, P. O’Dea, S. Leedy, S. Viswanath, and A. Madabhushi, “Interplay between bias field correction, intensity standardization, and noise filtering for T2-weighted MRI,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2011, pp. 5080–5083. doi: 10.1109/IEMBS.2011.6091258.
[99] N. J. Tustison *et al.*, “N4ITK: Improved N3 Bias Correction,” *IEEE Trans. Med. Imaging*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.
[100] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, “A nonparametric method for automatic correction of intensity nonuniformity in MRI data,” *IEEE Trans. Med. Imaging*, vol. 17, no. 1, pp. 87–97, Feb. 1998, doi: 10.1109/42.668698.
[101] E. W. Weisstein, “B-Spline.” https://mathworld.wolfram.com/ (accessed Jun. 02, 2022).
[102] “N4 Bias Field Correction — SimpleITK 2.0rc2 documentation.” https://simpleitk.readthedocs.io/en/master/link_N4BiasFieldCorrection_docs.html?highlight=n4 (accessed Jul. 08, 2022).

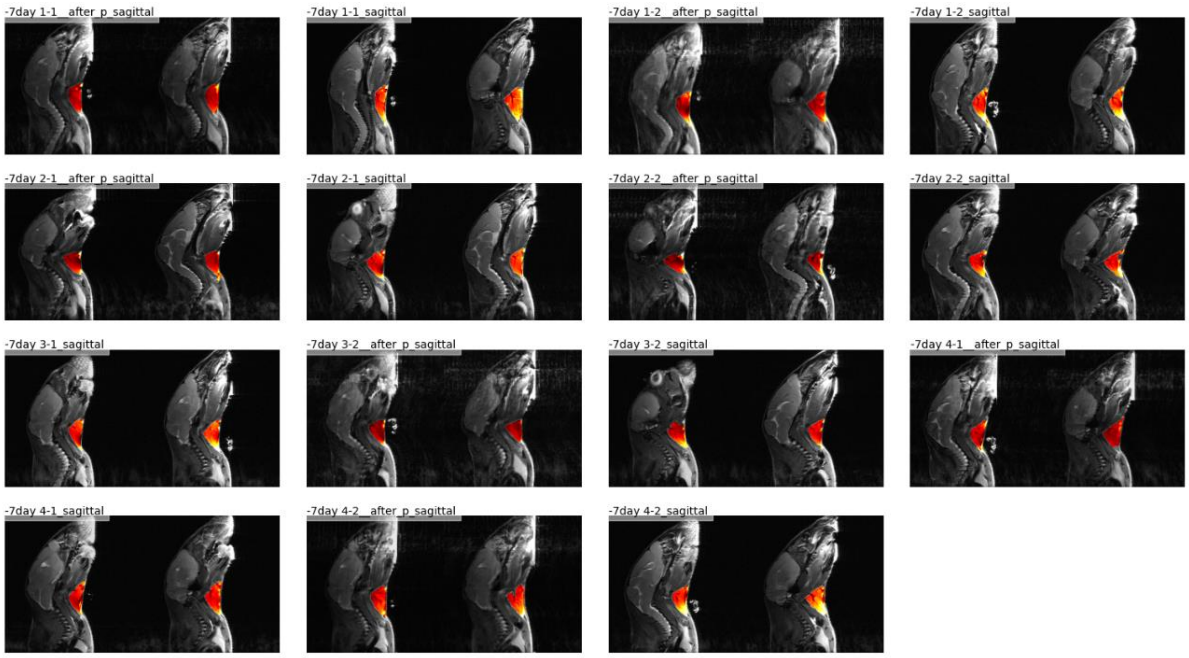
[103] G. Collewet, M. Strzelecki, and F. Mariette, “Influence of MRI acquisition protocols and image intensity normalization methods on texture classification,” *Magn. Reson. Imaging*, vol. 22, no. 1, pp. 81–91, Jan. 2004, doi: 10.1016/j.mri.2003.09.001.
[104] L. G. Nyul, J. K. Udupa, and X. Zhang, “New variants of a method of MRI scale standardization,” *IEEE Trans. Med. Imaging*, vol. 19, no. 2, pp. 143–150, Feb. 2000, doi: 10.1109/42.836373.
[105] L. Duron *et al.*, “Gray-level discretization impacts reproducible MRI radiomics texture features,” *PLOS ONE*, vol. 14, no. 3, p. e0213459, Mar. 2019, doi: 10.1371/journal.pone.0213459.
[106] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.
[107] V. Nardone *et al.*, “Delta radiomics: a systematic review,” *Radiol. Med. (Torino)*, vol. 126, no. 12, pp. 1571–1583, Dec. 2021, doi: 10.1007/s11547-021-01436-7.
[108] A. Crombé *et al.*, “T_2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy.: Sarcoma Δ-Radiomics Response Prediction,” *J. Magn. Reson. Imaging*, vol. 50, no. 2, pp. 497–510, Aug. 2019, doi: 10.1002/jmri.26589.
[109] Z. Zhao, R. Anand, and M. Wang, “Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform.” arXiv, Aug. 14, 2019. doi: 10.48550/arXiv.1908.05376.
[110] smazzanti, “smazzanti/mrmr.” Aug. 03, 2022. Accessed: Aug. 05, 2022. [Online]. Available: https://github.com/smazzanti/mrmr
[111] M. Wada *et al.*, “Circadian clock-dependent increase in salivary IgA secretion modulated by sympathetic receptor activation in mice,” *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, Aug. 2017, doi: 10.1038/s41598-017-09438-0.
[112] W. Rogers *et al.*, “Radiomics: from qualitative to quantitative imaging,” *Br. J. Radiol.*, vol. 93, no. 1108, p. 20190948, Apr. 2020, doi: 10.1259/bjr.20190948.
[113] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, and E. Chouzenoux, “Impact of Preprocessing and Harmonization Methods on the Removal of Scanner Effects in Brain MRI Radiomic Features,” *Cancers*, vol. 13, no. 12, p. 3000, Jun. 2021, doi: 10.3390/cancers13123000.
[114] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proceedings of the 22nd international conference on Machine learning - ICML ’05*, Bonn, Germany, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.
[115] “QIBA Profile Stages - QIBA Wiki.” https://qibawiki.rsna.org/index.php/QIBA_Profile_Stages (accessed Sep. 19, 2022).

