Outcome predictors for Gamma Knife radiosurgery on vestibular schwannoma

Outcome predictors for Gamma Knife radiosurgery on vestibular schwannoma

Outcome predictors for Gamma Knife radiosurgery on vestibular schwannoma

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op vrijdag 18 december 2020 om 13:30 uur

door

Patrick Petrus Johannes Hendrikus Langenhuizen

geboren te Nijmegen

Dit proefschrift is goedgekeurd door de promotor en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. M. Matters-Kammerer
1e promotor:	prof.dr.ir. P.H.N. de With
copromotor(en):	dr. S. Zinger
	dr. H.B. Verheul (Elisabeth-TweeSteden Ziekenhuis Tilburg)
leden:	prof.dr. M. Smits (Erasmus MC)
	prof.dr.ir. B.M. ter Haar Romeny
	prof.dr. S. Leenstra (Erasmus MC)
	prof.dr. H.P.M. Kunst (Maastricht UMC+)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

To my wife, for her boundless support, encouragement and love.

Outcome predictors for Gamma Knife radiosurgery on vestibular schwannoma

Patrick Petrus Johannes Hendrikus Langenhuizen

Cover design: Patrick Langenhuizen
Printed by: ProefschiftMaken

ISBN 978-90-386-5173-6
NUR-code 959

Summary

Outcome predictors for Gamma Knife radiosurgery on vestibular schwannoma

Vestibular schwannomas (VSs) are benign intracranial tumors originating from the eighth cranial nerve. Since most symptoms usually do not improve after treatment, the main treatment goal has shifted in the last decades from complete removal of tumor tissue to functional preservation of the vestibulocochlear nerve and other adjacent cranial nerves. However, selecting the optimal treatment modality, i.e. microsurgery or radiosurgery, remains ambiguous. Each modality has its own specific advantages and disadvantages, and the scientific body of evidence concerning which strategy is most suited, is hampered by the lack of randomized control studies. As such, it remains a difficult task to select the optimal treatment strategy, especially on an individual basis. Currently, in the majority of cases, radiosurgery is selected for small- to medium-sized VSs and microsurgery for large VSs. However, it is unclear whether this strategy provides the optimal treatment.

In order to aid the physician and VS patient in selecting the optimal treatment strategy, this thesis investigates the possibility of *a-priori* prediction of the Gamma Knife radiosurgery (GKRS) treatment response of the VS tumor on an individual basis. This investigation involves five distinct fields. First, to enable evaluation of the GKRS treatment response, a large database containing data on many patients is created. Since the response to GKRS involves slow, time-consuming processes, the follow-up times need to be long enough for the included patients to enable objective assessment of the treatment response. This resulted in the inclusion of 735 unilateral sporadic VS patients, treated between 2002 and 2014 with the Gamma Knife at the ETZ in Tilburg.

Second, because evidence-based ground-truth on transient swelling and on true tumor progression following GKRS are unavailable, clear and objective definitions of the various short- and long-term treatment responses are designed. Tumor volume changes, calculated on over 4,000 follow-up MRI scans, are employed for these measures. Therefore, inter- and intra-observer variabilities in tumor contouring are assessed to analyze the impact of multi-reader annotations. Our experiments show that for non-small tumors, the annotation variability is well below 10% in volume.

Third, in Chapter 3, the influence of the readily available, patient-specific pre-treatment growth rate on the GKRS treatment response is evaluated. To this end,

pre-treatment MRI scans of 311 patients are obtained and growth rates are calculated by so-called *volume doubling times* (VDTs). The experiments show that relative short-term volume changes have no statistically significant correlation to the pre-treatment VDTs. However, for the long-term treatment response, Kaplan-Meier survival analyses have revealed that slow-growing tumors, with a VDT equal to or longer than the median VDT of 15 months, have calculated 5- and 10-year tumor control rates of 97.3% and 86.0%, respectively, whereas fast-growing tumors have tumor control rates of 85.5% and 67.6%, respectively. This difference proves to be statistically significant using the log-rank test ($p < 0.01$). The influence of the VDT on tumor control is also determined by employing Cox regression analyses, obtaining a significant ($p < 0.05$) effect of the VDT on the hazard rates for loss of tumor control. The resulting model enables the calculation of the risk at treatment failure on an individual basis.

Fourth, in Chapter 4, the impact of the treatment planning on the GKRS response is investigated. This chapter is split into three different parts. First, the global treatment-related parameters, such as the planning quality scores and radiation doses to the tumor margin, are statistically evaluated for the complete patient cohort. The obtained results show that there is no significant impact of the global parameters on the final treatment outcome. Second, the variability in tumor contouring is evaluated on a small dataset, where 20 tumors showed a significant short-term volume decrease, and 20 tumors in which the treatment resulted in a loss of tumor control. In this limited dataset, we have determined that the variability in tumor contouring is larger in the significant volume reduction cohort, compared to the variability in the treatment failure cohort. Third, the impact of the heterogeneous dose distribution on the treatment outcome is evaluated. We introduce a novel approach for assessing the dose distribution. To this end, three-dimensional histograms of oriented gradients are employed in conjunction with support vector machines and principal component analysis. The resulting model obtained an accuracy of 77.5%, suggesting the influence of the treatment on the outcome itself.

Fifth, in Chapters 5, 6, and 7, the visual properties of VS tumors on MRI scans are examined and captured with quantitative image features. These chapters investigate the potential of predicting the individual GKRS treatment response of VS tumors employing these *radiomic* features. These features capture the intrinsic tumor texture and shape differences that may reflect variations in the underlying tumor biology. In Chapter 5, the first experiments into relating these image features to the treatment response are conducted. In the limited dataset of 40 tumors previously employed in Chapter 4, experiments show that the tumor shape is not suitable for predicting the treatment response. Furthermore, various quantitative tumor texture features are extracted from these 40 tumors for evaluating their correlation to the treatment outcome. It is determined that second-order statistical metrics distilled from gray-level co-occurrence matrices (GLCMs) and run-length matrices are suitable for describing texture differences, but are slightly outperformed by simple first-order statistics. The obtained prediction accuracy is

about 85%.

Next, in Chapter 6, the predictability of TTE employing MRI tumor texture features is assessed. To enable comparison between different MRI scans, a tissue-specific MRI normalization method is introduced. Furthermore, the impact of imbalance in the data is investigated, and the influence of different tumor sizes on the prediction results are examined. The experiments show that a set of four GLCM-features lead to the best discrimination between patients suffering from TTE and those not suffering from TTE. Moreover, these results increase for larger tumor volumes, obtaining sensitivity and specificity values of 0.77 and 0.89, respectively. The impact of the imbalance in the data is also visible in the results: models trained on lower performing features show results that are skewed to the majority class. However, for the higher performing models, the imbalance leads to improved prediction results due to the availability of more data for training.

Finally, in Chapter 7, the impact of MRI-based tumor texture features on the long-term tumor control is considered. By exploiting the strict and objective definitions for treatment failure and long-term tumor control, two distinct patient cohorts are selected from the unique large ETZ database. The experiments lead to a model that ultimately can predict whether the GKRS treatment will result in long-term tumor control, or a failed treatment where tumor progression is not stopped. The resulting prediction model obtained accuracy, sensitivity, and specificity scores of 0.77, 0.71, and 0.83, respectively. Again, these results depend on the size of the tumor. By increasing the lower bound for the tumor volume, the obtained results improve. Including tumors larger than 5 cc result in the best-performing model, obtaining accuracy, sensitivity, and specificity values of 0.83, 0.83, and 0.82, respectively. These results clearly show the possibility of predicting the long-term GKRS treatment response, utilizing MRI tumor texture features.

The work in this thesis demonstrates that computer-aided methods can facilitate the physicians and patients in establishing an optimal treatment strategy on an individual basis. The research has revealed that variations in the intrinsic tumor biology are most likely causing the differences in treatment response of vestibular schwannomas following Gamma Knife radiosurgery. The promising results obtained in this work show the feasibility of predicting the short- and long-term treatment response of vestibular schwannomas on an individual basis, using MRI-based data such as the pre-treatment growth rate and tumor texture features. These results can be exploited for further research into creating a clinical decision-support system, facilitating physicians and patients to select a personalized optimal care path.

Samenvatting

Resultaatvoorspelling van Gamma Knife radiochirurgie op vestibulair schwannomen

Vestibulaire schwannomen (VS) zijn goedaardige hersentumoren die ontstaan vanuit de achtste hersenzenuw. Aangezien de meeste symptomen niet verbeteren na behandeling, is het doel van de behandeling in de laatste decennia veranderd van het compleet verwijderen van de tumor naar het behoud van zenuwfunctionaliteit. Het selecteren van de optimale behandelstrategie (radiochirurgie of microchirurgie) blijft echter moeilijk. Beide opties hebben elk hun eigen voor- en nadelen en wetenschappelijk bewijs voor welke strategie de beste is, is nauwelijks beschikbaar omdat er geen gerandomiseerde studies zijn gedaan. Hierdoor is de keuze voor de beste behandelstrategie een dilemma, zeker op individueel niveau. Momenteel wordt in de meeste gevallen gekozen voor radiochirurgie bij kleine tot middelgrote tumoren en voor microchirurgie bij grote tumoren. Het blijft onduidelijk of dit de optimale strategie is op individuele basis.

Om de behandelend arts en de VS patiënt te helpen bij het selecteren van de optimale behandelstrategie, wordt in deze thesis onderzocht of het mogelijk is om de behandelresultaten van Gamma Knife radiochirurgie (GKRS) vooraf te voorspellen op individuele basis. Dit onderzoek omvat vijf verschillende aandachtsgebieden. Ten eerste is er een grote database gecreëerd met data van veel patiënten. Aangezien de behandelresponsie een langdurige periode omvat, is een lange opvolgingstijd na radiochirurgie vereist om objectief de behandelresultaten te kunnen beoordelen. Dit heeft geresulteerd in de inclusie van 735 patiënten met een eenzijdige VS, welke behandeld zijn tussen 2002 en 2014 met de Gamma Knife in het ETZ in Tilburg.

Ten tweede, aangezien de ware toestand van een tumor voor tijdelijke zwelling of voor uiteindelijke tumorcontrole niet te bewijzen is, zijn er duidelijke en objectieve definities opgesteld voor de behandelreacties op korte en lange termijn. Veranderingen in tumorvolume, welke gebaseerd zijn op analyses van meer dan 4000 MRI scans, zijn gebruikt om deze definities te bepalen. Tevens is er onderzocht wat de impact is van variabele annotaties tussen de diverse medische waarnemers op de tumorvolumes. Deze experimenten laten zien dat de variaties voor de middelgrote en grotere tumoren ruim onder de 10% liggen.

Ten derde is in Hoofdstuk 3 de invloed van de beschikbare patiënt-specifieke VS groeisnelheid voor behandeling op de GKRS behandelreactie onderzocht. Hiervoor zijn MRI scans van 311 patiënten van voor de behandeling verkregen en

groeisnelheden berekend volgens de zogenaamde volumeverdubbelingstijden (VDTs). De experimenten tonen aan dat de relatieve korte-termijn volumeveranderingen geen statistisch significante correlatie hebben met de VDTs. Kaplan-Meier overlevingsanalyses laten echter zien dat de langetermijnreacties voor langzaam groeiende tumoren, met een VDT gelijk aan of groter dan de mediaan VDT van 15 maanden, 5- en 10-jaars tumorcontroleratio's hebben van respectievelijk 97,3% en 86,0%, terwijl snelgroeiende tumoren tumorcontroleratio's van respectievelijk 85,5% en 67,6%, laten zien. Dit verschil is statistisch significant volgens de log-rank test ($p<0,01$). De invloed van de VDT op tumorcontrole wordt ook bevestigd door de Cox regressieanalyse, die een significant effect ($p<0,05$) van de VDT heeft op het relatieve risico van verlies op tumorcontrole. Het resulterende model maakt het mogelijk om het risico op een mislukte behandeling te berekenen.

Ten vierde is in Hoofdstuk 4 de impact van de behandelingsplanning op de GKRS-responsie geanalyseerd. Dit hoofdstuk is in drie delen gesplitst. Als eerste zijn de globale behandelparameters, zoals de kwaliteitsindices en de bestralingsdoses op de rand van de tumor, statistisch geëvalueerd. De behaalde resultaten laten zien dat er geen significante impact is van deze globale parameters op de uiteindelijke behandelresultaten. Als tweede is de variabiliteit in tumorcontouren geëvalueerd in een kleine dataset, waarin 20 tumoren een zeer sterke volumeafname in korte tijd laten zien en 20 tumoren waarbij de behandeling is mislukt. In deze gelimiteerde dataset is bepaald dat de variatie in tumor-annotaties groter is in de groep met significante krimp vergeleken met de variatie in het cohort waarbij de behandeling is mislukt. Als derde is de impact van de heterogene dosisverdeling op de behandeluitkomsten geëvalueerd. Er is een nieuwe methode geïntroduceerd om de heterogene dosisverdeling te analyseren. Hierbij is gebruik gemaakt van histogrammen van georiënteerde gradiënten in drie dimensies, samen met SVM en principale component analyse. Het resulterende model behaalt een nauwkeurigheid van 77,5%, wat suggereert dat heterogeniteit van de dosisverdeling invloed heeft op de behandelresultaten.

Ten vijfde wordt in de Hoofdstukken 5, 6 en 7 de visuele aspecten van de tumoren op de MRI scans onderzocht en berekend door middel van kwantitatieve beeldaspecten. Deze hoofdstukken onderzoeken de mogelijkheid om de individuele behandelreactie te voorspellen aan de hand van deze *radiomics* aspecten. Deze aspecten leggen de intrinsieke tumortekstuur- en vormverschillen vast, welke de verschillen in de onderliggende tumorbiologie mogelijk reflecteren. Allereerst worden de eerste experimenten naar meting van deze verschillen beschreven in Hoofdstuk 5. In de kleine dataset van 40 tumoren laten de resultaten zien dat de vorm van de tumor geen goede parameter is voor het voorspellen van de behandelresultaten. Vervolgens worden diverse kwantitatieve kenmerken van de tumortekstuur getest op dezelfde dataset. Hierbij is ontdekt dat tweede-orde statistische waarden bepaald door middel van zogenaamde *gray-level co-occurrence matrices* (GLCMs) en *run-length matrices* geschikt zijn om verschillen in tekstuur te beschrijven. Alleen worden deze enigszins overtroffen door eerste-orde statistische eigenschappen voor tumortekstuur zoals het gemiddelde, standaardafwijking en

mediaan van de MRI waarden. De behaalde nauwkeurigheid is 85%.

Vervolgens wordt in Hoofdstuk 6 gekeken of met behulp van de kwantitatieve eigenschappen voor tekstuur, de transiënte zwelling kort na behandeling te voorspellen is. Om de onderlinge verschillen in MRI waarden beter te kunnen vergelijken wordt in dit hoofdstuk een weefselspecifieke MRI normalisatiemethode geïntroduceerd. Bovendien wordt de invloed van de onbalans in de hoeveelheid patiënten op de machine learning techniek onderzocht en wordt gekeken naar de impact van de tumorgrootte op de voorspellingsresultaten. De resultaten van deze experimenten laten zien dat een set van vier GLCM-kenmerken het beste onderscheid kan maken tussen patiënten die een transiënte zwelling laten zien en de patiënten die dit niet laten zien. Daarbovenop nemen deze resultaten toe voor grotere tumoren, en worden sensitiviteit- en specificiteitsscores behaald van respectievelijk 0,77 en 0,89. De impact van de onbalans in de data is ook zichtbaar in de resultaten. Voor modellen die getraind zijn op minder goede kenmerken voor tekstuur worden de resultaten in de richting van de meerderheid getrokken, terwijl de resultaten van de goedwerkende modellen verbeteren door de beschikbaarheid van meer data voor het trainen.

Als laatste wordt in Hoofdstuk 7 de impact van de MRI-gebaseerde karakteristieken voor tumortekstuur op de langetermijnresultaten van de behandeling in de dataset van het ETZ onderzocht. Door het toepassen van strikte en objectieve definities voor mislukte behandeling en succesvolle behandeling kunnen twee onderscheidende groepen patiënten geselecteerd worden. De experimenten op basis van karakteristieken voor tumortekstuur leiden tot een model dat uiteindelijk het mislukken of het slagen van de behandeling kan voorspellen. De resulterende modellen behalen een nauwkeurigheid, sensitiviteit en specificiteit van respectievelijk 0,77, 0,71 en 0,83. Deze resultaten zijn wederom afhankelijk van de grootte van de tumoren. Het verhogen van de volume-ondergrens zorgt ervoor dat de resultaten verbeteren. Het alleen includeren van tumoren groter dan 5 cm^3 geeft de beste voorspellingsresultaten: een nauwkeurigheid, sensitiviteit en specificiteit worden behaald van respectievelijk 0,83, 0,83 en 0,82. Deze resultaten laten duidelijk zien dat het mogelijk is om de langetermijnresultaten van de GKRS behandeling te voorspellen op basis van tumortekstuur vanuit MRI beelden.

Het werk dat is beschreven in deze thesis laat zien dat computerondersteunde methodes doktoren en patiënten kan helpen in het bepalen van de optimale behandelstrategie op een individuele basis. Het onderzoek heeft aangetoond dat variaties in de onderliggende intrinsieke tumorbiologie zeer waarschijnlijk de verschillen tussen behandelresultaten van Gamma Knife radiochirurgie kunnen verklaren. De veelbelovende resultaten behaald in dit werk geven de haalbaarheid aan van het voorspellen van de korte- en langetermijnreacties van vestibulaire schwannomen op individuele basis, gebruikmakend van MRI-gebaseerde data, zoals groeisnelheid en kenmerken voor tumortekstuur. Deze resultaten kunnen gebruikt worden voor verder onderzoek naar het ontwikkelen van een klinisch beslissingssysteem, dat door doktoren en patiënten kan worden geraadpleegd om het beste persoonlijke zorgpad te kiezen.

Contents

Chapter 1 Introduction

1.1 Background

Vestibular schwannomas (VSs) are benign, usually slow-growing brain tumors, that arise from the Schwann cells of the vestibulocochlear nerve. An illustration of the tumor and its location in the cerebellopontine angle is presented in Figure 1.1. These tumors make up 8% of the primary brain tumors diagnosed in the United States [1]. In the Netherlands, the incidence rate is increasing and it grew to 15.5 cases per million inhabitants in the last decade [2]. The main reason for this increase is thought to be the ever-improving access to high-quality diagnostic tools, such as magnetic resonance imaging (MRI).

Typical symptoms associated with these tumors are unilateral or asymmetrical hearing loss, tinnitus (ringing in the ear) and vertigo (dizziness or loss of balance). As tumors grow larger, they can interfere with other cranial nerves, such as the trigeminal nerve causing facial numbness, and the facial nerve causing facial weakness or paralysis. Further enlargement of the tumor eventually can lead to pressure on nearby critical brain structures, such as the cerebellum and the brain stem, thereby becoming life-threatening in nature.

A VS is often difficult to diagnose in the early stages, since symptoms are most likely to be subtle and develop gradually over time. Furthermore, the typical symptoms related to these tumors are also associated with other middle- and inner-ear afflictions. This causes a significant variation in diagnostic strategies across medical centers [4]. However, the gold standard to diagnose a VS is performing an MRI of the brain, followed by a radiographical analysis.

Although the diagnosis is usually indisputable, the subsequent steps in clinical practice show great diversity. An international standard concerning the management of VS tumors is lacking [4]. In the last decades, the main treatment goal for VSs has shifted from complete removal of the tumor to functional preservation of the facial, trigeminal and cochlear nerves [5]. Especially the introduction of less invasive treatment options and their reduced risks at post-treatment morbidities has led to this substantial shift [6]. Moreover, various medical centers justify a conservative “wait-and-see” approach, instead of up-front active treatment because the relatively mild symptoms usually do not improve after treatment [7]. Active treatment is only considered if the tumor appears to progress on MRI scans. How-

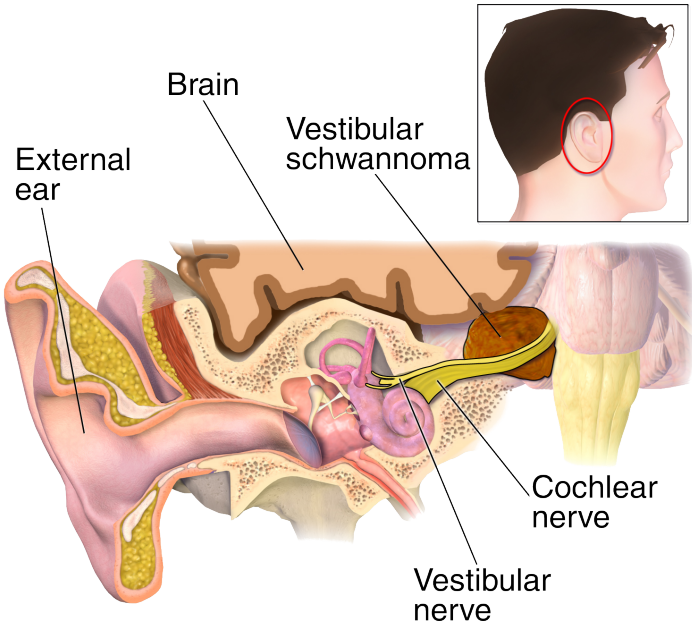


Figure 1.1 — *Illustration of a vestibular schwannoma in the cerebellopontine angle. [3]*

ever, other medical experts advocate against the wait-and-see strategy, because postponing active treatment may lead to a worsening of symptomatic tinnitus [8] and it exposes the patient to an elevated risk of hearing degradation [9].

If active treatment is selected, either at diagnosis or after progression of the tumor is observed, radiosurgery is generally preferred over microsurgery for small-to-medium-sized VSs [10]. The reason for this preference is the highly invasive nature of microsurgery, which results in (1) higher risk at mortality, (2) inferior preservation of the facial nerve function, (3) decreased hearing preservation, and (4) lower quality of life [11]–[13]. Moreover, neurosurgical treatments invoke a significantly larger overall cost on average, compared to radiosurgery [14], [15].

However, for large VSs the discussion concerning the best treatment strategy is still ongoing. Most medical centers consider microsurgical resection as the optimal treatment strategy for these large tumors, since it effectively averts the compression of surrounding critical brain structures such as the brain stem, the cerebellum and the previously mentioned cranial nerves [16]. Because the risks involved in microsurgery can be contra-indicative for this strategy, less invasive treatment strategies such as radiosurgery and radiotherapy have been considered increasingly in the last decade. These procedures have shown good results for large VSs and obtained acceptable radiation-induced morbidities [17]–[24].

Nevertheless, according to the latest guidelines on the treatment of adults with VSs, scientific evidence on which treatment option is best should be classified under Class III, i.e. *unclear clinical certainty* [25]. Furthermore, many care providers and institutions remain highly biased towards one particular therapy [26], and

some even consider that stereotactic radiosurgery can cause radiation-induced tumors or malignant transformation. However, this risk is limited and should not be used to justify the selection of an alternative treatment approach [27].

1.2 Gamma Knife radiosurgery

There are several radiation treatment modalities available that can treat VS tumors. These include Cyberknife, linear accelerator, and Gamma Knife. Of these modalities, most centers opt for Gamma Knife radiosurgery (GKRS), since it results in a 2 to 4 times lower dose to surrounding healthy brain tissue compared to other stereotactic modalities [28]. Due to this lower extratumoral dose, GKRS can spare surrounding critical brain structures, while subjecting the tumor to high radiation dose-levels. Since the start of employing GKRS for treating VS tumors, the prescribed dose has dropped significantly from 18 – 20 Gy to 12 – 14 Gy, while still demonstrating equivalent local control [29]. The main reason for this reduction in dose was the observed radiation-induced toxicities in numerous patients. The prescribed radiation dose depends mostly on the size of the tumor and the distance to the cochlea, especially when patients still have functional hearing [30], [31].

1.2.1 Gamma Knife technique

A Gamma Knife treatment device, as seen in Figure 1.2, employs γ-radiation beams of 192 individual Cobalt-60 sources. These sources are arranged in a cylindrical configuration in five concentric rings, such that they all converge into a single point. This point of convergence is called an isocenter. A 120-mm thick Tungsten collimator array ring is used for concentrating the individual beams in eight identical independent sectors. Three different collimator sizes are available: 4 mm, 8 mm, and 16 mm. Furthermore, each of the eight collimator sectors can be blocked, resulting in four collimator settings per sector. Using these 32 different collimator settings, each isocenter can be changed in shape and size. By combining multiple isocenters, a radiation-dose plan can be achieved that accurately conforms to the three-dimensional shape of a target. Due to the convergence of the 192 individual beams, there is a steep drop-off in radiation dose outside an isocenter. Hence, an accurate conformity to the tumor margin, results in a low radiation dose to surrounding tissues, while subjecting the tumor tissue to high levels of radiation. An example of a treatment plan, highlighting the steep dose gradient, is presented in Figure 1.3.

Each isocenter has a set of three Cartesian stereotactic coordinates that correspond to a three-dimensional space defined by a rigidly fixed stereotactic frame, called the Leksell G-frame (Figure 1.4). Thus, these coordinates are related to the position of the frame. To create these stereotactic coordinates, a fiducial box is placed over the frame during the MRI or computed tomography (CT) scanning of the head of a patient. This fiducial box has a Z-shaped canal at the left and right

Figure 1.2 — *Gamma Knife treatment machine, called Gamma Knife Icon. Courtesy Elekta AB, Stockholm, Sweden*

sides that is filled with copper sulfate. This creates bright marks on each MRI or CT scan, such that the planning software can determine the stereotactic coordinates of each MRI or CT volume element (voxel) with respect to the G-frame. By making use of the so-called patient positioning system, the planned isocenter coordinates can be placed in the point of convergence of the 192 radiation beams. As such, the complete treatment plan is executed fully automated.

1.2.2 Workflow

A typical workflow of a single Gamma Knife treatment is depicted in Figure 1.5. The patient comes in early for preparation by the nurse. After administering local anesthesia, the neurosurgeon mounts the G-frame to the skull of the patient. With the frame mounted, the patient goes into the MRI machine, where thin-sliced T1-weighted, T2-weighted, and T1-weighted contrast-enhanced (T1CE) MRI scans are obtained. These scans are sent to the planning system, where the neurosurgeon accurately delineates the target volume. Next, isocenters are placed within the target volume and by adjusting the size, shape, and weight of each isocenter, the target is conformed by a specific isodose line. Generally, an isodose value of approximately 50% is selected for this line, because the dose drop-off is the steepest at half the maximum dose. Using three different quality indices, namely selectivity, gradient, and Paddick conformity, the treatment plan is evaluated and further optimized. Once the neurosurgeon is satisfied with the treatment plan, it is reviewed by a radiotherapist and a medical physicist. After their approval of the plan, the patient is brought to the treatment device where the frame is fixated

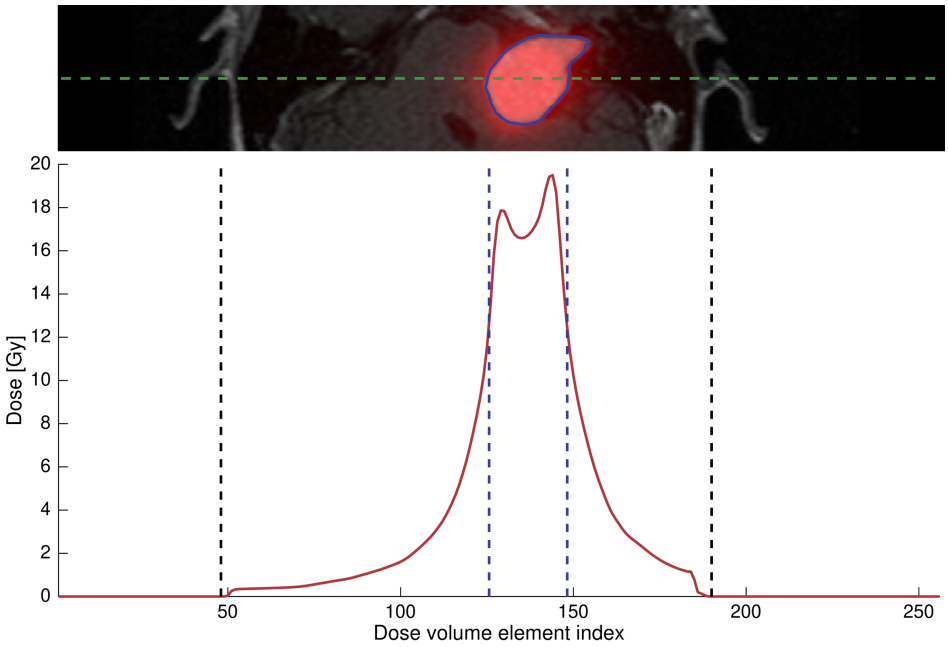


Figure 1.3 — *Example of a treatment plan. The top part of the figure displays the T1-weighted, contrast-enhanced MRI, with in red transparency, the dose distribution. The green line represents a cross-section, of which the dose levels are given in the graph at the bottom part of the figure. In this graph, the red line depicts the dose level. The dashed blue lines correspond to the tumor contour in the upper part of the image, and the dashed black lines represent the edges of the head. From this graph, it is clearly visible that the dose drops to low levels within a small distance from the tumor location.*

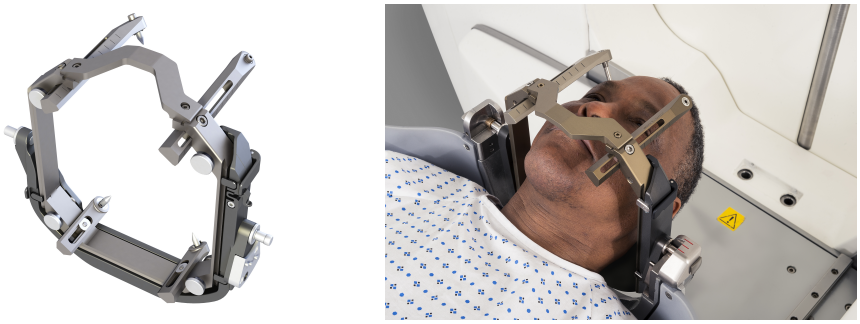


Figure 1.4 — *Leksell G-frame. Left: The Leksell G-frame as used in Gamma Knife radiosurgery. Right: The G-frame is mounted onto the head of a patient, after which it is locked in the machine during treatment. Courtesy Elekta AB, Stockholm, Sweden.*

Figure 1.5 — *Typical stage diagram of a single Gamma Knife treatment workflow.*

in the frame holder. After the treatment plan is completed, the patient is taken off the table and the frame is removed by the nurse. The patient is then discharged until follow-up visits.

1.3 Gamma Knife treatment response

GKRS is a safe and effective treatment strategy for VSs, obtaining local tumor-control rates in the range of 90% [29]. However, it has two major drawbacks. The first and foremost disadvantage involves the success of the treatment. Since the main treatment goal is to stop tumor progression, the definition of treatment success would be "no further tumor growth". However, it remains uncertain whether a VS will continue to grow, or even start regrowing several years after treatment. It has been reported that failures may occur even after 10 years following GKRS [32]. Therefore, each patient is subjected to lifelong surveillance by annual MRI scans. If continued growth or regrowth of the tumor is observed, a secondary treatment may be warranted. This can either be a re-treatment by GKRS, or microsurgical excision of the tumor.

The second drawback of stereotactic radiosurgery concerns the short-term volumetric treatment response of the tumor. During the first two to three years following treatment, a VS can suffer a radiation-induced transient tumor enlargement (TTE) [33]. This adverse effect, which is also known as pseudo-progression, is a temporary swelling of the tumor which usually subsides within the same time period. For small-to-medium-sized tumors, this is not a cause for concern, although patients can suffer from a temporary worsening of their complaints. However, when tumors are already pressing against critical brain structures, this transient enlargement can lead to significant issues that even necessitate microsurgical intervention [16].

As both of these major drawbacks can lead to microsurgical intervention, it negates the previously mentioned reasons for electing stereotactic radiosurgery over microsurgery in the first place. Moreover, microsurgical excision of tumor tissue following GKRS is considered to be more difficult than if the tumor was not irradiated. This is caused by the fact that in most cases of salvage surgery, dense adhesions between the tumor and the structures surrounding the tumor are observed [34]–[38].

Therefore, it can be concluded that GKRS is the preferred treatment modality for VS, but only if the tumor responds well to this treatment. However, this re-

sponse is not known in advance, and a reliable prediction of this response would be highly attractive. This is the main topic of research in this thesis.

1.4 The potential for predicting the Gamma Knife treatment response

Prediction of the Gamma Knife radiosurgical treatment response of vestibular schwannoma is twofold. It has both a clinical potential, as well as a technical potential. These are both discussed in the following subsections.

1.4.1 Clinical potential

Currently, it remains unknown why some tumors are unresponsive to the GKRS treatment, while others show significant volume reductions. Several studies have investigated possible factors that influence the GKRS treatment response [18], [31], [32], [39]–[59]. Numerous research papers concluded that the tumor volume is significantly correlated with treatment outcome [32], [39], [43], [47], [48], [51], [52], [60]. Therefore, many medical centers do not treat large VSs with Gamma Knife, since the chance at treatment failure is considered significantly higher, either for failed long-term tumor control or due to TTE. However, other research groups have shown that large and giant VS tumors can be safely treated using the Gamma Knife [17]–[22], [24], [61]. As such, the tumor volume is not a suitable informative parameter that can be employed for individual treatment outcome prediction.

To enable such prediction of a treatment response on an individual basis, all available patient-specific information should be considered. This information ranges from macroscopic-scale structure of the tumor to genetic profiling [62] obtained by performing a biopsy. However, in the case of a VS, this is an undesired and extremely risky procedure. One of the most frequently encountered complications of such a procedure is post-biopsy hemorrhage, which in the case of a VS tumor can cause even death due to its risky intracranial location.

Therefore, predictive parameters for radiation-treatment responses in individual VS patients have to be obtained from readily available clinical data. Examples of such data are tumor size, tumor shape, and MRI tumor characteristics. Using the tumor size at different time instances, the tumor-specific growth rate can be calculated. With this characteristic, Marston *et al.* recently determined that fast growing tumors are less likely to obtain long-term tumor control in GKRS-treated VS [49]. Furthermore, it is well known that the MRI findings in VSs are highly variable, not only for size and shape, but also in the gray-level inhomogeneity of the tissue itself. The VS tumors can appear micro- or macro-cystic [63], hemorrhagic [64], and with variable contrast-enhancement patterns after gadolinium. Some examples of different enhancement patterns of VS tumors can be found in Figure 1.6. These MRI differences reflect variations in histology, such as cell proliferation and micro-vessel density [65], [66]. Furthermore, the assumed biological effect of radiosurgery on VS tumor cells is a combination of acute inflammation and vascular occlusion [67], [68]. As such, the observed MRI differences may pro-

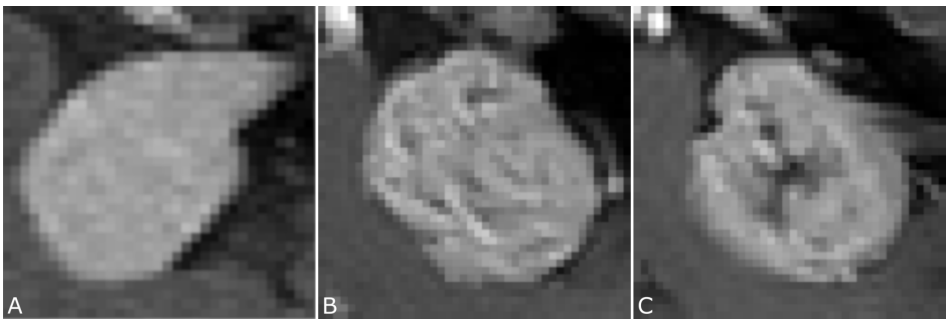


Figure 1.6 — *Examples of different appearances of vestibular schwannomas on contrast-enhanced MRI scans. Part A: Near-homogeneously enhanced lesion. Part B: Small irregularities in texture. Part C: Heterogeneously enhanced lesion with an apparent hypo-intense area.*

vide information on variations in tumor biology, thereby enabling the creation of a patient-specific tumor model that can be employed for predicting the Gamma Knife treatment response on an individual basis.

1.4.2 Technical potential

In a review by Gillies *et al.* [69], the authors conclude that medical images can provide useful data for computer-aided support (e.g. detection, diagnosis, survival prediction), using so-called radiomic features. These features, coupled with artificial intelligence, can serve as biomarkers and can be exploited to construct prediction models that have the potential to achieve significant improvements over the largely qualitative approaches currently performed in the clinical setting [70]. Employing imaging features for detection started in the 1960s, in the initial stage of computer vision [71]. The first large-scale and systematic research and development of these techniques in medicine was in the 1980s and focused mainly on computer-aided detection or diagnosis (CAD) in cancer research [72]. The initial CAD systems utilized size, shape, pixel- or voxel intensities, and textures of the regions of interest. These radiomic features can be described as handcrafted or engineered, and are used to obtain image-based phenotypes of the cancerous tissue. For example, the larger the entropy of enhanced texture, the more heterogeneous the pattern within the tumor, potentially reflecting the heterogeneous nature of angiogenesis and treatment susceptibility, serving as a location-specific "virtual digital biopsy" [73].

Many different radiomic features have been proposed in the past decades. Most of these are based on tumor shape [74]–[77], or tumor texture [78]–[85]. One widely applied texture feature extractor is based on the work by Haralick *et al.* [86]. They employed gray-level co-occurrence matrices (GLCMs) to calculate information about relative positions of pixels having similar gray-level values. Several variations of this texture feature extractor have been proposed, such as gray-level run-length matrices (RLMs), and gray-level size zone matrices (GLSZMs). Gray-level RLMs provide sizes of consecutive, collinear pixel lines with the same

intensity value for each gray-level in the image [87], whereas GLSZMs calculate the number of different zone sizes of equally valued intensity pixels [88]. Another group of radiomic features are biologically inspired features. These descriptors are based on the recognized radiological knowledge and can therefore improve the model accuracy [89], [90].

However, handcrafting features has a significant drawback. Useful features are difficult to design and often take the collective efforts of many researchers over years or even decades to optimize, and are in most cases domain- or problem-specific [91]. Nevertheless, in the last two decades, the number of publications concerning machine learning and radiomics in medicine has grown exponentially. A PubMed search on the terms "machine learning" and "radiomics" revealed that in 2019 over 10,000 articles were published.

As a possible solution, deep learning has proven in the past years that it can overcome the problems involved in handcrafting features. In the last 5 – 6 years, a shift to these featureless learning methods can be noticed. This revolution started in computer vision and is now also considered extensively in medicine. In 2010, only 9 articles appeared concerning deep learning, whereas in 2019 this has exploded to 4,470 articles, according to PubMed where "deep learning" was used as search term. Such deep learning methods no longer require feature calculation and extraction steps, but use the image itself as input to automatically determine useful features. These approaches employ multiple neural-like processing layers with several levels of abstraction [89], and have obtained significant improvements in computer vision tasks over the feature-based methods [92]. However, a major downside of deep learning is the required amount of training data, thereby limiting the implementation in studies with small datasets [70]. Furthermore, deep learning requires input data that are manually annotated by medical experts, which supervises the learning process. As such, the annotations need to be of high-level quality, since it provides the medical ground truth. To develop a large annotated dataset is therefore highly time-consuming and expensive. This aspect is another significant downside of using deep learning methods, and this explains why actual research concentrates on self-supervised learning. Nevertheless, numerous publications have shown the high performance of these types of models in medical image segmentation [93], computer-aided detection [94] and diagnosis [95], and survival prediction [96].

The above discussion highlights interesting developments that shows the potential of predicting the Gamma Knife treatment response of vestibular schwannoma. Since these tumors show highly varying enhancement patterns on MRI, it is hypothesized that these variations can be employed as radiological phenotypes, which in turn can explain the differences in the GKRS treatment response.

1.5 Challenges of developing a prediction algorithm

Currently, studies on predicting the Gamma Knife treatment response for vestibular schwannomas on an individual basis are not yet available. Therefore, the

feasibility of a prediction system has not yet been evaluated. Furthermore, medical experts are uncertain about which factors influence the individual treatment response. Several studies have reported that the tumor size at treatment has high correlation to the final treatment outcome. Furthermore, some investigations attribute the radiation dose as influencing factor, while other research groups have investigated the contributions of basic tumor appearances on MRI. Their results all highlight crucial information from a clinical point of view and are employed in calculating so-called "survival rates". These rates form a basis for determining the optimal treatment strategy for groups of patients. Although these factors definitely provide clues for a prediction system, most are expressed in a *qualitative* fashion and are evaluated at a cohort level. As a logical next step, these factors need to be transformed into *quantitative* features and need to be evaluated on an individual basis.

Availability of data: One of the serious challenges in studies involving rare pathologies is found in the number of available data samples. Having too few data points severely limits the accuracy of classifier models. A reasonable rule of thumb is, that for each included feature in a binary classification model, 10 data samples are required [69]. As such, a significant amount of data is needed for finding a binary prediction classifier. Since VS tumors are considered to be rare, this forms a major challenge for the considered research. In addition, the amount of data needed for treatment-prediction evaluation increases even further, because treatment failures are only occurring in approximately 10% of all VS patients. Furthermore, since Gamma Knife radiosurgery invokes a slow process lasting several months to years, treatment results can only be noticed after a long time period, i.e. treatment failures can occur as late as 10 years after treatment [32]. Consequently, this requires the data set to have an extended follow-up, further limiting the availability of data for creating a well-defined binary classifier.

Treatment outcome definitions: Another challenge in this work is related to the classification labeling of the data. For the development of a binary classification system, clear definitions of both output labels are needed. In the case of Gamma Knife treatment on VS tumors, this forms a significant challenge because there is no clear ground truth available. Therefore, trade-offs are unavoidable and need to be carefully made. First, the definition of treatment success is highly obscure, since the main treatment objective of GKRS is to stop tumor progression. Because it is impossible to prove that tumor progression has stopped indefinitely, concessions in creating the definition for treatment success are needed. Second, the definition of treatment failure varies significantly in literature. The employed definitions range from "an increase in diameter of more than 2 mm" [49] to "secondary treatment needed" [41], [97]. Furthermore, most medical teams rely on linear measurements in the clinical setting for determining tumor progression. This explains why subtle loss of tumor control can go unnoticed. Moreover, even if tumor progression is observed, salvage treatment may be (1) unwanted if patients are deemed unfit, or (2) are considered not necessary since the progression is considered to be small with respect to the time period in which it took place. To

enable a clear definition of treatment failure, an objective measure needs to be constructed. However, such an objective measure may have a limited clinical value and is formed to fit the scientific requirements of the technical machine learning approach.

Clinical patient-specific features: Next to the availability of data, the challenge of determining individual, patient-specific features is significant. This is especially the case when a ground truth based on genetics is unavailable. In prior research, the VS tumor size itself is an often evaluated and interesting characteristic. However, same-sized tumors may display contrasting behavior such as a difference in natural growth rate. Because the growth rates differ significantly between individual VS tumors, these may form an interesting feature for individual treatment-outcome prediction. However, calculation of the tumor growth rate is ambiguous due to the availability of various growth models, thereby further challenging the evaluation of this feature.

Treatment planning factors: Another challenge arises from the multi-faceted treatment planning. Different factors all have their contributions to the final treatment plan, as presented in Section 1.2. These factors may show a causal relation to the treatment outcome. One of the concerns with respect to the treatment planning is whether undertreating the tumor margin may be the cause of a decreased treatment efficacy. This undertreatment can originate from a lower dose to the tumor margin, but also from the variability in tumor contouring, which is performed by different neurosurgeons during the course of time. Another concern involves the high complexity of the treatment plan itself. Such a plan is created by placing multiple isocenters with varying shapes and weights, making it a highly flexible confounder that is difficult to evaluate.