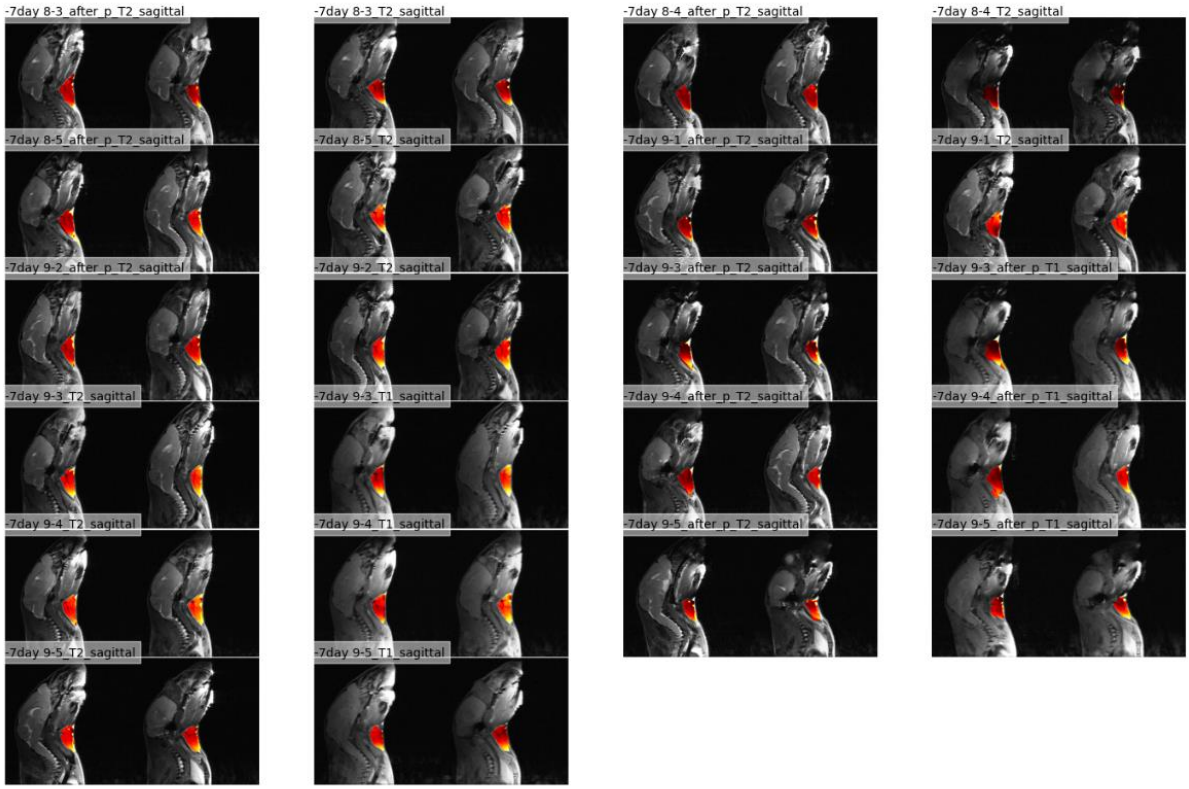
Appendix A: Segmentation hyperparameters