Radiomic feature selection: However, the biggest challenge in evaluating radiomic features lies in the appropriate feature selection. These features can be calculated on all possible imaging modalities, thereby creating a multitude of tumor imaging aspects. These may or may not provide quantitative information related to the treatment response. Furthermore, there are many possible features, each highlighting a different imaging characteristic. Since the amount of data limits the number of features that can be included in the machine learning approach, feature selection is a crucial step. Several researchers have opted for calculating clinically inspired features, thereby including the already present and recognized clinical experience in the technical machine learning approach. However, in the field of radiosurgically treated VS tumors, medical experts are undecided in what influences the treatment response, particularly with respect to the variations in MRI tumor appearances. Therefore, it is impossible to generate clinically inspired features, which requires that different radiomic strategies have to be evaluated.

Radiomic confounders: Finally, various parameters have an influence on the comparison of quantitative image characteristics. First, to enable direct comparison of image-based features, the image data need to be normalized. However, in the case of MRI data, this is non-trivial and therefore constitutes a significant challenge in this work. Furthermore, due to the employed MRI imaging protocols, the individ-

ual voxel sizes are fixed. As such, smaller tumors may present with fewer texture details, thereby increasing the difficulty of creating suitable texture features. However, volume-based selection of the data severely impacts the number of available data points. Consequently, careful analyses are needed to evaluate the impact of the tumor volume on the radiomic features.

1.6 Problem statement and research questions

This section describes the problem statement based on the observations from the previous sections and formulates specific research questions following from this problem definition.

Problem statement

It is our objective to investigate the possibility to *a-priori* predict the Gamma Knife treatment response of vestibular schwannomas, on an individual basis. If such treatment prediction can be implemented in a clinical workflow, a personalized treatment and follow-up strategy can be selected for each individual patient. This will improve the overall results in clinical outcome, it will most likely reduce healthcare costs, and it will enhance the individual and overall quality of life for patients suffering a VS tumor. The key problem in this research is the lack of clinical knowledge concerning the reason behind treatment failures. Since there are many factors that may potentially influence the treatment response, this investigation is multi-faceted. As such, several research questions are addressed, investigating the multiple factors that potentially could have influenced the treatment response.

Research questions

From the above statement, a number of specific research questions (RQs) can be derived, which are formulated below.

RQ1: Data and treatment response measurements

Since vestibular schwannomas are a relatively rare disease, the availability of data on these types of tumors is limited. To enable treatment prediction evaluation, a significantly large database needs to be created. Furthermore, many medical centers employ various measurements and definitions for determining the treatment response of these tumors. Therefore, the following research questions can be defined, which will be discussed in Chapter 2.

- RQ1a: *Which patients and how many need to be included for determining predictive parameters of the GKRS treatment outcome?*

- RQ1b: *What are good clinical metrics for determining the different treatment outcomes?*

RQ2: Influence of the pre-treatment growth rate on long-term tumor control
Due to the shift in treatment goal from complete removal of the tumor to preservation of cranial nerve functionality, the *wait-and-see* strategy is considered an important tool in the clinical decision-making process concerning treatment options for VS tumors. This specific strategy has led to an increase in data on patient-specific information concerning tumor-size changes over time. Because variations in tumor proliferation may reflect differences in the underlying tumor biology, this temporal information may have predictive value. This leads to the following research questions, which will be addressed in Chapter 3.

- RQ2a: *Is the pre-treatment growth rate influencing the rate of volume reduction following treatment?*
- RQ2b: *How does the pre-treatment growth rate relate to the long-term tumor control?*
- RQ2c: *In what way does the adopted clinical methodology influence the obtained prediction model results?*

RQ3: Influence of the treatment planning on the treatment outcome
There are many parameters that could potentially influence the outcome. Some studies have suggested that the dose to the tumor margin has an impact on the obtained treatment results, whereas other studies did not find this relation. Nevertheless, because the dose drop-off in a Gamma Knife treatment is steep, one can argue that undertreatment of the tumor margin may lead to an increased risk of treatment failure. Furthermore, since treatment plans are evaluated on specific quality indices that are related to the tumor contouring, the inter-observer variability in tumor segmentation can also be considered as a confounding factor. Moreover, due to the huge amount of possibilities in creating a conformal treatment plan, the resulting dose distributions are highly heterogeneous. As a result of different hot- and cold-spots in such plans, tumors may respond differently after GKRS. These observations lead to the following research question, which will be addressed in Chapter 4.

- RQ3a: *Does the marginal dose influence the long-term tumor control?*
- RQ3b: *Is there an influence of the specific heterogeneous dose distribution on the long-term treatment outcome?*
- RQ3c: *How does the inter-observer tumor segmentation variability work out on the Gamma Knife treatment response?*

RQ4: Selection of informative MRI-based quantitative features for predicting the GKRS treatment response
It is hypothesized that the differences in GKRS treatment response originate from variations in the underlying intrinsic VS tumor biology. However, in the case of VSs, biopsy carries a significant risk of complications due to the surrounding

critical brain structures, making it an undesirable procedure. Therefore, predictive tumor-specific information needs to be obtained from readily available clinical data, such as MRI scans. The differences in MRI appearances of VS tumors may reflect variations in histology and may therefore provide information that enables the prediction of the GKRS treatment response. Prediction of the various treatment responses of VS tumors following Gamma Knife radiosurgery can significantly improve overall treatment success, increase individual and overall quality of life, and reduce healthcare costs. Furthermore, it can help physicians and patients in selecting the most-suited treatment option on an individual basis. However, currently it is not known what influences the various treatment outcomes and how to define treatment success. These observations lead to the following research questions, which will be addressed in Chapters 5, 6 and 7.

- RQ4a: *Can quantitative tumor shape descriptors enable the prediction of the GKRS treatment response?*
- RQ4b: *Which texture features are informative for the various treatment responses?*
- RQ4c: *What is the influence of the imbalance in data and the variations in tumor volumes on the prediction results?*
- RQ4d: *Is it possible to develop models that can predict transient tumor enlargement and the long-term treatment success, based on MRI texture features?*

1.7 Contributions

This section provides an overview of the scientific contributions presented in this thesis. These contributions can be linked to four categories, which are elaborated below.

Contributions to data and treatment outcome measurements

Since Gamma Knife treatment of VS tumors in the Netherlands is only executed in a single center, a unique large database has been constructed using the data of that institution. The database is one of the largest in the world, enabling strict inclusion criteria to obtain a dataset that is well defined. This database includes 735 patients with a median follow-up of 72 months. For the creation, we have annotated over 4,000 MRI scans to enable careful evaluation of the various treatment responses. Furthermore, for the assessment of transient tumor enlargement, we have included an additional 22 patients for the analysis. Finally, in order to calculate the pre-treatment growth rate, the pre-treatment MRI scans of 311 patients in our database have been obtained.

Furthermore, since a significant amount of patients have their follow-up scans at the same institution, accurate analyses of the volumetric treatment responses are possible. Therefore, clear and objective treatment response definitions have been constructed. First, to include the potentially missed failures due to linear measurements employed in the clinical setting, we have introduced an objective

measure for treatment failure. The definition evaluates whether there is twice a significant increase in tumor volume within three consecutive follow-up MRI scans at least two years after treatment. Second, because the employed dataset includes data of patients with a significantly long follow-up period, we have been able to carefully assess the treatment responses, thereby enabling the creation of a data-driven treatment success definition. Treatment is considered successfully executed if the tumor progression has stopped for at least 129 months after treatment. Third, by employing data of patients that had available MRI scans during the first year after treatment, we have been able to create a definition for transient tumor enlargement. For this definition, we have included MRI scans obtained at about 6 months after treatment. If the tumor volume on this MRI scan shows a significant increase with respect to the treatment volume, followed by a reduction to at least the tumor volume in the succeeding follow-up MRI scans, it is considered that this tumor presented a transient tumor enlargement.

Contributions to the evaluation of the pre-treatment growth rate influence

With the first large-scale evaluation on temporal pre-treatment data, we have investigated the influence of the tumor-specific growth rate of a VS prior to treatment, on the tumor volume response after treatment. The conducted experiments have shown that fast growing tumors are less likely to obtain long-term tumor control. A Cox-regression-based model has been developed and can be implemented in a clinical setting for calculating the risk at treatment failure. The experiments on the pre-treatment growth rate have highlighted that the clinical failure definition, as well as the so-called *volumetric* failure definition, correlate to the pre-treatment growth rate. Furthermore, we have evaluated whether the pre-treatment growth rate influences the short-term volumetric response after Gamma Knife treatment. Our experiments have shown that this so-called *bending-the-curve* effect is not present in our unique large dataset.

Contributions to the evaluation of the treatment planning influence

Due to a change in treatment planning strategy, we have evaluated whether the marginal dose influences the long-term tumor control. The conducted experiments have shown that there is no statistical significant difference between a cohort treated with less than 12 Gy compared to a cohort treated with more than 12 Gy. Furthermore, experiments on the heterogeneous character of the dose distribution have highlighted that there is a correlation to the treatment outcome. In a limited dataset, we have employed a novel strategy in measuring the heterogeneity of a Gamma Knife treatment plan, using a computer vision-based method called *three-dimensional histograms of oriented gradients*. A machine learning technique is exploited to determine whether short-term significant volume reductions can be separated from treatment failure using this measure of heterogeneity. The obtained results have shown that the calculated heterogeneity features are correlated to the treatment response in a limited dataset. Finally, due to inter-observer variations in delineation of tumor contours, a treatment plan may result in undertreating

the tumor margin. Our experiments have shown that for tumors responding with significant short-term volume reductions, the variations in contouring are significantly larger than for tumors that show a treatment failure, when comparing two different delineations created on separate moments in time.

Contributions to the evaluation of tumor-specific MR image features in correlation to the short- and long-term GKRS treatment responses

With the first evaluation on MR image features, we have investigated the discriminating properties to distinguish between VS tumors presenting a significant volume decrease and those that showed a treatment failure. To the best of our knowledge, we are the first to evaluate quantitative tumor shape descriptors on clinical data of GKRS-treated VS tumors. Our experiments have shown that shape appears to be a weak predictor of the short-term significant volume reduction. Furthermore, we have contributed with the first explorations into quantitative tumor texture features on conventional MR images to enable prediction of the treatment response in VS tumors. The machine learning experiments have achieved promising results offering high prediction value, warranting further research in the predictive value of MRI tumor texture.

Exploiting our unique, large-scale database on all treated VS patients, and the aforementioned explorations in feature selection, we have extensively investigated the possibility of creating prediction models for the clinical GKRS treatment responses. First, we have examined the short-term treatment response known as radiation-induced transient tumor enlargement. The conducted experiments show that gray-level co-occurrence matrix features lead to the best results for capturing the textural differences between two cohorts. The obtained model enables to predict whether a large VS tumor will either show a temporal swelling, or remain stable or shrink in the same time period. Additionally, due to the imbalance in the amount of tumors presenting transient enlargement and those that did not, a balancing step is introduced in the machine learning algorithm. This has helped in determining which features lead to robust prediction models, without preferring the majority class.

Second, for evaluating the long-term tumor control prediction, a definition for treatment success is introduced in this thesis. Using this definition, in conjunction with our objective measure for treatment failure, we have conducted experiments on creating a model that is able to distinguish between treatment failure and treatment success, i.e. long-term tumor control, using MRI tumor texture features. This model is trained by support vector machines on retrospective clinical MRI data. Our experiments have shown that gray-level co-occurrence matrix features are most informative for distinguishing long-term tumor control from treatment failures.

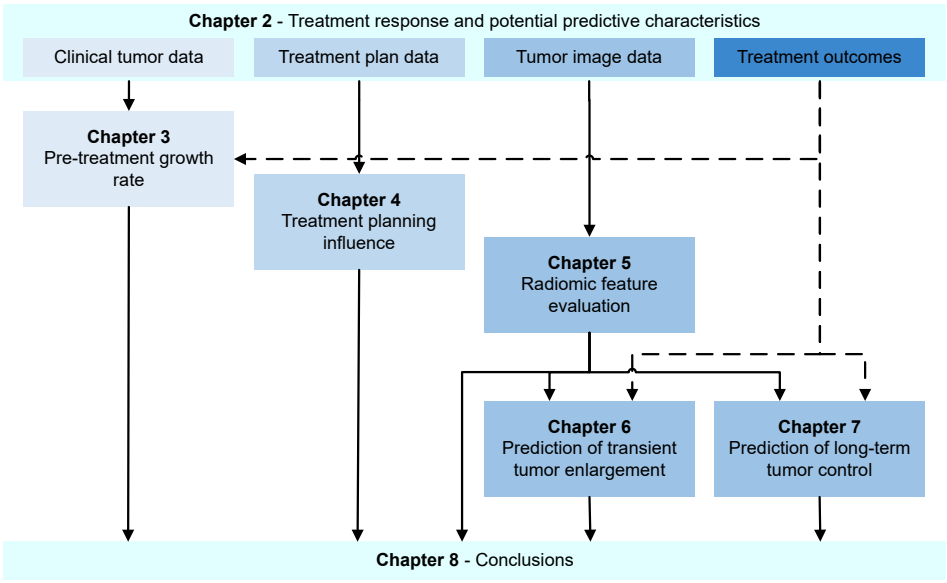


Figure 1.7 — *Schematic overview of this thesis. Solid arrows represent direct relations, while the dashed arrows indicate the link between treatment outcome definitions and the various chapters.*

1.8 Thesis outline and scientific background

This section presents an outline of the chapters in this thesis and briefly discusses the contributions of each chapter, including the underlying scientific publications. Figure 1.7 shows a schematic layout of this thesis. Chapter 2 presents a state-of-the-art overview, and introduces the unique large database and several key elements for evaluating the factors that may influence the treatment response, such as the treatment response definitions. Furthermore, it provides a technical background in the image analysis tools employed in this work. Chapter 3 investigates the influence of the pre-treatment growth rate on the volumetric response of the VS tumor to Gamma Knife treatment. The possible impact of the variations in treatment planning is discussed in Chapter 4. Chapter 5 describes the first experiments on quantitative MRI image features, such as shape descriptors and tumor texture characteristics. In Chapter 6, these experiments are extended to the prediction of the short-term volumetric treatment response. The resulting machine learning approach on short-term volumetric data is then employed for evaluating the predictability of long-term treatment success and treatment failure in Chapter 7, using the strict definitions introduced in Chapter 2. The remainder of this section introduces the focus of the individual chapters, including references to the corresponding publications.

Chapter 2 introduces the unique large Gamma Knife database of VS patients treated at the ETZ hospital in Tilburg. Furthermore, it provides an overview of the state-of-the-art in evaluating risk factors associated with the Gamma Knife

treatment response of vestibular schwannomas and it introduces the strict and objective treatment response definitions employed in this work. Moreover, since these definitions are based on tumor volume changes, the impact of inter- and intra-observer variations from multi-reader tumor annotations are discussed. Finally, this chapter introduces the background of the employed feature extraction methods and machine learning approaches.

Chapter 3 investigates the influence of the pre-treatment growth rate of vestibular schwannomas on the post-treatment volumetric tumor response. First, the strict inclusion criteria for determining the influence of the growth rate are presented. The data of the obtained patient cohorts are then analyzed for determining group differences, followed by the evaluation of the impact on short-term volumetric responses, i.e. the "bending-the-curve effect". Finally, the experiments on assessing correlations between the growth rate prior to treatment and the long-term tumor control are presented and discussed. The contributions of this chapter were presented at the Int. Stereotactic Radiosurgery Society meeting (ISRS) in 2017 [98], the winter meeting of the Dutch society for neurosurgery (NVvN) in 2018 [99], and were published in the Journal of Neurosurgery in 2019 [97].

Chapter 4 addresses the different factors that influence the creation of a treatment plan and the resulting treatment parameters. Each of these factors may have an impact on the treatment efficacy. For this purpose, the chapter explores variations in global treatment parameters, such as the prescribed dose to the tumor margin, to evaluate their impact on the long-term tumor control. Next, experiments are conducted that assess whether the complexity of a single treatment plan and its resulting heterogeneous dose distribution impacts the treatment response. To this end, an image-based feature extractor is employed for evaluating differences in the dose distribution data, and the resulting features are evaluated in a machine learning environment. Finally, the accuracy of the tumor annotations during treatment planning are evaluated for their influence on the treatment response. Contributions to this chapter were presented at the Leksell Gamma Knife Society meeting (LGKS) in 2016 [100], at SPIE Medical Imaging in 2019 [101], and were published at the Int. Symp. on Information Theory and Signal Processing in the Benelux (SITB) in 2017 [102].

Chapter 5 describes the first experiments into MRI-based tumor features and their correlation to the Gamma Knife treatment response of VS tumors. It presents the results on experiments employing tumor shape descriptors and MRI-based tumor texture features. The experiments lead to the creation of several classification models, possibly enabling the prediction of the treatment response. These resulting classification models are evaluated and a selection of best-performing models is given. The contributions to this chapter were presented at BioMedica in 2017 [103] and at SPIE Medical Imaging in 2018 [104].

Chapter 6 examines the predictive value of the MRI tumor texture features for the transient tumor enlargement effect that may occur during the first two to three years following GKRS. First, two MRI normalization techniques are discussed. Second, a method for coping with the imbalance in the data is described. Third, a machine learning algorithm is constructed based on support vector machines, to evaluate the predictability of transient tumor enlargement using the quantitative MRI image features. Finally, this chapter evaluates the impact of the tumor size on the amount of tumor texture information. The contributions to this chapter were presented at the Int. Stereotactic Radiosurgery Society meeting (ISRS) in 2019 [105], the Int. 8th Quadrennial Conf. on Vestibular Schwannoma and other CPA tumors in 2019 [106], and were published in the Medical Physics journal in 2020 [107].

Chapter 7 investigates the predictability of the long-term tumor control of VS tumors that are treated with Gamma Knife radiosurgery. First, this chapter introduces a data-driven objective definition of long-term tumor control. Second, it evaluates the effect of the imbalance present in the data and the influence of the tumor volume on the prediction results. Finally, it assesses the possibility to train a model that enables the *a-priori* prediction of long-term tumor control on an individual patient basis. The contributions to this chapter were presented at the Int. 8th Quadrennial Conf. on Vestibular Schwannoma and other CPA tumors in 2019 [106], and is accepted for publication in the journal Otology and Neurotology in 2020 [108].

Chapter 8 summarizes the most important results of this thesis and addresses the research questions formulated in Section 1.6. Finally, this chapter provides a brief outlook regarding treatment prediction algorithms for application in Gamma Knife radiosurgery for vestibular schwannomas.

Chapter 1

Chapter 2 Data and Methods

2.1 Introduction

The previous chapter has outlined the scope of this thesis by introducing the vestibular schwannoma tumor and the Gamma Knife treatment modality. Furthermore, it has emphasized the benefit of predicting the GKRS treatment response of these tumors and it has addressed the inherent challenges.

In this chapter, the basis for the research conducted in this thesis will be presented. First, the current state-of-the-art in determining predictive parameters for the Gamma Knife treatment response of vestibular schwannomas is summarized in Section 2.2. Although many papers describe several parameters that correlate to the treatment response, an algorithm for predicting the GKRS treatment response for VS tumors on an individual basis is currently not available. Nevertheless, in recent years, a variety of image-based investigations into radiation treatment responses on different tumors have been conducted. These studies are also discussed in this section.

Because the vestibular schwannoma is a relatively rare intracranial tumor, studies concerning this disease are rather limited in patient numbers. In the Netherlands, Gamma Knife treatment of VSs have been executed since the start of the Gamma Knife center at the ETZ hospital in Tilburg in 2002. Up till now, this is the only Gamma Knife center in the Netherlands that has taken care of VS patients, and therefore has been treating a large number of patients from all over the Netherlands. The data enclosed in this large patient database is the basis of this thesis and will be described in Section 2.3.

Another important aspect in determining predictive parameters for any treatment outcome, is the employed definitions of the treatment responses. In current literature, as described in Section 2.2, there is no clear consensus on the definitions for the different treatment outcomes of Gamma Knife-treated VSs. Therefore, Section 2.4 introduces novel and objective definitions of the clinically relevant treatment outcomes.

In order to evaluate the individual treatment responses, we employ the MRI data of all follow-up visits of each included patient. By determining the VS tumor volumes on each MRI, it is possible to analyze the individual treatment responses accurately and objectively. However, since the volume calculations are based on tumor annotations that are obtained manually, the inter- and intra-observer varia-

tions are critical in the treatment outcome evaluation. To assess these variations, an inter- and intra-observer study is conducted, which is presented in Section 2.5.

2.2 State of the art on influencing parameters

Vestibular schwannomas have been treated with Gamma Knife radiosurgery since the first treatment by Leksell in 1969 [109]. Since then, numerous publications have discussed this treatment modality for these types of tumors. The treatment protocol has changed significantly over the past decades, as discussed in Section 1.2, due to investigations into the efficacy of this treatment. Furthermore, many research papers discuss numerous parameters that may influence the treatment response, thereby shaping the treatment protocol to what is currently applied in the clinical care path. In current state of the art, many papers describe cohort-based investigations into parameters that influence the GKRS treatment outcome of VS tumors. These so-called risk factors can be classified into the following three main areas: patient-related, treatment-related, and tumor-related areas. The following sections discuss, per main area, several clinical investigations correlating certain risk factors to the treatment response. In Table 2.2.2, a summary of the papers is given. In this table, the implemented methodologies and the investigated risk factors and their results are highlighted per article.

2.2.1 Patient-related risk factors

Many papers describe patient-related characteristics, such as patient age and gender, that do not influence the long-term treatment outcome. However, Varughese *et al.* [42] postulated that a 10-year increase in age gave 2.18 times higher odds of tumor control. They speculate that this is caused by failing DNA repair mechanisms, leading to the VS becoming more sensitive to radiation with increasing age. Conversely, Wangerid *et al.* [45] showed a tendency in younger patients (less than 60 years old at time of treatment) for higher control rates. Nevertheless, it appears that age, as well as gender, does not influence the volumetric treatment response in these tumors.

2.2.2 Treatment-related risk factors

In the case of treatment-related characteristics, various conclusions are found in the literature. Most of the characteristics that have been investigated, include the following:

- *Radiation dose to the tumor margin*: minimum level of absorbed radiation in Gray (Gy) within the tumor margin;
- *Prescription isodose line*: planned isodose line that covers the tumor margin on which a specific dose is prescribed;
- *Radiation dose to specific percentages of the tumor volume*: minimum level of absorbed radiation in Gy for a specific percentage of the tumor volume;

- *Maximum radiation dose*: maximum level of absorbed radiation in Gy within the tumor volume;
- *Number of shots, or isocenters*: the total number of planned radiation shots per treatment;
- *Beam-on time*: total number of minutes that is spend in the Gamma Knife machine, while it was administering radiation;
- *Gradient index*: a measure for the dose drop-off at the tumor margin;
- *Selectivity*: a measure for the dose delivered to the tumor, relative to the dose delivered to normal tissue;
- *Paddick conformity index*: a measure for the conformity of the treatment plan.

Many articles that examined one or more of these parameters, concluded that the influence on the long-term tumor control is limited. As reported in Section 1.2, the prescribed dose has dropped significantly in the last decades. This resulted in considerably less radiation-induced toxicities, while maintaining equivalent long-term tumor control. Therefore, it can be hypothesized that small differences in marginal dose do not influence the long-term tumor control. Nevertheless, both Hasegawa *et al.* [39] and Lim *et al.* [57] concluded that a slightly higher dose to the tumor margin is correlated to increased long-term tumor control rates. Contrarily, others did not find this correlation. Notwithstanding, the Gamma Knife treatment modality has many settings and options, resulting in a high variation of underlying differences in treatment plans: no two plans with the same prescribed dose to the tumor margin are the same. Therefore, Millar *et al.* [110] investigated the role of the concept biologically effective dose (BED) in treatment planning. They concluded that BED calculations, taking into account the repair of sublethal damage, may indicate the importance of reporting overall treatment time, to reflect the biological effectiveness of the total physical dose applied. Due to the high complexity of calculating the BED for multi-isocenter treatment plans, reports have not yet been published that investigate the influence of the BED on the long-term treatment outcome in VS patients. Nevertheless, it is expected to observe only a limited influence on the treatment results.

	Methodology				Risk factors					
Publication	No. Ptns (failures)	FU time (months)	Tmr. size assessmt.	Definition of treatment failure	Age	Tumor size	GKRS plan	Growth (-rate)	Cystic	MRI
Hasegawa *et al.* 2005 [39]	317 (22)[1]	median: 93	diamtr	Tumor enlargement > 2 mm	-	✓	✗	-	✓	-
Arthurs *et al.* 2011 [40]	70 (4)[2]	mean: 26	diamtr	Growth > 1 mm	-	✗	✗	-	-	-
Timmer *et al.* 2011 [41]	100 (8)	mean: 27	volume	(1) Additional treatment, (2) $> 20\%$ decrease	✗	✗	✗	✗	-	-
Varughese *et al.* 2012 [42]	45 (13)	mean: 50	volume	Positive post-GKRS growth rate	✓	✗	-	✗	-	-
Hasegawa *et al.* 2013 [32]	73 (9)[1]	median: 135	diamtr	Tumor enlargement > 2 mm, or edema requiring surgery	✗	✓	✓	-	-	-
Williams *et al.* 2013 [43]	73 (6)	median: 48.5 median: 100.9[3]	volume	Second treatment	✗	✓	✗	-	-	-
Boari *et al.* 2014 [31]	379 (11)	mean: 68.3	volume	Second treatment	✗	✗	-	-	-	-
Larjani *et al.* 2014 [44]	63 (7)	median: 32	volume	No stabilization or regression > 24 months post-GKRS	✗	✗	✗	✗	-	-
Wangerid *et al.* 2014 [45]	128 (10)	median: 86	diamtr	Tumor growth > 2 mm	✗	✗	✗	✗	-	-
Frisch *et al.* 2016 [46]	40 (4)[4]	median: 64	diamtr	Tumor growth > 2 mm	-	-	-	-	✗	-
Klijn *et al.* 2016 [47]	420 (45)	median: 61	volume	Additional treatment	✗	✓	✗	-	-	-
Lee *et al.* 2016 [48]	511 (36)	median: 52	diamtr	Increase of > 2 mm in diamtr	-	✓	-	-	✓	-
Marston *et al.* 2016 [49]	68 (9)	median: 43.5	diamtr	Tumor growth > 2 mm	-	✗	-	✓	-	-
Bowden *et al.* 2017 [50]	219 (8)	mean: 50 – 58	volume	Additional treatment	-	✗	-	-	✗	-
Camargo *et al.* 2017 [51]	20 (11)[5]	median: 42.5	volume	Non-responders: $< 20\%$ decrease	-	✓	-	-	-	✓
Kim *et al.* 2017 [52]	235 (14)	median: 34	volume	Sustained increase $> 20\%$	✗	✓	✗	-	✓	-
Wu *et al.* 2017 [53]	187 (17)	median: 60.8	volume	Tumor growth $> 10\%$ at last FU	✗	✗	✗	-	✓	✓

Huang *et al.* 2018 [18]	35 (5)[1]	48	volume	Need for resection	✗	✓	✗	-	✗	-
Borghei-Razavi *et al.* 2019 [54]	48 (N.A.)	mean: 39	volume	Non-responders: $< 20\%$ decrease in volume	-	-	✗	-	-	-
Chang *et al.* 2019 [55]	62 (17)[1]	40	volume	(1) $> 20\%$ volume increase (2) Need for salvage treatment	✗	✗	-	✗	-	-
Frischer *et al.* 2019 [56]	426 (6)[1]	61	diamtr	Resection	-	✗	✗	-	-	-
Lim *et al.* 2019 [57]	24 (6)	mean: 55.8	diamtr	Increased volume between One year post GKRS and last FU	✗	✗	✓	-	✗	-
Smith *et al.* 2019 [58]	227 (N.A.)	median: 29	diamtr	(1) Increase > 2 mm after 3 years (2) Salvage therapy	✗	✗	✗	-	-	-
Speckter *et al.* 2019 [59]	23 (2)[1]	mean: 42.7	volume	Constant progression	-	-	-	-	✗	✗

Table 2.1 — *Differences in methodological definitions and results among various studies evaluating the Gamma Knife treatment outcome of vestibular schwannoma. It is evident that the number of patients (No. Ptns), the length of follow-up (FU time), the tumor-size assessment, and the treatment-failure definitions vary significantly. Risk factors were either correlated with treatment outcome (✓), not correlated (✗), or not reported on (-).*
[1]*Combination of patients primarily treated with Gamma Knife and those who had one or more prior surgical interventions for their VS.*
[2]*Including neurofibromatosis Type 2 patients and patients who had prior microsurgery.*
[3]*Different median follow-up for large VS (48.5) and small VS (100.9).*
[4]*Matched cohort study including 20 cystic and 20 solid case-matched tumors.*
[5]*Including treatments with CyberKnife and linear accelerators.*

2.2.3 Tumor-related risk factors

The most-likely source of the variations in treatment responses may be found in the tumor-specific parameters. Indeed, many articles report that tumor size is predictive for the long-term tumor control[32], [39], [43], [47], [48], [51], [52], [60]. Large tumors tend to have significantly lower tumor control rates according to these publications, although a clear cut-off is difficult to render. However, there are numerous papers that did not find this distinction [31], [40], [41], [44], [45], [49], [50], [53], [54], [56] or even concluded the opposite, where larger tumors had greater odds of tumor control [42].

Even though tumor size can be considered a tumor-specific parameter, it does not reflect differences in intrinsic tumor biology. For instance, two different VS tumors with the same size may have a contrasting growth behavior, since growth rates are highly variable in VS tumors [111]. Several clinical investigations have reported on the influence of the pre-treatment growth rate on the long-term GKRS treatment outcome. This will be further highlighted in Chapter 3. Presently, six publications have investigated the influence of the pre-treatment growth to the outcome of Gamma-Knife treatment. Four of these concluded that it has no influence, while the two other publications determined the opposite.

Another promising tumor-specific characteristic that is directly related to tumor biology is whether a tumor is cystic or not. During the last decades, it was considered that cystic VSs did not obtain a good response to radiosurgery [112]. Contrarily, recent publications have shown that there is no significant difference in long-term treatment outcome between cystic and non-cystic tumors [18], [39], [46], [50], [52], [57], [59] and that cystic VSs even present a rapid decrease in tumor volume post-GKRS [39], [46], [50], [53], [59].

2.2.4 Methodologies

From the presented results in the previous sections, it can be concluded that determining the treatment response *a priori* is highly problematic. Some resulting conclusions are not concurred by its peers, while others even deduce the opposite. It is a difficult topic, and the authors of the new guidelines published in 2018 determined that all evidence could only be classified as *unclear clinical certainty* at best [25]. The difficulty in comparing results obtained in these mostly retrospective studies, lies in the fact that many papers describe data on a limited number of patients and follow-up times. Since the reported tumor control rates of GKRS on VS lie in the range $80-100\%$, the absolute number of failures is small in cohorts of limited sizes. Furthermore, treatment failures can occur many years after treatment [32], resulting again in a limited number of failures if the follow-up time is insufficient. Therefore, the statistical power of these studies is low. Moreover, the differences between treatment failure definitions cause significant methodological inconsistencies [42], [47], [113]. These different definitions may substantially impact the number of failures per study, resulting in incomparable results that cannot be generalized over the complete VS patient population. There are numerous other issues that cause methodological imperfections, making comparison even

more difficult. Some examples are variations in: (1) tumor measurement methods, (2) growth classifications, (3) distribution of tumor sizes, and (4) classification definitions for cystic tumors.

2.2.5 Radiomics

Because the assumed biological effect of radiosurgery on VS cells is a combination of acute inflammation and vascular occlusion [67], [68], it is hypothesized that the treatment response depends mainly on the individual tumor biology. Therefore, a prediction model should incorporate biological tumor characteristics. Because tumor biopsy is not necessary for diagnosis and poses significant risks for intracranial hemorrhage, tumor-tissue information needs to be obtained from readily available clinical data. One source of biological information is through imaging techniques, such as MRI. For VS tumors, MRI images are readily available as this imaging technique is generally used for diagnosis.

Only a few studies to date have investigated the possible influence of differences in VS tumor biology on the GKRS treatment response by means of quantitative MRI tumor characteristics, i.e. radiomics. Two studies considered the apparent diffusion coefficient (ADC), calculated from MRI with diffusion-weighted imaging (DWI). The ADC value is a measure of the magnitude of water molecule diffusion within tissue, thereby reflecting differences in tumor biology. Camargo *et al.* [51] concluded that the pre-treatment ADC values of VS tumors were lower in responders than in non-responders. By employing a minimum ADC value of 800×10^{-6} m^2/s, they correctly classified 18 out of 20 patients. These results were obtained in a limited dataset containing 11 responders and 9 non-responders. Wu *et al.* [53] similarly determined that ADC values were predictive for the VS treatment response. They surmised that the maximum ADC value of patients with tumor regression or stabilization (at last follow-up) was significantly higher than in those with tumor progression. Since advanced techniques like DWI are usually not applied in clinical practice, Speckter *et al.* [59] employed a first-level texture recognition analysis for determining a prediction model. The comparison of first-order statistical texture features (e.g. mean, standard deviation, kurtosis) between the group of 14 progressors and the 9 regressors did not show significant differences. Nevertheless, they were able to obtain a sensitivity and specificity of 71% and 78%, respectively, in their complete cohort by employing a separating value based on the kurtosis of the T2-weighted MRI.

Speckter *et al.* [114] not only investigated the impact of tumor texture on the GKRS treatment outcome for VS tumors, but also for meningioma tumors. They concluded that, if only routine MRIs are available, the standard deviation of the T2-weighted MRI can be used for predicting treatment success. Radiomics are also used in creating models that can predict the radiosurgical treatment outcome of malignant brain tumors, such as brain metastases and gliomas. Tiwari *et al.* [82] were able to distinguish cerebral radionecrosis from recurrent brain tumors on multi-parametric MRI. Peng *et al.* [84] and Zhang *et al.* [83] evaluated texture features calculated on MRI sequences, to find predictive models distinguishing

radionecrosis from true tumor progression following radiosurgery on brain metastases. Wang *et al.* [85] demonstrated that multi-modality MRI imaging and radiomics analysis have potential to identify early treatment response of malignant gliomas treated with concurrent radiosurgery and bevacizumab.

These studies all highlight the importance of incorporating tumor-specific information, possibly reflecting underlying biological differences, for creating a treatment prediction model. Nevertheless, a well-designed methodology is paramount in such an investigation. It requires a large number of patients with a significantly long follow-up time, in order to create a robust and generic prediction algorithm. Furthermore, clear and objective definitions of treatment failure, transient tumor enlargement, and long-term tumor control need to be implemented. In the following sections, the patient database, the treatment outcome definitions, and the tumor measurement methods used for the work described in this thesis, will be discussed in more detail.

2.3 Patient database

Within the Gamma Knife center at the ETZ hospital in Tilburg, patients with vestibular schwannoma have been treated since its start in 2002. Up till now, more than 1,200 patients have received GKRS for their VS. This includes (1) patients that previously underwent surgical resection, either partial, sub-, or near-total, (2) patients that suffered from neurofibromatosis Type 2 (NF2), and (3) patients that received a re-treatment of their VS after GKRS treatment failure. This is a unique large database which enables extensive and in-depth research in this disease and its response to the Gamma Knife treatment.

To create a database that is optimally defined, all patients are excluded that previously underwent treatment for their VS, either (micro-)surgical or radiosurgical. Furthermore, all NF2 patients are omitted, as their conditions are considered to have significant differences in tumor biology compared to non-NF2 patients [115]. One of the signs that a patient may suffer from NF2 is presenting bilateral VS tumors. As such, only unilateral VS patients are included. Finally, since the treatment response is generally a slow process, a long follow-up time is needed for evaluating the overall treatment response. Therefore, all patients that received treatment after 2014 are excluded from the long-term treatment-response evaluation, and all patients treated after 2015 are excluded from the short-term treatment-response evaluation.

After these exclusions, a dataset of 735 patients is created for the long-term treatment-response evaluation. Table 2.2 highlights a number of aspects of this dataset, including patient- and treatment-related characteristics. Of these 735 patients, 75 (10.2%) were at some point in time lost to follow-up, 62 (8.4%) required a second treatment, and the remaining 598 (76.1%) remain under observation. Lost to follow-up is generally caused by disease-unrelated death, patient well-being, or patient preference. Of the 62 patients requiring intervention, 14 (22.6%) underwent microsurgical resection, 46 (74.2%) had a secondary GKRS treatment,

Characteristic	Median	IQR	Range
Patient age at treatment [yrs]	58	49 – 66	15 – 87
Tumor volume at treatment [mm^3]	1531	627 – 3810	14 – 18706
Pre-treatment observation time [mos]	19	14 – 30	6 – 105
Post-treatment follow-up time [mos]	72	49 – 108	0 – 199
Volume Doubling Time [mos]	15	10 – 26	3 – 344
Prescription dose [Gy]	13.0	12.5 – 13.0	8.4 – 19.4
Prescription isodose line [%]	58	48 – 64	36 – 100
Dose to 100% of tumor vol. [Gy]	11.3	11.0 – 12.0	8.8 – 13.0
Dose to 99% of tumor vol. [Gy]	11.8	11.4 – 12.6	9.5 – 14.8
Dose to 95% of tumor vol. [Gy]	12.6	12.2 – 13.6	4.5 – 15.3
Dose to 90% of tumor vol. [Gy]	13.0	12.6 – 14.2	5.9 – 16.2
Maximum dose to tumor vol. [Gy]	21.9	19.8 – 26.1	12.8 – 36.5
No. of isocenters	15	10 – 21	1 – 53
Beam-on time [mins]	42.2	31.0 – 55.6	4.9 – 132.2
Coverage [%]	91.0	89.0 – 99.0	0.0 – 100.0
Selectivity	0.89	0.82 – 0.94	0.00 – 0.99
Gradient index	2.94	2.70 – 3.34	2.44 – 9.73
Paddick conformity index	0.83	0.78 – 0.86	0.02 – 1.31

Table 2.2 — *Patient- and treatment-related characteristics, where IQR stands for inter-quartile range.*

and 2 patients (3.2%) were relieved of a growing cyst and fluid build-up. A total number of 6 neurosurgeons planned the GKRS treatment at the Gamma Knife center in the employed time-period.

The Gamma Knife center in Tilburg is a tertiary center, where patients are generally referred from other hospitals. Reason for referral is mainly tumor growth (441 patients, 60.0%). Other indications are the tumor size at diagnosis (159 patients, 21.6%), hearing preservation (84 patients, 11.4%), and preference of the patient (20 patients, 2.7%). Of the remaining 31 patients (4.2%), the reason for referral was not discernible from their medical records. Out of the 735 patients, 363 had their VS at the left side (49.4%).

To enable the evaluation of the treatment response, all follow-up MRI data are collected. From these scans, the tumor volume changes are calculated. A total number of 3,666 MRI scans are employed for determining these volume changes. Utilizing this data, it is possible to accurately determine the Gamma Knife treatment response of the included patients. This enables the creation of very strict inclusion criteria for the experiments conducted in this thesis. Furthermore, the extensive number of parameters that are included in the database allows for accurate assessment of all possible risk factors.

2.4 Treatment response definitions

In current clinical practice, tumor treatment responses are generally evaluated using medical imaging. The assessment is performed by means of linear measurements via comparing the size of the tumor on the latest post-treatment scan to its size of the pre-treatment scan or of the second-last post-treatment scan. In order to provide an objective measure for treatment responses, the Response Evaluation Criteria in Solid Tumors (RECIST) group developed a concept which is integrated in the guidelines for codification of tumor response evaluation [116]. In this concept, the response criteria are defined as follows:

- *Complete response*: disappearance of all target lesions;
- *Partial response*: at least 30% decrease in the sum of longest diameters;
- *Progressive disease*: at least 20% increase in the sum of longest diameters;
- *Stable disease*: otherwise.

In 2003, a consensus meeting on systems for reporting results in vestibular schwannoma was convened [117]. In that meeting, an agreement was made on only using linear measurements in millimeters instead of volumetric measurements. Nevertheless, in the classification of the radiotherapy treatment effect, a 10% decrease in volume (or 2 mm in longest diameter) is considered control of tumor growth. The tumor is classified as progressive if the tumor has increased in size by 2 mm or by 10% in volume.

In 2009, a revision of the RECIST criteria, called modified RECIST (mRECIST), was developed [118]. In this modification, disease progression classification now additionally requires a 5-mm absolute increase in the sum of largest diameters of the tumors, in order to cope with erroneous classification of progression when the sum of diameters is very small. Furthermore, a tumor is now considered measurable if the largest diameter is larger than 10 mm, instead of the previously adopted 20-mm threshold. At the moment of the mRECIST publication, the authors did not recommend adoption of volumetric assessments, although there was sufficient standardization and widespread availability. However, since then, several publications emphasize the need for volumetric tumor measurements in order to more accurately assess the treatment response [119]–[122]. The benefit of using volumes becomes evident when a comparison is made between the one-dimensional RECIST criteria and their volumetric counterparts: a 20% increase in largest diameter results in a 73% gain in volume, and a 30% linear reduction in a 65% decrease in volume [116]. This motivates why tumor volumes are employed in this work, to accurately assess the treatment response. Moreover, many of the treated tumors are smaller than 20 mm in longest diameter. As such, these small tumors should be considered immeasurable according to the RECIST and mRECIST criteria.

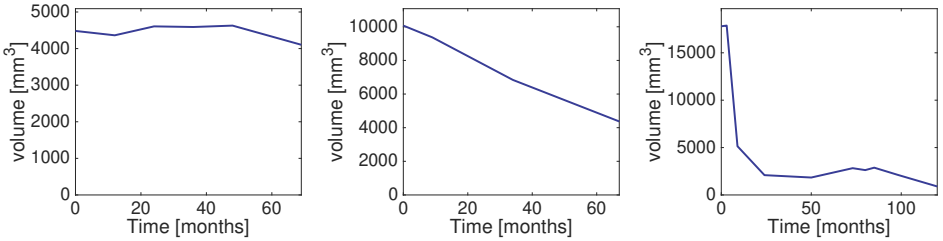
As previously described, the main difficulty in determining disease progression in the case of GKRS-treated VS tumors is that the treatment response is

generally slow and can show unpredictable behavior. For instance, multiple reports describe a radiation-induced transient tumor enlargement (TTE) that occurs in a broad range of 11 – 74% of VS patients treated with stereotactic radiosurgery, during the first 2 – 3 years after treatment [33], [52], [123]–[136]. With the availability of a large database and significant follow-up, we have been able to carefully investigate the individual treatment responses of these tumors. Indeed, from these assessments, various volumetric treatment responses have been observed. Some examples are highlighted in Figure 2.1. These graphs demonstrate that the response of a VS tumor to the GKRS treatment can change significantly throughout the follow-up duration and it is rarely linear or monotonous, making the classification of disease progression difficult at any given time-point. For example, in the case of a TTE, the increase in tumor size leads to a progressive disease classification according to the RECIST criteria. However, since TTE is a transient phenomenon and most tumors reduce to their original size after this temporary swelling, it is generally not considered as disease progression. Moreover, there are tumors that display volumetric stability following TTE. Again, RECIST classifies this as progressive disease, even though the radiosurgical treatment objective for VS tumors is to halt tumor progression, which is accomplished in these cases. As such, we consider that the RECIST criteria are not well suited for defining treatment failure and treatment success in this work.