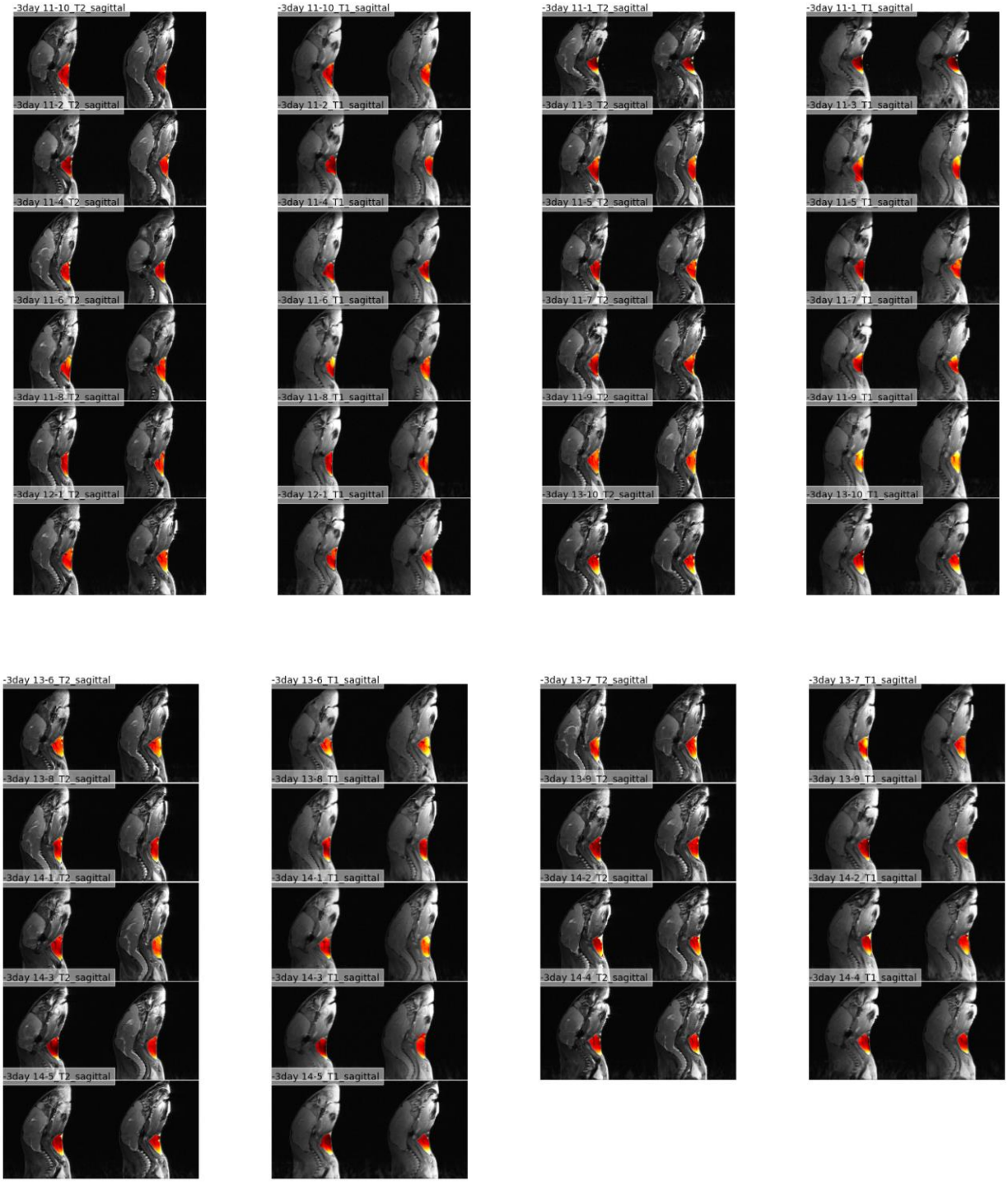
Segmentation parameter tuple (median disk-size, gradient disk-size, marker disk-size, marker threshold)	Number of images segmented
(3, 4, 13, 2)	279
(4, 5, 10, 2)	35
(2, 5, 10, 2)	8
(4, 4, 12, 2)	8
(2, 4, 10, 2)	2
(2, 8, 14, 2)	1
(3, 6, 6, 2)	1