2.4.1 Treatment failure

Many papers describing the Gamma Knife treatment response of VS tumors have been published. In these papers, the employed definitions of treatment failure vary significantly, as discussed in Section 2.2.4. Most medical centers classify the treatment as failed if intervention took place. There are two major issues with such a definition: (1) what criteria need to be satisfied prior to intervention is considered, and (2) if those criteria are met, is an intervention really desirable. Concerning the first highlighted issue, the criteria for intervention vary for each medical center and even for each patient, especially if worsening of symptoms is involved. Moreover, these criteria may change over time due to new insights. For example, the insights in the occurrence of TTE have changed the criteria for intervention [137]. Furthermore, in conjunction with the second issue, intervention may still be highly undesirable due to e.g. the fact that a patient is deemed unfit for salvage surgery, even if the criteria for intervention would be standardized.

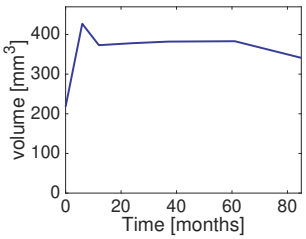
Nevertheless, the most important criterion for intervention is tumor size progression. Generally, for many medical centers, this is determined by employing linear tumor measurements. As discussed in the previous section, this is not a highly reliable and accurate method, as small progressions can be easily missed. Therefore, in addition to proven tumor progression followed by intervention, a mathematical model for determining treatment failure is employed in this work. This model simultaneously accounts for missed small progressions in the clinical setting and for undesirable interventions. The model may not be clinically rele-



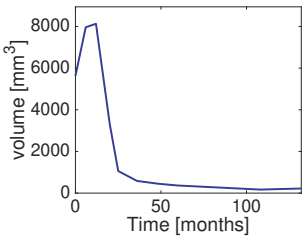
(a) *Stable response.*

(b) *Slow decay in volume.*

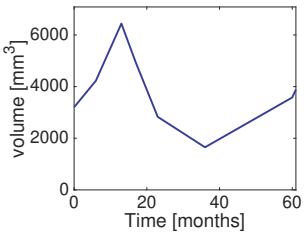
(c) *Rapid and significant volume decrease*



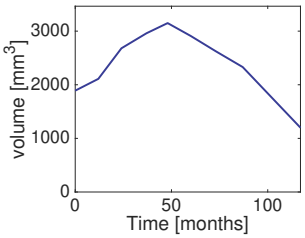
(d) *Small transient enlargement followed by stabilization.*



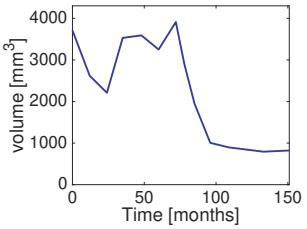
(e) *Significant transient enlargement followed by significant volume reduction.*



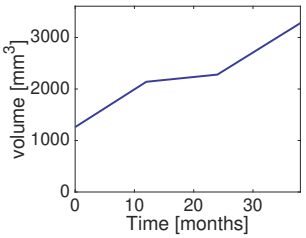
(f) *Significant transient enlargement followed by regrowth and intervention.*



(g) *Continuous growth that resolved in volume reduction four years after treatment.*



(h) *Oscillating volume changes.*



(i) *Continuous growth followed by intervention.*

Figure 2.1 — *Examples of different volumetric responses of vestibular schwannoma to the Gamma Knife treatment.*

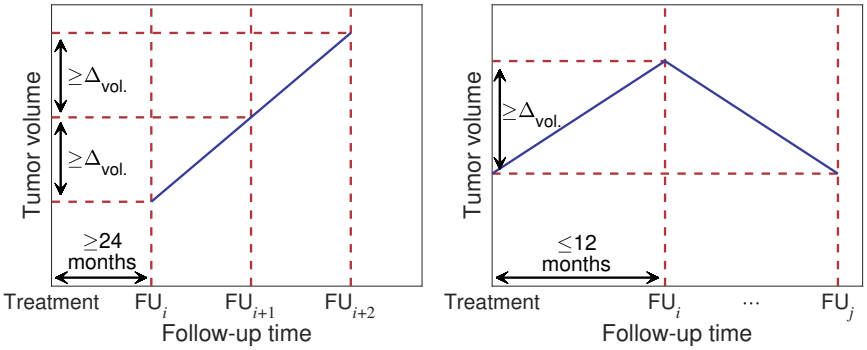


Figure 2.2 — *Graphical representation of (a) the volumetric failure definition and (b) the transient tumor enlargement definition. In these graphs, $\Delta_{vol.}$ represents the required significant volume change, and i, j are integers denoting the i-th or j-th follow-up (FU) scan, respectively.*

vant, but it provides an objective measure for determining treatment failure. This so-called *volumetric failure* is defined as two consecutive significant increases in tumor volume, among three consecutive follow-up MRI sessions. The definition of a significant volume change is based on the inter- and intra-observer variability, discussed in Section 2.5. To exclude changes in volume caused by TTE, only MRIs were used that are obtained after the first two years of follow-up. A graphical representation of this definition can be found in Figure 2.2. Furthermore, in order to avoid miss-classifying TTE occurrence as treatment failure, all patients with an intervention in the first two years following GKRS are excluded in our binary prediction evaluations.

2.4.2 Treatment success

Medical papers describing the GKRS treatment response of VS tumors employ various statistical tools for evaluating the treatment outcome. These tools measure the fraction of patients surviving for a certain amount of time after treatment. In the case of VS tumors, a non-failed treatment is considered as surviving. Therefore, a treatment success definition is not required. However, such a distinction from treatment failure is crucial in a machine learning approach for binary prediction of the treatment response. Defining treatment success is in the case of Gamma Knife-treated VS tumors highly challenging, specifically because the main treatment goal for VSs is to halt tumor progression. Therefore, treatment success may be defined as no more growth. However, it is impossible to determine whether a tumor has stopped growing permanently and will remain stable or decrease even further in size during the remainder of a patient's life. In theory, a treatment failure may occur many years after GKRS. Indeed, Hasegawa *et al.* [32] reported that only after ten years of follow-up, no more failures occurred. However, others reported the last-occurring failure at four years after GKRS [138]. Hence, treatment success cannot be defined without concessions. As a consequence, the treatment

success definition employed in this work is based on the latest occurring failure in the employed large database. This failure presented at 129 months following treatment, concurring with the data reported by Hasegawa *et al.* [32]. According to these strict criteria, 89 patients have been treated successfully in the database.

2.4.3 Transient tumor enlargement

A third relevant Gamma Knife treatment outcome is the previously mentioned phenomenon of transient tumor enlargement (TTE). This short-term treatment response is considered clinically important for large VS tumors that already exhibit contact to, or pressure on the neighboring critical brain structures, such as the brain stem. This post-radiation swelling of the tumor may cause severe, and in some cases, life-threatening adverse effects, necessitating emergency interventions to alleviate the mass effect, which further increases the risk of surgical complications. This phenomenon can occur in the first 2 – 3 years following treatment, with the peak volume between 6 – 15 months after radiation, followed by volumetric reduction [33], [49], [127], [132], [134]. For this reason, the TTE effect is defined in this work as a significant volumetric increase within the first 12 months after treatment, followed by volumetric reduction to at least the tumor volume at treatment. Again, the definition of a significant volume change was based on the inter- and intra-observer variability, as discussed in Section 2.5. Since the TTE can be resolved at 12 months following treatment, a follow-up scan around 6 months after treatment is required in order to prevent miss-classifying such a response as non-TTE. Part b of Figure 2.2 shows a graphical representation of the TTE definition.

2.5 Inter- and intra-observer variations

As previously discussed, volumetric measurements form an integral part of this work. The tumor volumes were determined on all available pre- and post-treatment MRI scans, in order to evaluate the pre-treatment growth rates and the post-treatment volumetric responses. Segmentation of the tumor was performed slice by slice with the Gamma Knife treatment software (GammaPlan® Versions 10 and 11, Elekta AB, Stockholm, Sweden). The segmentation tool in this software aids the operator by enabling a semi-automatic contouring. This method highlights and selects all voxels in an operator-selected intensity range, that are within a drawn circle on a single MRI slice. The resulting contours can be manually modified to optimize the tumor delineation. After the segmentation is completed, the software calculates the encompassed tumor volume. Because the contouring of each tumor depends on the operator, the obtained volumes may vary between operators and between contours of the same tumor by a single operator at different time-points. These operator-induced discrepancies are known as inter- and intra-observer variations, respectively.

These inter- and intra-observer variations are critical in many fields, especially in the medical field. Generally, in the medical community, a *gold standard* is established for annotations of anatomical regions in medical imaging. Nevertheless,

small differences can be crucial in automatic segmentation and detection. For instance, Van der Sommen *et al.* [139] introduced a so-called *sweet-spot* method for improving the annotations of detections and optimizing detection performance. In their paper, they utilize the intersection of all expert annotations as positive training samples and the complement of the union of all annotations as negative training samples. Another method for estimating the true segmentation is introduced by Warfield *et al.* [140]. Their method considers a collection of annotations and computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. In this thesis, it is important to know the inter- and intra-observer variations, since it influences the treatment outcome evaluation. Because we are not interested in automatic segmentation and all follow-up MRI scans are contoured by a limited number of operators, we assume that the obtained annotations are sufficient for volumetric response evaluation. Nevertheless, small variations between contours should be incorporated in the definitions for treatment responses.

Generally, for linear measurements, an increase of at least 2 mm is considered as growth by the RECIST criteria (see Section 2.4), for preventing miss-classification due to these operator-induced variations. However, this is not yet generalized for volumes. Some researchers employ a 20% cutoff [41], while others recommend to use a 10% volume threshold [117]. To determine a lower bound on volume change, an inter- and intra-observer variation experiment was designed. For the inter-observer variations, several operators evaluated the tumor volume by contouring it. These operators included two neurosurgeons with years of experience in contouring tumors on MRI, using the planning software. Two other operators were instructed on the contouring and had certain experience prior to this variability study. Furthermore, one operator who annotated most of the pre- and post-treatment volume annotations, performed the same contouring again several weeks after the first contouring, to determine the intra-observer variations.

This study is based on employing the T1-weighted, contrast-enhanced MRI scans of the first follow-up visits. The reason for selecting these scans is found in the tumor appearance, since this can vary significantly between patients. During the first year after treatment, a radiation effect that results in tumor necrosis may be present inside the tumor. This will create hypointense areas within the tumor, making the segmentation more labor-intensive because the computer-aided method cannot be applied in these cases. This enforces the operators to segment the tumor completely manually, leading to increased differences between operators, compared to employing the treatment scan. This results in an over-estimation of the observer variations.

Since the volume range in the complete dataset is large, i.e. ranging from 15 mm^3 to 18,720 mm^3, the scans employed in this analysis are not randomly chosen, but selected according to the distribution of all tumor sizes. The volume distribution of all VS tumors in the database is depicted in Figure 2.3, where it can be distilled that the distribution is highly skewed towards the left side of the

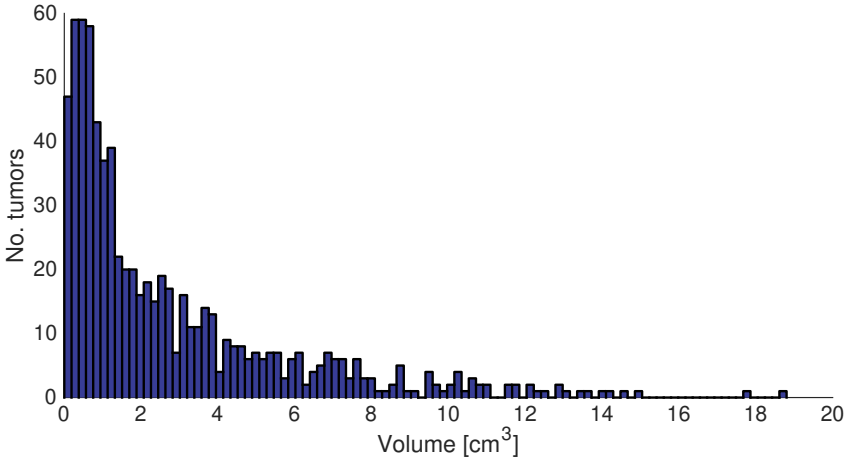


Figure 2.3 — *Histogram-based distribution of all tumor volumes in the dataset.*

volume range.

The inter- and intra-observer rates can be visualized using a Bland-Altman plot [141]. Generally, this plot is employed for displaying the absolute differences. However, in this case, the relative difference between the volumes is of interest. For doing so, the following calculations are performed:

$$\text{Mean}_\text{v} = \frac{\text{Volume}_1 + \text{Volume}_2}{2}, \tag{2.1}$$

$$\text{Relative difference} = \frac{\text{Volume}_1 - \text{Mean}_\text{v}}{\text{Mean}_\text{v}}. \tag{2.2}$$

These calculations allow the comparison of the volumes created by two different operators (*inter-observer*), or of two different annotations from the same operator (*intra-observer*). The obtained results are depicted in Figure 2.4. It can be clearly observed from the graphs that the variations depend on the tumor size. This was hypothesized, because small tumors consist of a small number of voxels, while large tumors are composed of a large number of voxels. The difference between two operators may be relatively small in the absolute number of voxels. However, the relative difference with respect to the tumor size, becomes larger for smaller tumors. Indeed, in the bottom graph of Figure 2.4 it is clear that the relative difference for smaller tumors is increased compared to the larger tumors.

Another influencing factor in annotating the tumor is the MRI image quality. The first treatment in the database dates from 2002. Since then, the scanning protocols of the MRI and the machine itself may have changed. For instance, in the earlier years, thicker-sliced MRIs were employed for the treatment planning. Furthermore, the signal-to-noise ratios have increased due to improved software and hardware. As such, the MRI image quality has improved significantly over

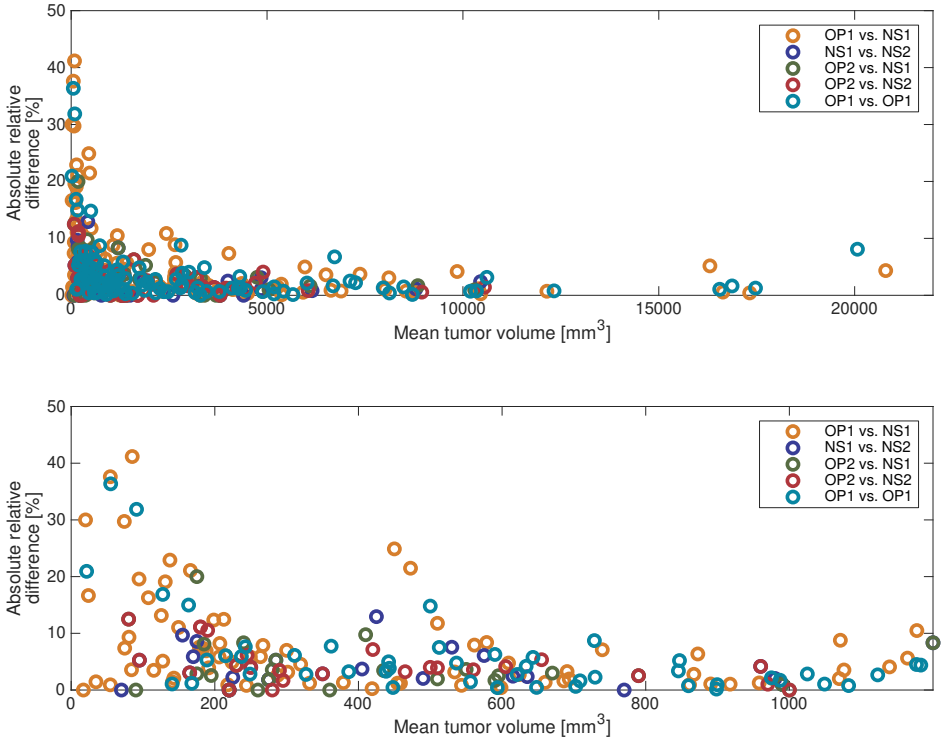


Figure 2.4 — *Results of the inter- and intra-observer variation study in volume annotations. Here, the absolute values of the relative differences are plotted against the calculated means. Terms OP1 and OP2 denote the two instructed operators with certain experience, NS1 and NS2 are the experienced neurosurgeons. Note that the top figure depicts the results over the complete volume range, while the bottom figure zooms in on the smaller tumors.*

the last decades. To analyze this impact, two operators additionally investigated the inter- and intra-observer variations in older scans compared to more recent scans. To this end, two different datasets were created. One contained images from 2002 up to 2009, while the second set contained images from 2011 up to 2014. The results of this experiment can be found in Figure 2.5. Again, smaller tumors show a larger variability than larger tumors. These results also highlight that the variability is slightly higher in the older scans, although this dependency is limited to the smaller tumors.

From the results given in Figures 2.4 and 2.5, it is extrapolated that the variability for larger tumors remains well below the 10% relative difference. Therefore, 10% is chosen to be the lower bound on volume change for larger tumors. However, for smaller tumors, this bound does not hold. At the bottom part of Figures 2.4 and 2.5, it can be distilled that for tumors smaller than 250 mm^3, the lower bound on volume change should be higher than 10%. In this study, a lower bound of 20% is selected for these tumors. Although it can be argued that this

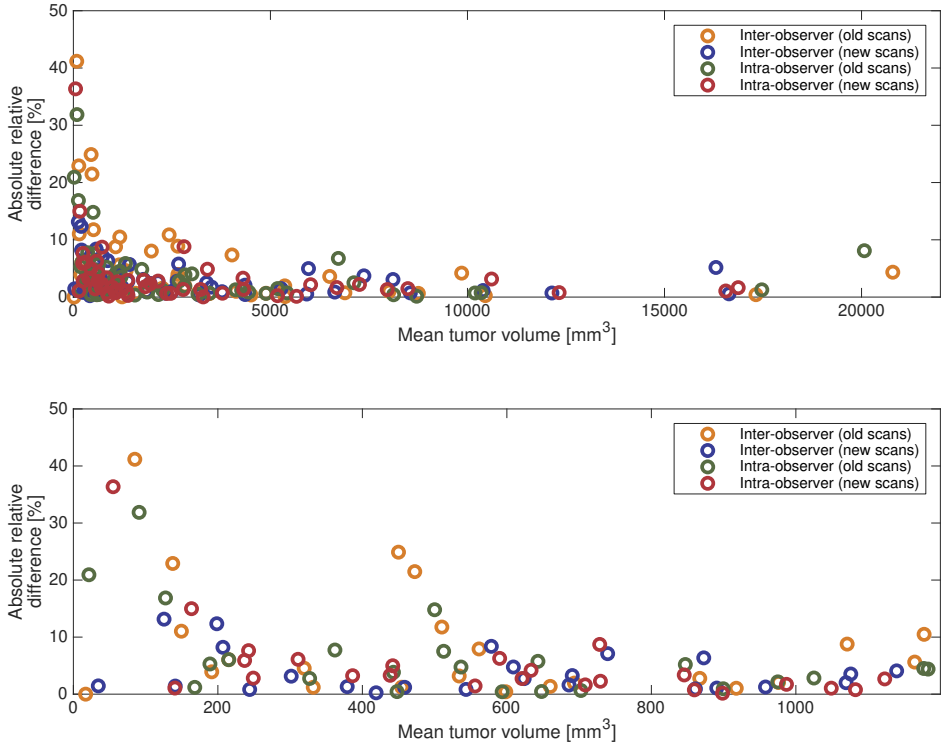


Figure 2.5 — *Plot of the inter- and intra-observer variations, comparing old and recent MRI scans. The absolute values of the relative differences are plotted against the calculated means. Note that the top figure depicts the results over the complete volume range, while the bottom figure zooms in on the smaller tumors.*

bound value is too low, the volumetric failure definition is safeguarded against misclassification, by requiring twice such an increase.

2.6 Machine learning techniques

At present, machine learning is a hot topic in research and industry, where new methods are developed continuously. It enables the discovery of models that are contained within the data, thereby facilitating the construction of many applications. Some examples are image recognition (e.g. face detection, license plate detection), automatic language translation, stock-market trading, online fraud detection, spam and malware filtering, self-driving cars, product recommendations, speech recognition, etc.

In the medical field, the applications of machine learning are based on disease detection and diagnosis, automatic segmentation of regions of interest, and prognosis prediction, and its use has increased rapidly [142]. Some examples of these applications are detection of early Barret's neoplasia [143], discriminating molecular subtypes of glioblastomas and the corresponding 12-month survival

status [81], and cancer detection in histopathology whole-slide images [144]. In recent years, the number of publications employing machine learning in clinical data has grown exponentially and the methods have increased in complexity.

Generally, there are two different machine learning strategies: feature-based and featureless. The first strategy employs calculated image characteristics as input to learn a mathematical model, which relates these characteristics to an output variable, e.g. disease classification. This strategy generally explores a hyperplane that separates the input vectors of different output classes with a high accuracy. In medical image analyses, the input characteristics are also known as *radiomic* features. These features, often based on the expertise of clinicians, are considered to describe biological properties of the tissue, thereby –at least partially– representing biological information. In a review by Gillies *et al.* [69], the authors have reported on the potential power of medical image analysis using radiomics, to facilitate improved clinical decision making. Indeed, many studies describe the ability of employing computer-aided diagnosis using medical imaging for classifying disease and treatment response [73].

The second strategy, i.e. featureless learning, employs images as inputs instead of handcrafted features, thereby skipping the feature extraction and selection steps. This method, also known as *deep learning*, is currently outperforming the previously described strategy in many fields of research, including the medical imaging field [91]. It employs several layers of parameters to capture nonlinear patterns in the data. However, for the best performance, this technique requires large-scale data input and a significant amount of computational power, since it involves self-tuning of many parameters within huge network architectures. Furthermore, due to the number of hidden layers and parameters, it is difficult to interpret the resulting deep learning models [91].

Since medical data on the rare vestibular schwannoma is inherently limited especially in combination with Gamma Knife radiosurgery, training deep learning networks from scratch is impossible. Furthermore, because this work provides the first experiments into evaluating the predictability of GKRS on VS tumors, employing feature-based methods instead of featureless methods may provide more insight in the resulting model, thereby providing a basis for more complex methods. Hence, we are going to opt for employing feature-based machine learning techniques to evaluate the predictability of the GKRS treatment response. Because the ultimate goal is to predict the treatment response, we will utilize supervised machine learning techniques in this work. Therefore, the input data needs to have a correct classification labeling, such that the resulting model can classify new unseen inputs with high accuracy, as discussed in Section 2.4.

Several supervised machine learning strategies are available for binary classification of the input data. These include a.o. logistic regression, decision trees, support vector machines, and random forests[1]. Since logistic regression employs a

[1]Random forests are typically an ensemble of decision trees, which average the results of multiple trained decision trees.

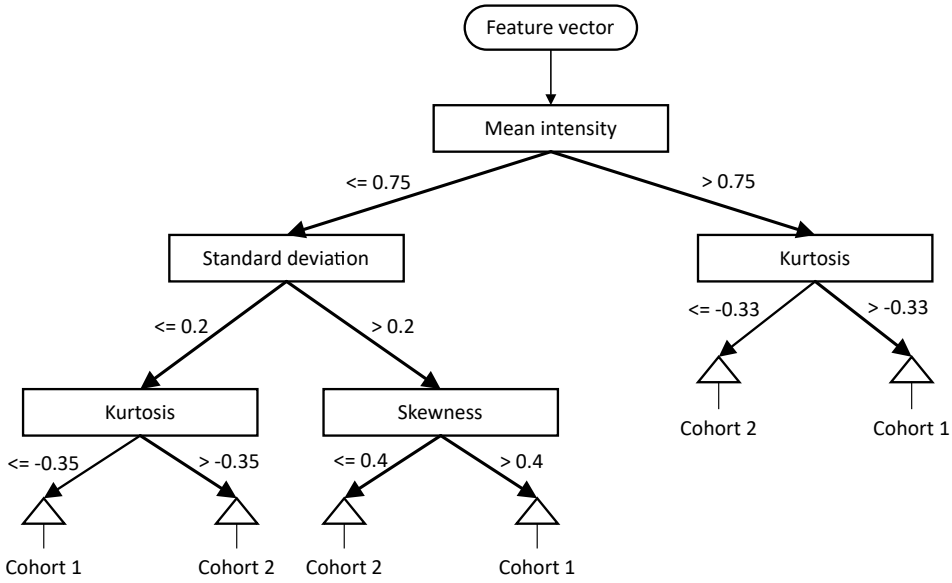


Figure 2.6 — *Graphical example of a decision tree. In this tree, the input is divided into subsets using first-order statistical features of gray-level intensities within an image, e.g. MRI. For instance, the first "if-then" node in this tree represents the question "is the mean intensity value lower or higher than 0.75?". If the value is higher than 0.75, the input progresses to the right branch, otherwise it proceeds to the left branch. This is continued until the input reaches a terminating leaf (triangle), after which the classification label is known.*

linear decision surface, i.e. the separating surface between inputs, it is considered not powerful enough for our research. Therefore, we will experiment on the data using decision trees and support vector machines, because these methods provide a higher multi-dimensional freedom in constructing the decision boundary. The choice for these two types of machine learning is based on the ease of interpretation and their successful results in medical image classification. Both methods are explained in more detail in the following subsections.

2.6.1 Decision trees (DTs)

Decision tree learning is one of the supervised machine learning methods that enables classification from observations in the data. A DT generates hypotheses that consist of multiple "if-then" statements, thereby building up a decision tree that is easy to interpret. It divides an input space, spanned by the calculated features, into multiple non-overlapping regions. A resulting tree model consists of nodes, branches, and leafs. A node represents the attribute that is tested, and a branch is the outcome of that test. The leafs are the end-points of the tree, representing the final classification label. This is graphically depicted in the example of Figure 2.6.

Training such a tree results in a division of the input data into subgroups, such that the distribution of class labels in the subgroups are as homogeneous as possible. The depth of the tree is an important parameter and can be controlled by

setting the maximum number of branch nodes, the maximum number of consecutive branches, and the minimum number of leaf-node observations. The creation of such a tree is performed by splitting the input feature vector into subsets. This process is recursively repeated on each derived subset until either a single subset at a node has all the same class labels, or when splitting no longer adds information to the classification. This process is known as top-down induction of decision trees [145]. The features that are employed for splitting the data are determined using the information gain provided by the individual features. The individual feature-specific information gain $I(A)$ is calculated by the difference between the entropy H of the training set S and the entropy of the feature A:

$$I(A) = H(S) - \sum_{v \epsilon A_{\text{vals}}} \frac{|S_v|}{|S|} H(S_v), \text{ where} \tag{2.3}$$

$$H(S) = -\sum_{i=0}^{N-1} p_i \log_2 p_i. \tag{2.4}$$

Furthermore, $H(S_v)$ is computed according to Eq. (2.4) where the input variable is the probability distribution of subset S_v. In the above equations, parameter N is the number of output classes, p_i denotes the probability of class i within the (training) set, v is a specific feature value, A_{vals} are all possible values of feature A, and S_v corresponds to the subset of S where feature A has a value v [70]. Features are then ranked by their information gain from highest to lowest and are subsequently employed for splitting the data.

2.6.2 Support vector machines (SVMs)

Another important classification method of supervised machine learning is SVMs, which is still broadly applied in many machine learning applications. This method allows for training a binary classification model by finding the best hyperplane that separates all input features of one class from the feature values of the other class. It computes a decision boundary with a maximized marginal distance to the input data, such that it can provide a robust decision boundary enabling the obtained model to tolerate noisy test data [70]. In order to determine this decision boundary, SVM methods optimize between the maximum margin and the training error. The data points that are closest to the separating hyperplane are the so-called *support vectors*. Figure 2.7 shows a graphical representation of a separating hyperplane and the corresponding support vectors. There are several methods for determining the separating hyperplane. If data is linearly separable, a hard margin can be calculated. With a normalized dataset, the parallel hyperplanes can be described by $(\vec{w} \cdot \vec{x}) - b = +1$ and $(\vec{w} \cdot \vec{x}) - b = -1$, where $\vec{x}$ represents the data point, $\vec{w}$ is the normal vector to the hyperplane, b determines the offset of the hyperplane from the origin, and -1 and $+1$ are the class labels. The margin distance equals $2/\|\vec{w}\|$, so that for the maximum distance between the two parallel hyperplanes, $\|\vec{w}\|$ needs to be minimized. Furthermore, data points

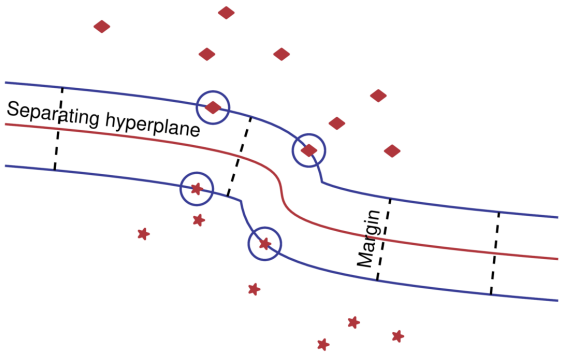


Figure 2.7 — *Graphical representation of separating hyperplane. Here, the support vectors, i.e. the data points that are on the maximum marginal distance, are highlighted by circles.*

cannot fall within the margin. Therefore, the constraint $y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\geq 1$ for all data points i is included in the optimization problem. For data that is not linearly separable, the hinge loss function is introduced, based on the expression $\max\left(0,1-y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\right)$. This results in the following minimization problem:

$$\underset{\vec{w},\lambda}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=0}^{n-1}\max\left(0,1-y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\right)+\lambda\|\vec{w}\|^2\right], \tag{2.5}$$

where parameter λ determines the trade-off between increasing the margin size and ensuring that all $\vec{x}_i$ do not lie between the two parallel hyperplanes. This problem can be rewritten as a constrained optimization problem, so that

$$\begin{aligned}&\text{minimize}\quad \frac{1}{n}\sum_{i=0}^{n-1}\max\left(0,1-y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\right)+\lambda\|\vec{w}\|^2, \\ &\text{subject to}\quad y_i((\vec{w}\cdot\vec{x}_i)-b)\geq 1-\max\left(0,1-y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\right)\ \text{ and} \\ &\qquad\qquad \max\left(0,1-y_i\left((\vec{w}\cdot\vec{x}_i)-b\right)\right)\geq 0.\end{aligned} \tag{2.6}$$

By solving the Lagrangian dual of this problem, it can be simplified to:

$$\begin{aligned}&\text{maximize}\quad f(c_0,\ldots,c_{n-1})=\sum_{i=0}^{n-1}c_i-\frac{1}{2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}y_ic_i(\vec{x}_i\cdot\vec{x}_j)y_ic_i, \\ &\text{subject to}\quad \sum_{i=0}^{n-1}c_iy_i=0,\ \text{where } 0\leq c_i\leq\frac{1}{2n\lambda}.\end{aligned} \tag{2.7}$$

Here, the variables c_i for $i=0,1,\ldots,n-1$ (where n is the number of data points) are defined such that $\vec{w}=\sum_{i=0}^{n-1}c_iy_i\vec{x}_i$. Moreover, $c_i=0$ for $\vec{x}_i$ that lie on the correct side of the margin, and $0\leq c_i\leq(2n\lambda)^{-1}$ for $\vec{x}_i$ lying on the parallel hyperplanes. This problem can be solved using quadratic programming algorithms.

In order to create nonlinear classifiers, Boser *et al.* [146] suggested to apply

kernels in the optimization problem. Some common kernels that are employed in SVM training are polynomials, Gaussians, and hyperbolic tangents. These kernels are implemented to transform feature space S to a higher dimensional space S_K, using the transform $\varphi(\vec{x}_i)$. The classification vector $\vec{w}$ is also transformed to S_K, thereby becoming $\vec{w} = \sum_{i=0}^{n-1} c_i y_i \varphi(\vec{x}_i)$. In the problem given in Eq. (2.6), the inner product $(\vec{x}_i \cdot \vec{x}_j)$ is transformed to $(\varphi(\vec{x}_i) \cdot \varphi(\vec{x}_j))$. The coefficients c_i can be solved again by using quadratic programming.

2.6.3 Validation

After the training step, each resulting model needs to be evaluated for its ability to separate the classes. This validation step can be performed utilizing several different strategies and provides an unbiased evaluation of the model fit on the training data. Generally, it calculates the following performance metrics.

- *Accuracy*: Percentage of the correctly classified data samples out of all data samples, i.e. true positives plus true negatives divided by all available samples.
- *Sensitivity*: Ratio between correctly classified positives over all actual positives, i.e. true positives divided by the sum of true positives and false negatives.
- *Specificity*: Ratio between correctly classified negatives over all actual negatives, i.e. true negatives divided by the sum of true negatives and false positives.
- *Area under the receiver operating characteristic*: An aggregate measure of performance across all possible classification thresholds. It is the probability that a classifier will rank a randomly chosen positive data sample higher than a randomly chosen negative sample (under the assumption that positives rank higher than negatives).

Here, positives and negatives represent the two classification classes. These above-described measures are chosen for their interpretability and are well-known for their ability to evaluate model performance.

One of the possible validation strategies which is often employed is k-fold cross-validation. This method first splits the data into k randomly partitioned subsets, i.e. folds. Then, the model is trained on $k-1$ folds and tested on the left-out fold. This is executed k times, after which the results of each validation are averaged over the k calculations. A schematic overview of this validation strategy is given in Figure 2.8. Such a method of validation provides an assessment of the ability of the model to generalize towards independent datasets. However, it has a significant downside, since it trains only on $k-1$ folds of the available data. This disadvantage is reduced in another method that is often implemented in model validation, namely *leave-one-out* cross validation (LOOCV). This method performs the same steps as k-fold cross-validation, only with k equal to the number of data

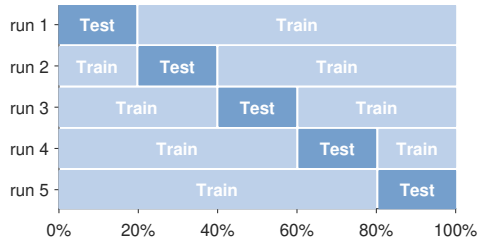


Figure 2.8 — *Graphical representation of a 5-fold cross validation strategy.*

samples. In other words, it leaves out only one sample for the training step, and then checks the classification result of the trained model on that left-out sample. This is executed for each individual data sample, after which all performance results are averaged.

Both above-described model validation methods result in models that have seen all data samples. This makes it hard to evaluate whether the final model is able to accurately classify new (unseen) data samples. To improve this, the complete dataset can be split into a training set and a test set. This validation method is known as hold-out validation, where the model is trained on the training set, possibly by including k-fold cross-validation, after which it is validated using the test set. Since the trained model has not yet seen the test data, the results obtained on these unseen data samples highlight the ability of the model to generalize on other novel datasets. However, this method reduces the number of data samples available for training even further, compared to k-fold cross-validation. In addition, this method may suffer from sampling bias because due to non-random sampling, some data points may be less likely included in training compared to other data points. This is particularly the case when the complete dataset consists of a limited number of data points. Therefore, this hold-out method requires a significant amount of training data. Indeed, it has been shown that hold-out validation is less suited for performance estimation given a finite sample amount [147]. For this reason, in this work, we will opt for LOOCV and k-fold cross-validation, where k is set to 10.

2.7 Feature extractors

Many feature extraction methods have been introduced in radiomics studies. These features may involve the shape of the regions of interest (ROIs), but also texture of the included ROIs. Some of the implemented features are based on clinical knowledge and are engineered for a specific task. Others are more general and have been implemented in other fields of research within computer vision, or even find their origin outside computer vision.

In the current research of vestibular schwannomas, it remains unclear which sets of radiomic features are related to the Gamma Knife treatment response. However, based on the supposition of the neurosurgeons who perform the GKRS

treatment of VS tumors in the Gamma Knife center in Tilburg, it is considered that contrast-enhancing tumors with inhomogeneous texture properties show different behavior than the homogeneous contrast-enhanced tumors. More specifically, inhomogeneity in the form of dark streaks and dark areas within the enhanced lesion are considered to be the most informative visual properties. Thus, a subset of radiomic features can be selected, that adequately quantify such forms of heterogeneity. These include features calculated on gray-level co-occurrence matrices (GLCMs), gray-level run-length matrices (RLMs), gray-level size zone matrices (GLSZMs), and Minkowski functionals (MFs). The following subsections describe these feature extraction methods in more detail.

2.7.1 Gray-level co-occurrence matrices (GLCMs)

One method for quantifying image texture is by employing GLCMs. These matrices describe the distribution of co-occurring pixel values at a given offset within an image. This method was introduced in 1973 by Haralick *et al.* [86], and has been exploited successfully in many fields of research [78], [79], [81], [82], [148]–[152].

Given a gray-level image I, this method computes how often pairs of pixels with a specific value and offset, under different viewing angles, occur within the image. Each element (i, j) in the resulting matrix $\mathbf{P}_{d,\theta}$ denotes the number of times that the i^{th} and j^{th} pixel values occur in the image, in the relation given by the offset d and angle θ. Furthermore, the image can be quantized in N_ℓ levels to include the ability to modify the level of detail in the image. As such, the GLCM $\mathbf{P}_{d,\theta}$ of image $I(x, y)$ quantized to N_ℓ levels can be calculated according to the following equation

$$\mathbf{P}_{d,\theta,N_\ell}(i,j) = \sum_{x=0}^{n-1} \sum_{y=0}^{m-1} \begin{cases} 1, & \text{if } I(x,y) = i \text{ and } I(x+d_x, y+d_y) = j\ , \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

In this equation, n and m denote the width and height of the image, respectively, and d_x and d_y can be calculated using offset distance d and angle θ. For instance, if $\theta = 45°$ and $d = 2$, then $d_x = 2$ and $d_y = 2$, whereas for $\theta = 0°$ this results in $d_x = 2$ and $d_y = 0$. This is further highlighted in Figure 2.9, where a basic image is employed for calculating a GLCM. For each of the resulting GLCMs, numerous statistical features can be calculated [153], which are then employed for training.

2.7.2 Gray-level run-length matrices (RLMs)

An alternative variant on the previously described GLCMs was introduced by Galloway [87] in 1975. In that work, gray-level run-length matrices were employed for terrain classification. A gray-level run is a set of consecutive, collinear pixels having the same gray-level value. The length of the run is the number of pixels contained in that run. Galloway employed statistical features calculated on the RLMs for terrain classification and obtained promising results [87]. Since it is an efficient tool for determining specific textures within a gray-level image, it has

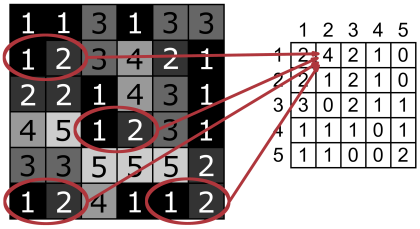


Figure 2.9 — *Graphical example representation of the calculations of the gray-level co-occurrence matrices (GLCMs). The GLCM is calculated by counting the inter-pixel relations. These relations depend on (1) the inter-pixel distance (in this example equal to unity), (2) the inter-pixel angle (in this illustration zero degrees), and (3) the number of quantization levels, which is the number of different pixel values (here equal to 5). The resulting GLCM is calculated by counting the number of each specific combination of pixel pairs. In this example, the pairs "1-2" are highlighted. The resulting value in the corresponding position of the final matrix is equal to "4", as there are 4 pairs of "1-2".*

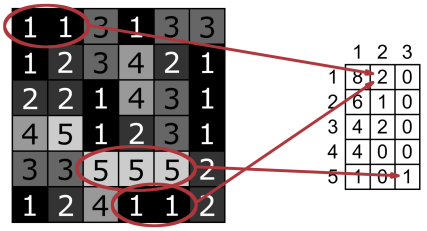


Figure 2.10 — *Graphical example representation for calculating a gray-level run-length matrix (RLM). For the RLM, the number of equally valued connected pixels is counted in a specific direction given by θ. In this example, θ equals zero degrees. The pixel values depend on the number of quantization levels and for different levels, distinct RLMs can be calculated. In this figure, some example run-lengths in the horizontal direction are highlighted. These run-lengths consist of connected pixels with values "1" and "5" of size 2 and 3, respectively. Since there are two run-lengths with pixel value "1" of size 2, the corresponding position in the resulting matrix becomes "2". The same counting can be performed for the run-length with the value "5". This results in a "1" on the corresponding position in the matrix.*

proven its application in other fields of computer vision, including medical image analysis [80], [81], [152], [154].

Calculation of a run-length matrix $\mathbf{R}$ is straightforward: each matrix element (i, j) specifies the number of times that the image contains a run of length j, in the direction given by angle θ, consisting of pixels having gray-level i. Numerous distinct RLMs can be calculated on each image by varying the viewing angle θ. Furthermore, by utilizing image quantization, various levels of detail can be incorporated in the resulting RLMs, denoted by $\mathbf{R}_{\theta,N_\ell}$. Figure 2.10 gives a visual representation of the RLM computation on the same example image employed for the GLCM calculation in Figure 2.9. For each individual RLM $\mathbf{R}_{\theta,N_\ell}$, a single feature vector is computed incorporating the above-described statistics, which is then employed for training.

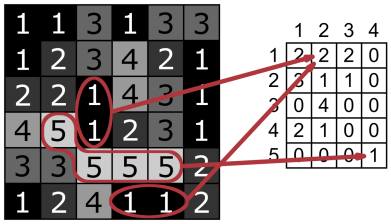


Figure 2.11 — *Graphical example representation for calculating the gray-level size zone matrices (GLSZMs). For the GLSZM, the number of zones with equally valued connected pixels and specific size is counted. The pixel values depend on the number of quantization levels, and for different levels distinct GLSZMs can be calculated. In this figure, some example zones are highlighted. These zones consist of connected pixels with the values "1" and "5" of size 2 and 4, respectively. Since there are 2 zones with pixel value "1" of size 2, the corresponding position in the resulting matrix becomes "2". The same can be done for the zone with the value "5". This results in a "1" on the corresponding position in the matrix.*

2.7.3 Gray-level size zone matrices

The previous two texture feature extractors, i.e. GLCM and RLM, only consider neighboring pixels in a specific direction in the calculation of the resulting matrices. However, in many medical cases, texture may not be limited to specific parallel structures (e.g. roads in terrain images). For example, homogeneous texture consists mostly of large areas of similar intensities, whereas heterogeneous textures can contain streaks and patches in arbitrary directions. As such, it is interesting to include neighboring pixels in multiple directions simultaneously, thereby analyzing a neighborhood of pixels. In a paper by Thibault *et al.* [88], the authors modified the RLM to incorporate the neighborhood of pixels, instead of only relying on the neighboring pixels in a specific direction. This resulted in a novel metric: the gray-level size zone matrix (GLSZM). The authors employed GLSZMs, next to GLCMs and RLMs, to determine whether the nuclei of cells have a homogeneous or a heterogeneous texture. The authors determined that GLSZMs improve the classification rate, compared to results obtained with GLCMs and RLMs. Other papers concurred with their findings, concluding that GLSZMs provide valuable texture features [150]–[152].

A GLSZM is calculated according to the RLM principle. Each matrix element (i, j) specifies the number of times that the image contains an area of size j, consisting of pixels having gray-level i. Again, this can be executed for various levels of quantization in order to influence the level of detail. A visual example of the GLSZM calculation is illustrated in Figure 2.11. For each individual GLSZM per quantization level N_ℓ, a single feature vector is calculated. This feature vector incorporates the above-described statistics. This feature vector is then employed for training.

2.7.4 Minkowski functionals

Another method for texture analysis, finally presented here, that can incorporate neighborhoods of pixels is the application of Minkowski functionals (MFs). These functionals have been defined in the field of integral geometry [155] and analyze spatial structures, by simultaneously describing the morphology and shape of regions within an image [156]. These regions are obtained by thresholding the gray-scale image, resulting in a binary image. MFs have been applied in various fields of research, such as in cosmology [157], materials [158], and even for biometry of tree positions in forests [159]. MFs are also employed in medical image analysis, where they have yielded interesting results [160]–[163].