Appendix B: Segmented regions of interests at baseline




-7day 1-1_ after_p_sagittal
-7day 1-1_sagittal
-7day 1-2_ after_p_sagittal
-7day 1-2_sagittal
-7day 2-1_ after_p_sagittal
-7day 2-1_sagittal
-7day 2-2_ after_p_sagittal
-7day 2-2_sagittal
-7day 3-1_sagittal
-7day 3-2_ after_p_sagittal
-7day 3-2_sagittal
-7day 4-1_ after_p_sagittal
-7day 4-1_sagittal
-7day 4-2_ after_p_sagittal
-7day 4-2_sagittal


-7day 8-3_after_p_T2_sagittal
-7day 8-3_T2_sagittal
-7day 8-4_after_p_T2_sagittal
-7day 8-4_T2_sagittal
-7day 8-5_after_p_T2_sagittal
-7day 8-5_T2_sagittal
-7day 9-1_after_p_T2_sagittal
-7day 9-1_T2_sagittal
-7day 9-2_after_p_T2_sagittal
-7day 9-2_T2_sagittal
-7day 9-3_after_p_T2_sagittal
-7day 9-3_after_p_T1_sagittal
-7day 9-3_T2_sagittal
-7day 9-3_T1_sagittal
-7day 9-4_after_p_T2_sagittal
-7day 9-4_after_p_T1_sagittal
-7day 9-4_T2_sagittal
-7day 9-4_T1_sagittal
-7day 9-5_after_p_T2_sagittal
-7day 9-5_after_p_T1_sagittal
-7day 9-5_T2_sagittal
-7day 9-5_T1_sagittal



* left / right segmented ROI seen as separate slice for each mouse at baseline times (day -7 or -3)

Appendix C: Register of radiomic features by filters and type

IDX	FILTER_TYPE		IDX	FILTER_TYPE
1-9	original_shape		376-380	lbp-2D_ngtdm
10-27	original_firstorder		381-402	original_glcm
28-43	original_glszm		403-418	original_glrlm
44-57	original_gldm		419-436	wavelet-H_firstorder
58-62	original_ngtdm		437-458	wavelet-H_glcm
63-80	square_firstorder		459-474	wavelet-H_glrlm
81-96	square_glszm		475-490	wavelet-H_glszm
97-110	square_gldm		491-504	wavelet-H_gldm
111-115	square_ngtdm		505-509	wavelet-H_ngtdm
116-133	squareroot_firstorder		510-527	wavelet-L_firstorder
134-149	squareroot_glszm		528-549	wavelet-L_glcm
150-163	squareroot_gldm		550-565	wavelet-L_glrlm
164-168	squareroot_ngtdm		566-581	wavelet-L_glszm
169-186	logarithm_firstorder		582-595	wavelet-L_gldm
187-202	logarithm_glszm		596-600	wavelet-L_ngtdm
203-216	logarithm_gldm		601-622	square_glcm
217-221	logarithm_ngtdm		623-638	square_glrlm
222-239	exponential_firstorder		639-660	squareroot_glcm
240-255	exponential_glszm		661-676	squareroot_glrlm
256-269	exponential_gldm		677-698	logarithm_glcm
270-274	exponential_ngtdm		699-714	logarithm_glrlm
275-292	gradient_firstorder		715-736	exponential_glcm
293-308	gradient_glszm		737-752	exponential_glrlm
309-322	gradient_gldm		753-774	gradient_glcm
323-327	gradient_ngtdm		775-790	gradient_glrlm
328-345	lbp-2D_firstorder		791-812	lbp-2D_glcm
346-361	lbp-2D_glszm		813-828	lbp-2D_glrlm
362-375	lbp-2D_gldm			