In this thesis, the MFs are calculated based on the work by Hadwiger [164]. For a binarized image I_T, thresholded at level T, the following elementary geometric shape objects can be extracted: (1) number of white pixels, (2) sum of the boundary lengths of all white shapes, and (3) number of white shapes minus the number of encompassed black shapes within the white shapes. As an example, these computations are visualized in Figure 2.12. The geometric shape objects can be extended to three-dimensional data, such as MRI scans, where the following shape objects can be computed: (1) number of cubes (voxels) N_c, (2) number of open faces N_f, (3) number of open edges N_e, and (4) number of open vertices N_v [162]. These four objects are then employed in the calculation of the following four MFs: foreground volume M_0^T, surface area M_1^T, curvature M_2^T, and Euler number M_3^T. These functionals are specified by the following equations:

$$M_0^T = N_c \,, \tag{2.9}$$

$$M_1^T = -6N_c + 2N_f \,, \tag{2.10}$$

$$M_2^T = 3N_c - 2N_f + N_e \,, \tag{2.11}$$

$$M_3^T = -N_c + N_f - N_e + N_v \,. \tag{2.12}$$

The calculated MFs are highly scale-dependent. Therefore, the MFs need to be normalized with respect to the tumor volume. This is performed by dividing the functionals by the maximum included tumor volume.

Furthermore, by varying the threshold T, multiple distinct MFs can be calculated. Some examples of thresholded MRI slices can be found in Figure 2.13, where it becomes apparent that each threshold can cause significant differences in the shapes of objects.

2.8 Conclusions

This chapter has discussed several aspects that form the foundation of evaluating the predictability of the Gamma Knife radiosurgical outcome of vestibular schwannomas.

First, the current state of the art in this field is described. Numerous papers

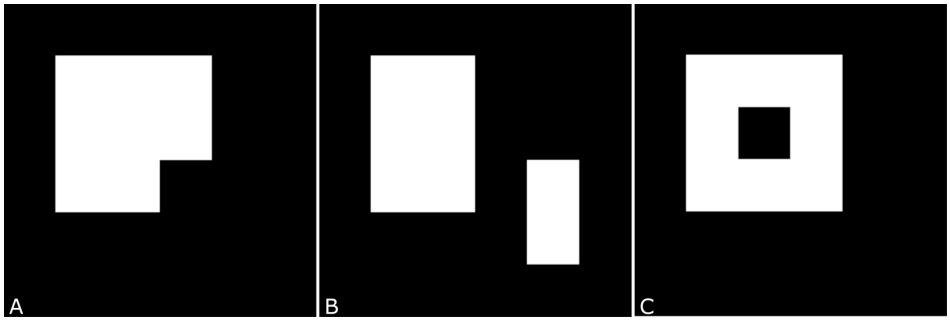


Figure 2.12 — *Visualized example for calculating the Minkowski functionals in two-dimensional images. In these basic images of 6×6 pixels, the following features are computed [164]: (1) number of white pixels K, (2) sum of all boundary lengths L, and (3) the number of white shapes minus the number of encompassed black shapes M. As such, each image is represented by three values. Part A: $K = 8$, $L = 12$, and $M = 1$. Part B: $K = 8$, $L = 16$, and $M = 2$. Part C: $K = 8$, $S = 16$, and $M = 0$.*

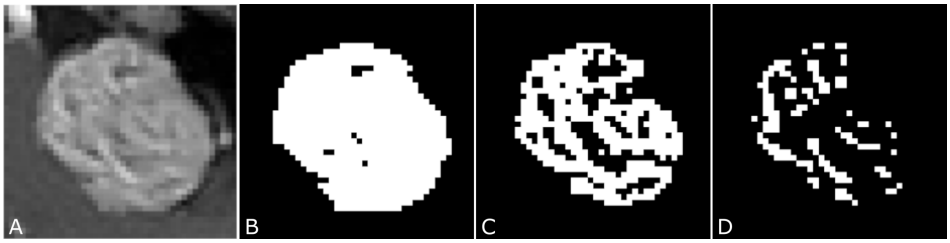


Figure 2.13 — *Visualization of the effect of imposing various thresholds on an MRI tumor image. Part A represents the original T1-weighted contrast-enhanced MRI image. Parts B, C, and D were generated with thresholds T equal to 0.5, 0.7, and 0.8, respectively, where the final image only contains pixels within the tumor segmentation.*

have assessed the impact of various risk factors on the treatment outcomes. However, the results remain inconclusive. Several researchers have distilled that the tumor size is an important factor associated with treatment failure, although other studies suggest the safe and efficient treatment of large and giant VS tumors using GKRS. Furthermore, tumor growth rate, various treatment-related parameters, and tumor appearances on MRI have been considered, but correlation to the treatment outcome is lacking due to conflicting results. In the following chapters will evaluate the previously considered risk factors on our data, and we will improve upon these parameters by introducing a novel approach for assessing the treatment planning in Chapter 4 and by employing radiomic features in Chapters 5, 6 and 7.

Second, our unique large database of all VS patients treated at the Gamma Knife center in Tilburg is introduced. Due to the substantial number of patients included in this database, it offers the unique opportunity to assess the correlation of the numerous parameters on various treatment outcomes. As such, we are able to reliably assess one highly interesting tumor-related parameter, i.e. the tumor

growth rate. The results of these experiments are presented in Chapter 3.

Third, a discussion on the treatment response definitions is presented. We have debated that objective treatment outcome definitions are required for accurate and robust GKRS treatment outcome evaluation. As such, mathematical models for defining treatment failure and transient tumor enlargement are proposed. Furthermore, a data-driven definition for long-term tumor control is introduced. These definitions form the basis of the machine learning approaches in Chapters 6 and 7, where training data are labeled accordingly.

Fourth, this chapter has evaluated the importance of inter- and intra-observer variations of the tumor annotations. Since we employ tumor volume changes over time on several thousands of MRI scans, it is paramount that the observer variations are analyzed and included in the definitions for the various treatment outcomes. After conducting several experiments, we have determined lower bounds on these variations, which have been included in the treatment response definitions.

Finally, the technical methods for machine learning and feature extraction employed in this work have been introduced. These methods form the basis for the machine learning approaches described in Chapters 4, 5, 6, and 7.

The next chapter will evaluate the impact of a clinically highly interesting parameter, i.e. tumor growth rate, on the long-term treatment outcome. Due to the unique large number of patients in our database that were followed-up prior to GKRS treatment, it is possible to robustly ascertain the influence of this tumor-specific parameter on the effect of the GKRS treatment.

Chapter 3

Pre-treatment growth rate

3.1 Introduction

The previous chapter has given an overview of the current state-of-the-art in the risk factors associated with the long-term tumor control following Gamma Knife stereotactic radiosurgery on vestibular schwannomas. It has discussed the three main areas in which these factors can be classified. From the presented results was deducted that patient-related risk factors do not influence the treatment response. Concerning the treatment-related parameters, it was determined that these have shown some correlation to the outcome in a number of papers. However, the presented results were found to be contradicting with other studies. The final category of risk factors, i.e., tumor-related characteristics, has presented some promising results. More specifically, tumor size and tumor appearance on MRI have shown correlation to the treatment response.

Furthermore, the previous chapter also highlighted a number of significant concerns in determining predictive factors associated with the volumetric treatment response of vestibular schwannomas. Methodological inconsistencies tend to blur the results, such that comparing different studies becomes difficult and conclusions cannot be generalized to other patient cohorts and Gamma Knife centers. Regardless of this, the chapter has provided interesting evaluation factors for developing a treatment prediction model.

Concerning some of the methodological inconsistencies in the current state-of-the-art, Section 2.3 has elaborated on the creation of a unique database with a large number of patients having an extensive follow-up time. A total of 735 patients that were treated at least 5 years ago, were included. Furthermore, a separate database was constructed for developing a prediction model of the short-term treatment response, specifically for predicting transient tumor enlargement. The previous chapter also discussed in Section 2.4 the development of clear and objective GKRS treatment outcome definitions for VS tumors. These definitions were based on in-house determined inter- and intra-observer variations in tumor volume measurements, and provide a solid basis for acquiring robust and generalizable outcome prediction models.

As presented in Section 2.2.3, the most promising class of risk factors consists of tumor-specific characteristics. One of the parameters in this class that has been considered previously in literature is the pre-treatment growth rate. Since

Varughese *et al.* [111] determined in their prospective study that VS tumors show highly variable growth rates, the pre-treatment growth may contain predictive information related to the treatment outcome. Nevertheless, reporting is sparse on the effect of it to stereotactic radiosurgery in patient cohorts [41], [42], [44], [49], [55], [165]. Recently, Marston *et al.* [49] indicated that fast-growing VSs are more likely to continue to grow after GKRS. However, other studies did not confirm this effect [41], [42], [44].

Again, the conflicting results of these previous studies can be possibly explained by methodological imperfections such as limited patient numbers, two-dimensional tumor measurements, insufficient follow-up times, and inconsistencies among treatment failure definitions [47], [166]. Therefore, any meaningful analysis of the effect of pre-treatment growth rates on the GKRS treatment outcome should include a sufficient number of cases in which tumor control was not achieved. The low number of failures after GKRS and their possible late occurrence implies that such studies should include large patient numbers and long follow-up times. Furthermore, Varughese *et al.* [111] determined that pre-treatment growth rates can best be modeled by employing an exponential model to calculate volume doubling times (VDTs). This implies that volumetric measurements should be obtained. In addition, these measurements are also necessary to accurately determine post-treatment tumor progression [119], [120].

The possible influence of pre-treatment growth rates on the radiosurgical outcome can be important for clinical decision-making. A pronounced effect of the pre-treatment growth rate, with faster growing tumors exhibiting lower control rates after GKRS, might be an argument for a different treatment strategy in these cases.

The objective of this chapter is to obtain insight into the efficacy of GKRS in growing VSs and to provide information on the possible effect of the pre-treatment growth rate on the treatment efficacy. As a refinement of this general problem statement, we therefore list the following aspects.

- Since the number of treatment failures is limited to approximately 10% of the cases, data from a large patient cohort with proven radiological pre-treatment progression and sufficiently long follow-up times after treatment need to be available.

- There are several models possible for calculating the growth rate prior to treatment. It is important to implement a clinically proven, accurate, and robust growth model in evaluating its impact on the treatment response.

- The treatment response of VS tumors is highly variable and it is difficult to determine an objective outcome measure. Therefore, different treatment outcomes will be evaluated. These include the short-term volumetric response, and the long-term tumor control, based on (1) clinical failures and (2) volumetric failures.

- Alternative to the different treatment outcomes, there are also numerous methods for evaluating the impact of the growth rate to the treatment outcome. We will describe these methods that will be employed for correlating the growth rate to the different outcome measures.

- The possibility to create a model that enables the treatment response prediction is evaluated. If possible, this model should be feasible for implementation in the current clinical workflow, since such a model is less complex than image processing algorithms and growth rate data are relatively easy to obtain.

This chapter is outlined as follows. First, Section 3.2 elaborates on the available growth rate models for untreated vestibular schwannomas, combined with the different treatment outcome measures that are employed. Next, the experimental setup is described for evaluating the correlation of the pre-treatment growth rates to the specific outcome measures in Section 3.3. Then, the experimental results are presented in Section 3.4, after which these are discussed in Section 3.5. Finally, Section 3.6 concludes this chapter.

3.2 Growth rate models and radiosurgical treatment outcome

This section gives a description of the pre-processing steps involved in the experiments of this chapter. First, the pre-treatment growth rate needs to be calculated using MRI scans obtained in the wait-and-scan period. Second, the post-treatment outcome measures need to be determined. The following subsections discuss these crucial elements.

3.2.1 Pre-treatment growth rate

There are several methods available for calculating the growth rate of untreated tumors. Essentially, these methods rely on either measurements of the tumor diameter, or measurements of the tumor volume. As discussed in Section 2.4, linear measurements of the tumor diameter are significantly less accurate than volumetric tumor measurements [119]–[122]. Also volume approximation methods using linear measurements are deemed undesirable, as irregular shapes are poorly suited for this [167], [168] and VS tumors have many different shapes [169]. Nevertheless, the most commonly employed method for growth-rate measurements is diameter-per-time [41], [49], [111], [170]–[175]. Other researchers describe growth rate in terms of volume-per-time [111], [176], clinical growth index [177], [178] and volume doubling time (VDT) [42], [111], [179]. From current literature, it can be concluded that there is no consensus on how to uniquely measure the growth rate of VS tumors.

In an *in-vitro* study by Sarapata *et al.* [180], several growth models were evaluated for different types of tumors. These models were all based on volumetric measurements. The resulting model-fit ranking showed that there is no "one-size fits all" growth model. In theory, an exponential model can be considered as ideal

for tumors where cells can divide without constraint and can continue to double indefinitely [181]. However, it is not appropriate for the long-term growth of solid tumors, due to limitations of the availability of nutrients, oxygen, and space [181]. Therefore, the authors performed an *in-vivo* examination for various tumors and concluded that for VS tumors, the 2/3 power law fits best. Nevertheless, the employed dataset was limited in the number of patients to 75. Furthermore, all employed tumor volumes were smaller than 3 cm^3, and volumetric measurements were performed on CT images, which is not the preferred modality for imaging of VS tumors [5].

In a prospective study by Varughese *et al.* [111], the authors compared three different growth models with respect to their fit with the actual clinical data. In their study, they evaluated whether a tumor will (1) increase by a set number of millimeters per year, (2) increase by a set number of cubic centimeters per year, or (3) double every set number of years. The first two models consider that the tumor increases linearly, while the third model assumes the growth as an exponential development. Their results, based on a cohort of 178 patients, showed that the tumor growth data fitted the best with an exponential volumetric model [111].

Therefore, in this chapter, the exponential volume doubling time (VDT) model proposed by Varughese *et al.* is employed. This exponential model implies that tumor volumes double every set number of months and is calculated using the following specification:

$$VDT = \log_{10}(2) \cdot \frac{T_{\text{treatment}} - T_{\text{pre-treatment}}}{\log_{10}(V_{\text{treatment}}) - \log_{10}(V_{\text{pre-treatment}})}. \tag{3.1}$$

In this equation, $T_{\text{treatment}}$ and $T_{\text{pre-treatment}}$ represent the dates of the treatment and pre-treatment MRI sessions, respectively. Parameters $V_{\text{treatment}}$ and $V_{\text{pre-treatment}}$ are the corresponding volumes. The pre-treatment tumor volumes and the treatment tumor volumes were determined using GammaPlan (Versions 10 and 11, Elekta AB, Stockholm, Sweden) on T1-weighted contrast-enhanced (T1CE) MR images. If such MR images were unavailable or of poor quality for the pre-treatment scan, instead thin-slice T2-weighted MR images were used. Furthermore, the pre-treatment MRI should have been obtained at least 6 months prior to treatment, in order to avoid the impact of short observation periods which would result in misleadingly large growth rates [111]. As an example of a VDT calculation, consider a tumor volume of 1 cm^3 one year before treatment and 2 cm^3 at treatment. The result is a VDT of 12 months.

3.2.2 Radiosurgical treatment and outcome

Gamma Knife radiosurgery was performed using either Leksell Gamma Knife model 4C or Perfexion (since November 2008; both Elekta AB, Stockholm, Sweden). A dose of 13 Gy was prescribed to the isodose line covering 90% (until May 2011) or 99% (since May 2011) of the tumor volume. Treatment data such as beam-on times, number of isocenters, dose to 99% of the tumor volume, and

Paddick conformity indices were collected from the treatment system.

After treatment, follow-up imaging was performed within a standard interval of one year using T1CE MRI with a slice thickness of 1 mm. The follow-up interval was shortened in case of suspected radiological progression or new or worsening symptoms. If the tumor displayed radiological regression or stability for several years, the follow-up interval was extended. Volumetric tumor measurements were performed on all follow-up scans using GammaPlan (Versions 10 and 11).

All patient records and volumetric tumor responses are reviewed to assess treatment failure. Loss of tumor control was always confirmed by the radiosurgical team, based on linear measurements. During the first two years after treatment, an increase in tumor volume is generally accepted and considered as transient tumor enlargement, unless tumor expansion is deemed too excessive. Tumor growth after this period is considered as failure. In addition, we look for potentially missed failures, i.e. discrete volume increases that are undetected by linear measurements employed in the clinical setting, but detected with the volumetric analysis performed for this research. For this purpose, the volumetric failure definition introduced in Section 2.4.1 is employed.

3.3 Experimental setup

In order to evaluate the predictive value of the VDT on the Gamma Knife treatment response of VS tumors, a number of experiments is conducted. These experiments involve the assessment of the VDT impact on the treatment response using two different approaches. First, in Section 3.3.1, we evaluate whether the GKRS treatment results in a so-called ”bending-the-curve” effect, by assessing the correlation of the pre-treatment growth rate to the short-term volume changes after GKRS. Second, the effect of the pre-treatment growth rate on the long-term tumor control is evaluated, which is discussed in Section 3.3.2. All statistical analyses in this chapter are performed using IBM SPSS statistics for Windows (Version 23, IBM Corp., USA).

3.3.1 Bending-the-growth curve

A recent publication by Fu *et al.* [182] discussed the efficacy of retreating VS patients with the Gamma Knife, after failure of their first GKRS treatment. They conclude that retreatment is an effective strategy in terms of tumor control. This, together with the hypothesis that fast growing VS tumors respond less well to GKRS, raises the question if this treatment slows down the growth rate of these tumors [166]. The presence of this so-called ”bending-the-growth” effect would therefore be very interesting for clinicians. To evaluate this effect, the correlation between the pre-treatment growth rate and the volume changes after GKRS are assessed. Since the volumetric response to GKRS treatment is non-uniform, as presented in Section 2.4, these volume changes are calculated by the (1) relative volume changes with respect to the treatment volume, and (2) the inverse of the VDT, i.e. the volume halving time (VHT). These volume changes are calculated

using the 6-month, 1-year, 2-year, and 3-year follow-up MRIs. The correlations between the pre-treatment growth rate (calculated as the VDT) and the post-treatment volume changes are analyzed with the Spearman's rank correlation coefficient. This method is selected because it assumes a monotonic relationship between the two variables, thus not limited to a linear relation, and it does not require the variables to be uniformly distributed. Since it correlates the ranking instead of the actual values of the variables, it is highly robust to strong outliers. Furthermore, the corresponding confidence intervals are calculated by performing bootstrapping with 1000 samples, using case-resampling with replacement from the original dataset. These confidence intervals propose a range of plausible values for the correlation coefficient. Using these ranges, the hypothesis that the correlation coefficients are significantly different from zero is tested. A p-value of less than 0.05 is considered to be statistically significant.

3.3.2 Long-term tumor control

There are multiple methods available for determining whether the VDT has an impact on the response to GKRS in terms of long-term tumor control. Since patients can exhibit a treatment failure at different points in time and also have variable follow-up times, it is important to incorporate the time component. This is accomplished with so-called survival analyses. In these methods, the expected time duration until one or more events happen is examined. It attempts to find the proportion of a population which will survive past a certain time. Patients that did not experience an event up to their last follow-up are censored from analysis at their last available follow-up. In the case of GKRS-treated VS tumors, such an event can be treatment failure.

There are two distinct statistical analysis methods for survival analysis: Kaplan-Meier, and Cox proportional hazards regression. Both methods are utilized in the experiments of this section and are discussed in detail below.

Kaplan-Meier

The Kaplan-Meier method estimates the survival function $S(t)$ of a cohort. This function represents the probability that a patient survives longer than time t. This analysis can be employed in cases where the predictor variable is categorical. By stratifying a cohort into multiple groups using a quantitative predictor variable, a comparison can be made among the resulting Kaplan-Meier survival curves. This is done using the log-rank test. This test determines if the observed number of events in each cohort is significantly different from the expected number of events, at each time point. If this log-rank test obtains a p-value less than 0.05, the cohorts are significantly different in terms of survival.

In this chapter, the patient cohort can be stratified using the median VDT, creating two approximately equally sized cohorts. VDT values equal to the median VDT, are assigned to the slow-growing cohort. Next, a comparison of Kaplan-Meier curves is made between the fast growing tumors and the slow growing tumors, using the log-rank test. However, the other possible predictor variables

that could potentially influence the treatment outcome, need to be assessed in both groups and tested for statistically significant differences. These include patient age, tumor volumes at treatment, doses to 99% of the tumor volume, number of isocenters, beam-on times, and Paddick conformity indices. To this end, Mann-Whitney U tests are performed. This test is selected, since it does not assume a uniform distribution of the input variable. Furthermore, the difference between the number of failures is evaluated using the Fisher's exact test, since this method is ideal for testing differences in categorical data.

The same Kaplan-Meier analysis can be extended to multiple stratifications, evaluating whether a linear trend is present in the complete patient database. However, due to the limited number of failures, the possible number of stratifications is also limited. We therefore split the complete cohort into three groups, using the 33$^{\text{rd}}$ and 67$^{\text{th}}$ percentile values of the VDT. If a linear trend is present, it indicates a distinct relation between the VDT and the GKRS treatment outcome in terms of tumor control. Again, the other possible predictor variables that could potentially influence the treatment outcome, need to be assessed in the three groups and tested for statistically significant differences. Since Mann-Whitney U tests are limited to two groups only, Kruskal-Wallis tests are performed because this allows for more than two groups. Furthermore, the difference between the number of failures is evaluated using the Chi-square test, enabling the evaluation of multiple groups.