Appendix D: Additional classification results

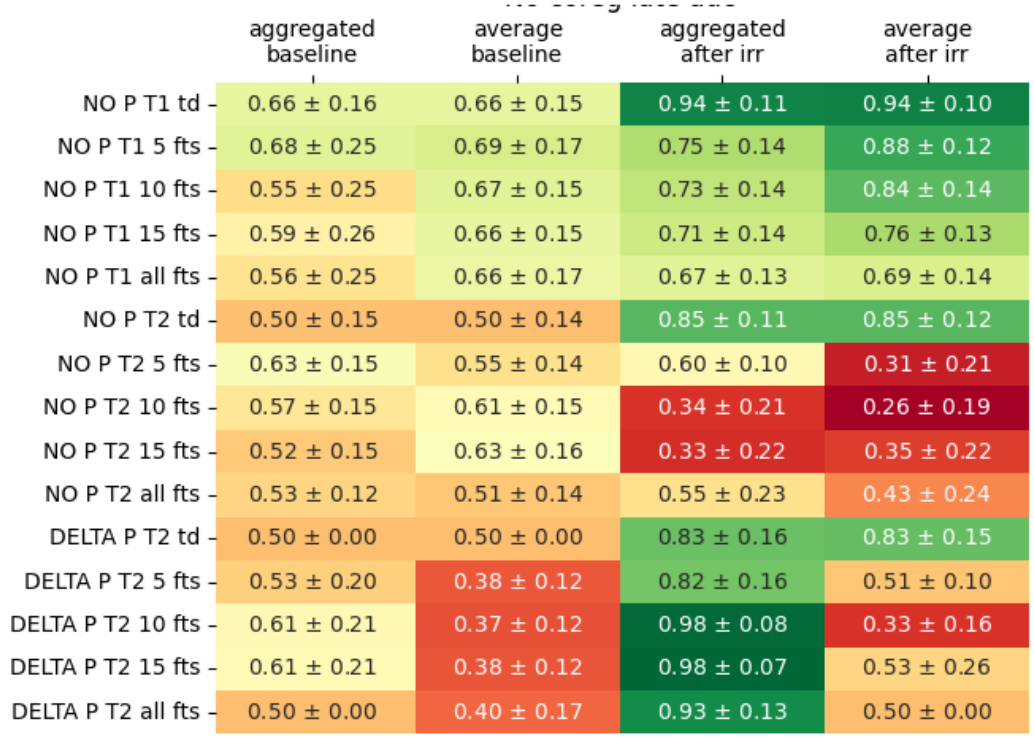


Figure 6-1: AUC results from prediction of late xerostomia using various amounts of top features from four feature-spaces, including only time and dose (td).

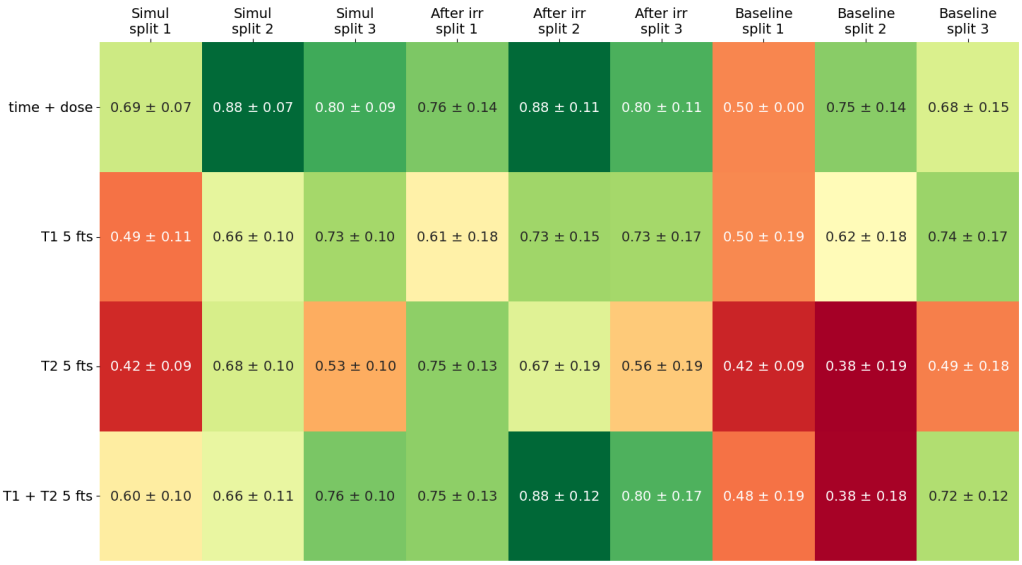


Figure 6-2: Mean of AUCs ± sd of AUCs for RF models using instances where both T1 and T2 images are present. Columns: prediction mode and test / train split. Rows: features used in model

Appendix E: Best k features across LOOCV evaluated regression models