Cox proportional hazard regression

The second survival analysis method, i.e. Cox regression, is able to evaluate the impact of quantitative predictor variables on the survival function. Using this type of method, no arbitrary stratification is needed for the analysis. First, univariate analyses are utilized to evaluate the influence of each of the predictor variables to the treatment outcome. Apart from the VDT, these also include the aforementioned characteristics. If more than one predictor variable obtains a p-value less than 0.05, a multivariate Cox regression analysis is performed to evaluate the interaction between the multiple predictor variables and their joint impact on the survival function. The Cox regression fits a so-called hazard function $h(t)$ to the data. This function can be interpreted as the risk of failure at time t and is estimated by:

$$h(t) = h_0(t) \cdot \exp(b_1 x_1 + b_2 x_2 + \ldots + b_k x_k). \tag{3.2}$$

In this function, the coefficients $(b_1, b_2, \ldots, b_k)$ measure the impact of the predictor variables $(x_1, x_2, \ldots, x_k)$ on the survival. The function $h_0(t)$ is called the baseline hazard.

3.4 Results

Using the experimental setup described in the previous section, two distinct experiments are conducted to investigate the impact of the VDT on the treatment response of VS tumors. In this section, first a description of the employed data is

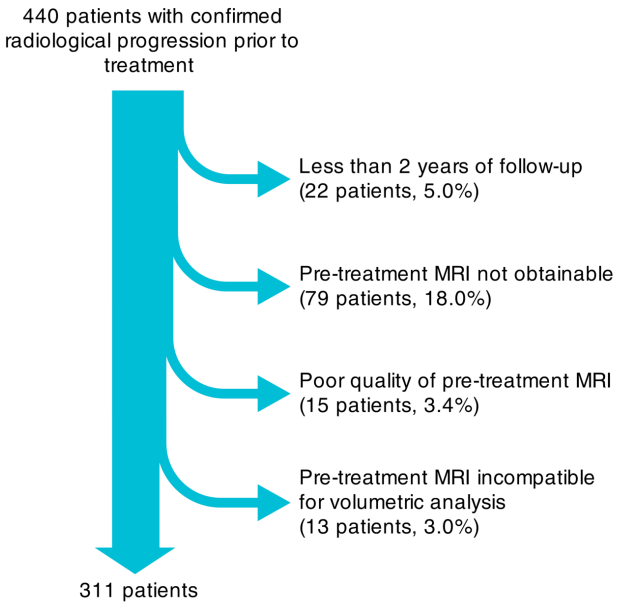


Figure 3.1 — *Patient exclusion criteria and their resulting numbers.*

given. Subsequently, the results of each experiment are presented.

3.4.1 Patient cohort

In the database from the Gamma Knife center at the ETZ hospital in Tilburg, 440 patients showed a confirmed radiological progression of the tumor prior to treatment, as described in Section 2.3. After reviewing the medical records, 129 patients (29.3%) are excluded for the experiments in this chapter (Figure 3.1). Reasons for exclusion are (1) less than two years of follow-up (22 patients, 5.0%), (2) pre-treatment MRI images not obtainable, unavailable, or less than 6 months prior to treatment (79 patients,18.0%), (3) pre-treatment MRI of poor quality (slice thickness less than 2.5 mm) (15 patients, 3.4%), or (4) pre-treatment MRI incompatible for accurate volumetric analysis in GammaPlan, due to external MRI formats such as non-square images (13 patients, 3.0%). This finally results in a patient cohort of 311 patients.

In this cohort, the median time between treatment and the available pre-treatment scan is 19 months. For the VDT calculations, we use 165 T1-weighted and 146 T2-weighted pre-treatment MR scans with a median slice thickness of 1 mm. The resulting VDT values for the cohort have a median of 15 months with a range of 3 – 344 months. Patient- and treatment-related characteristics are given in Table 3.1. The median post-treatment follow-up time is 60 months with a median time between two consecutive scans of 12 months. Lack of tumor control is observed in 35 cases (11.3%) within this cohort, resulting in the Kaplan-Meier

Characteristic	Median	IQR	Range
Patient age at treatment [years]	59	51 – 68	24 – 85
Tumor volume at treatment [cm^3]	1.16	0.62 – 2.54	0.06 – 12.18
Pre-treatment observation time [months]	19	14 – 30	6 – 105
Post-treatment follow-up time [months]	60	38 – 86	19 – 159
VDT [mos]	15	10 – 26	3 – 344
Prescription isodose line [%]	55	47 – 64	37 – 100
Dose to 99% of tumor volume [Gy]	11.9	11.5 – 13	9.5 – 13.6
No. of isocenters	13	9 – 19	1 – 43
Beam-on time [mins.]	41.5	30.9 – 54.7	8.3 – 112.0
Gradient index	2.92	2.68 – 3.30	2.47 – 6.74
Selectivity	0.87	0.81 – 0.93	0.50 – 0.99
Paddick conformity index	0.82	0.77 – 0.85	0.46 – 1.31

Table 3.1 — *Patient- and treatment-related characteristics, where IQR stands for inter-quartile range.*

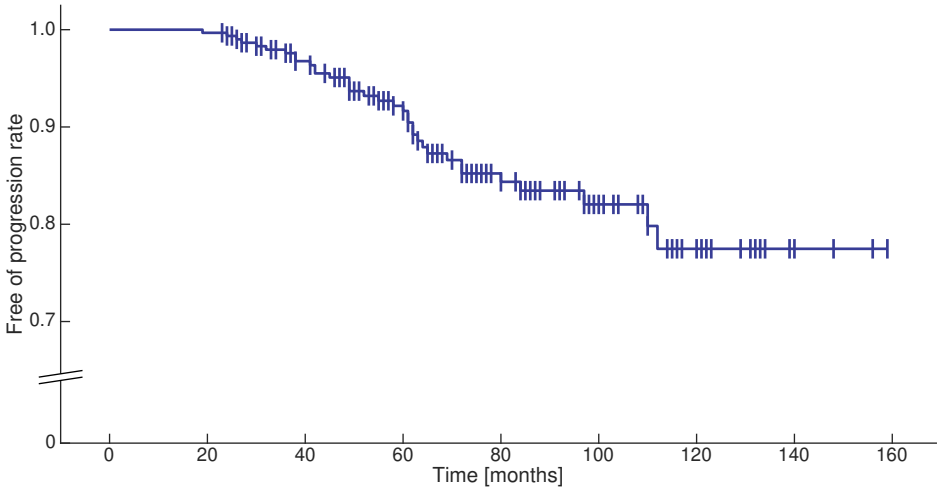


Figure 3.2 — *Kaplan-Meier curve for the complete cohort. The calculated 5- and 10-year control rates for this cohort are 91.6% and 77.5%, respectively. Tick marks indicate censored cases.*

curve depicted in Figure 3.2. One tumor exhibited obvious and excessive growth during the first two years after treatment, such that intervention was considered necessary. The calculated 5- and 10-year control rates of the cohort were 91.6% and 77.5%, respectively. Of these 35 cases, 14 were classified as a volumetric failure.

Follow-up times	No. of Patients	Median (IQR) of relative volume change [%]	p-value
6 months	54	-4.4 (-25.4 – 13.5)	0.84
1 year	282	10.6 (-11.9 – 34.3)	0.45
2 years	225	28.2 (8.3 – 43.5)	0.07
3 years	170	19.6 (3.1 – 37.5)	0.17

Table 3.2 — *Statistics of the relative volume changes after GKRS. A negative volume change denotes an increase in tumor volume. The second column represents the number of patients having an available MRI scan at the indicated time.*

Follow-up times	No. of Patients	Median (IQR) of tumor halving time [months]	p-value
6 months	54	-9.0 (-28.0 – 23.8)	0.71
1 year	282	13.0 (-29.0 – 35.0)	0.09
2 years	225	23.0 (-21.0 – 52.0)	0.55
3 years	170	31.5 (-22.5 – 62.3)	0.41

Table 3.3 — *Statistics of the tumor halving times following GKRS. A negative halving time denotes an increase in tumor volume. The second column represents the number of patients having an available MRI scan at the indicated time.*

3.4.2 Short-term volumetric response

To evaluate the relationship between pre-treatment growth rates and the short-term volumetric tumor response after GKRS, the post-treatment volume changes of each tumor are calculated based on the 6-month, 1-year, 2-year, and 3-year follow-up MR images, using two different models. The medians and inter-quartile ranges of the relative volume changes and for the tumor halving times are given in Tables 3.2 and 3.3, respectively. Correlations between these volume changes and the VDTs were determined according to the Spearman's rank correlation method. The relative volume changes were not significantly correlated with the pre-treatment growth rates, as can be seen in the last column of Table 3.2. For the tumor halving times, the same conclusion can be distilled, as all p-values are larger than 0.05 (Table 3.3).

3.4.3 Long-term tumor control

For analyzing the effect on the long-term tumor control, two different approaches are implemented. This section will present the results for both methods. First, the results of the Kaplan-Meier survival analyses are described, followed by the Cox regression analyses results. Finally, a robustness analysis will be applied to the obtained results.

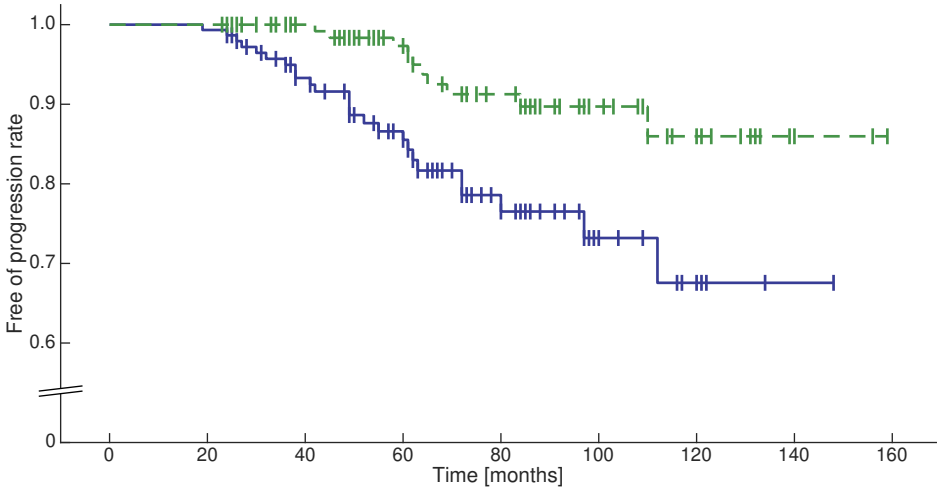


Figure 3.3 — *Kaplan-Meier curves for the fast growing cohort (blue continuous line) and slow growing cohort (green dashed line), stratified at the median of the pre-treatment growth rate. Tick marks indicate censored cases.*

A. Kaplan-Meier

For the first Kaplan-Meier survival analysis, the patient cohort is stratified into two groups, a slow growing and a fast growing tumor group, with a median VDT of 15 months as the separating value; ties ($VDT = 15$) are assigned to the slow growing cohort. This stratification results in slow growing and fast growing tumor cohorts of 162 and 149 patients, respectively. These two cohorts include 10 and 25 cases of treatment failure, respectively. This difference is statistically significant (Fisher's exact test, $p < 0.01$). The median times to loss of tumor control in both groups are 63 and 49 months for the slow growing and the fast growing cohorts, respectively. This difference is also statistically significant (Mann-Whitney U-test, $p = 0.04$), meaning that fast growing tumors tend to exhibit a failure earlier than slow growing tumors. The Kaplan-Meier curves for both cohorts are depicted in Figure 3.3. A comparison of these curves indicates a significant difference between tumor control rates of the cohorts using a log-rank test ($p < 0.01$). The calculated 5- and 10-year tumor control rates were 97.3% and 86.0% in the slow growing cohort, and 85.5% and 67.6% in the fast growing cohort, respectively. To evaluate if characteristics other than the VDT can explain this difference, we investigate possible distinctions in the main characteristics between these two groups. Statistical analyses indicate that there are no significant differences between the groups using the Mann-Whitney U test (Table 3.4).

Next, we perform the same Kaplan-Meier analysis using three groups by creating a slow growing, an average growing, and a fast growing tumor cohort. Separation is performed by splitting the total cohort according to the 33rd and 67th percentiles, where ties are assigned to the slow growing tumor cohort or the aver-

Characteristic	Median (IQR)		p-value
	Fast growing cohort	Slow growing cohort	
Patient age at treatment [yrs.]	60 (49 – 68)	59 (52 – 67)	0.86
Tumor vol. at treatment [cm^3]	1.06 (0.65 – 2.74)	1.27 (0.53 – 2.44)	0.67
Prescription isodose line [%]	55 (46 – 63)	53 (48 – 64)	0.54
Dose 100% of tumor vol. [Gy]	11.4 (11.1 – 12.2)	11.4 (11.0 – 12.0)	0.09
Dose 99% of tumor vol. [Gy]	12.0 (11.6 – 13.0)	11.9 (11.4 – 12.7)	0.11
Dose 95% of tumor vol. [Gy]	12.6 (12.4 – 13.8)	12.6 (12.3 – 13.8)	0.08
Number of isocenters	13 (10 – 19)	14 (9 – 19)	0.71
Beam-on time [mins.]	41.9 (32.1 – 55.2)	41.5 (29.7 – 54.4)	0.40
Gradient index	2.92 (2.70 – 3.26)	2.92 (2.68 – 3.37)	0.98
Selectivity	0.88 (0.82 – 0.94)	0.87 (0.80 – 0.92)	0.29
Paddick conformity index	0.82 (0.78 – 0.85)	0.82 (0.77 – 0.85)	0.24

Table 3.4 — *Comparison of patient- and treatment-related characteristics between the fast growing and slow growing cohorts.*

age growing tumor cohort, respectively. This results in three cohorts containing 106, 109, and 97 patients with 5, 12, and 18 cases of treatment failure, respectively. The Chi-square test resulted in a p-value less than 0.01, indicating a significant difference in number of failures among the three cohorts. The calculated 5- and 10-year tumor control rates are 98.8% and 91.4% for the slow growing cohort, 90.6% and 70.7% for the average growing cohort, and 84.6% and 66.4% for the fast growing cohort, respectively. The resulting curves are presented in Figure 3.4, where a linear trend can be observed (log-rank test, $p < 0.01$). Here, the Kruskal-Wallis tests reveal that again no other factor is significantly different among the three cohorts.

B. Cox regression

Finally, we have also investigated the effect of the main characteristics on the tumor control rates by implementing univariate Cox regression analyses. The results have shown that only the VDT has a significant effect ($p < 0.05$). None of the other patient- and treatment-related characteristics, shown in Table 3.4, display a statistically significant influence. Consequently, the Cox regression is constrained to performing a univariate analysis and multivariate analyses are not needed. The impact of the VDT on the proportional hazards ratio is given by $\exp(bx)$ (see Eq. (3.2)), where constant $b = 0.97$. This means that the risk of loss of tumor control for a tumor with a given VDT (the x parameter) will decrease with factor 0.97 for a tumor for which the VDT is one month larger, i.e. a slower growing tumor. For example, the 5-year loss of tumor control in the Kaplan-Meier analysis is 8.4% in the patient cohort for a tumor with a median VDT of 15 months.

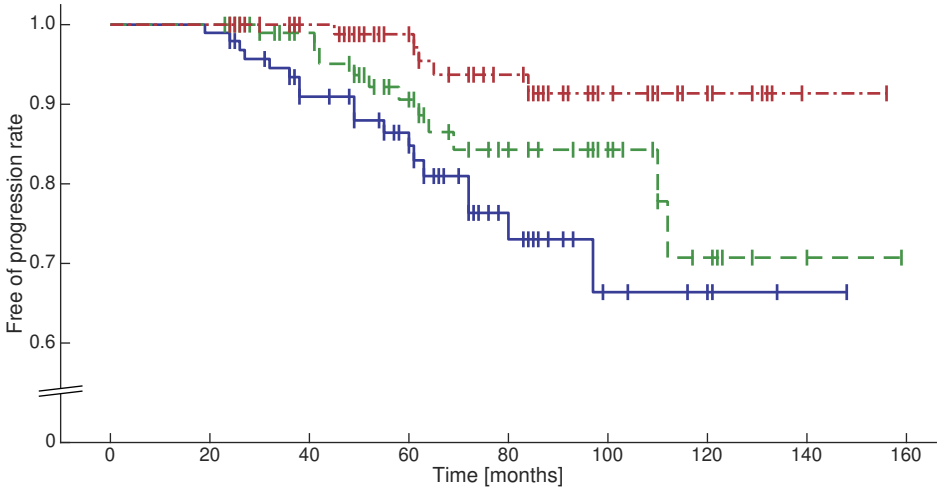


Figure 3.4 — *Kaplan-Meier curves for the fast growing cohort (blue continuous line), average growing cohort (green dashed line), and slow growing cohort (red dash-dotted line), stratified by the 33rd and 67th percentiles of the pre-treatment growth rate. Tick marks indicate censored cases.*

If the VDT increases by 12 months, the estimated 5-year loss of tumor control rate will be 5.8%.

C. Robustness of Our Findings

Since this work is mainly based on tumor volumes, the impact of the inter- and intra-observer inconsistencies of the volumetric assessments may be significant on the results. These inconsistencies are more critical in small tumors because relative errors become larger for decreasing volumes, as shown in Section 2.5. We have adopted a threshold on the minimum time required between scans in the VDT calculations. However, a threshold on the required minimum tumor volume is not imposed, which would reduce the impact of the relative volume errors. Hence, for small tumors the calculated VDT can be inaccurate. If we remove tumors smaller than 0.25 cm^3 in our cohort and redo the analyses, we observe a statistically improved result from the univariate Cox regression (286 patients, $p = 0.01$).

Furthermore, the inter- and intra-observer inconsistencies also have an impact on our additional volumetric treatment failure definition. If we redo the analyses without the so-called volumetric failures and only consider failures determined in the clinical setting, we still find a significant difference between the slow- and fast-growing tumor cohorts (Fisher's exact test, $p = 0.04$; Kaplan-Meier log-rank test, $p = 0.02$).

3.5 Discussion

This chapter has focused on investigating the relationship between the pre-treatment growth rate and the Gamma Knife radiosurgical efficacy. More specifically,

several experiments are conducted to examine the correlation of the VDT to the volumetric treatment response of VS tumors. This is relevant because if the growth rate prior to treatment influences the radiosurgical outcome, the question arises whether this justifies alterations in treatment management.

In this section, the different aspects of the implemented methodology and the obtained results of the pre-treatment growth-rate investigation are highlighted and discussed. First, the growth model is discussed in Section 3.5.1 and a comparison to the current state-of-the-art is made in Section 3.5.2. Next, in Section 3.5.3, our findings are summarized and we discuss possible explanations and consequences of these findings in Sections 3.5.4 and 3.5.5, respectively. Finally, some limitations of this investigation are reviewed in Section 3.5.6.

3.5.1 Growth model

The only way to accurately determine the influence of the pre-treatment growth rate to the GKRS treatment outcome is to study a cohort with quantifiable tumor progression prior to treatment. We have used the VDT model proposed by Varughese *et al.* [42] to quantify tumor progression. Since VSs do not necessarily grow at a regular rate, the VDT may not be a perfect fit for each individual case, as also stipulated by Talkington *et al.* [181]. However, we do not have the data to observe the growth of each individual tumor over a long period of time. Many tumors are treated when tumor growth is observed between two consecutive scans. With the pre-treatment data available in this work, the VDT is the most accurate way to describe VS growth, as Varughese *et al.* [42] concluded in their article, that specifically addresses the issue of determining growth rates of VSs.

3.5.2 State-of-the-art in methodology

Various previous studies have addressed the potential role of pre-treatment growth rates on GKRS treatment responses, but the results are conflicting [41], [44], [49], [55], [111], [165]. In our opinion, the reported inconsistent results can be explained by methodological shortcomings in these studies. Loss of tumor control is observed rarely and often occurs several years after treatment. This fact makes it mandatory that studies on this topic include large patient numbers and long follow-up times. All six aforementioned studies reported on a relatively low number of patients, and only three studies had median follow-up times significantly larger than the generally accepted time period for transient tumor enlargement (Table 3.5). Furthermore, it is important to apply volumetric measurements. Subtle loss of tumor control can go unnoticed when obtaining linear measurements. Such small changes may be irrelevant from a clinical perspective, i.e. not demanding an intervention. However, from a scientific viewpoint, it is important to identify all cases of tumor growth after radiosurgery. It allows for an accurate assessment of the correlation between the pre-treatment growth rate and the volumetric GKRS treatment response. Four of the six studies addressing the influence of pre-treatment growth rates on the treatment outcome utilized pre-treatment volumetric measurements. One study, by Timmer *et al.* [41], did not exploit the

actual linear pre-treatment measurements, but it stratified the cohort accordingly into growing and non-growing tumor cohorts. Chang *et al.* [55] also investigated whether there was a difference between growing and non-growing tumors. They found no statistical significant difference between both cohorts. Furthermore, our literature study has revealed that there are also inconsistencies among the definitions of treatment failure (Table 3.5). Therefore, in our research we have used volumetric tumor measurements in a large patient cohort with long follow-up times (Table 3.5).

3.5.3 Prediction model

Our data clearly illustrate that the pre-treatment growth rate correlates with the radiosurgical efficacy. Slow growing tumors in our cohort are more likely to exhibit tumor control than their fast growing counterparts. Using Kaplan-Meier analysis, the estimated 5- and 10-year tumor control rates are 97.3% and 86.0% for the slow growing tumors, and 85.5% and 67.6% for the fast growing tumors, respectively. This effect is also apparent if we stratify the data of this patient cohort into three groups, with most failures in the fastest growing cohort (n = 18), and an intermediate number of failures in the middle cohort (n = 12). This suggests a distinct effect of the pre-treatment growth rate on tumor control after GKRS. Indeed, the Cox regression analysis is significant for the pre-treatment growth rate expressed by the VDT ($p < 0.05$). In the resulting prediction model, the impact of the VDT on the risk of loss of tumor control equals a factor of 0.97. This means that increasing the VDT by one month results in a decrease with a factor of 0.97 of failure risk at a certain time. It is interesting to use this factor for evaluating the risk of loss of tumor control for various cases. Some examples of this evaluation can be found in Table 3.6.

3.5.4 Possible explanations of the findings

It has been suggested that a higher rate of loss of tumor control after radiosurgery for fast growing VSs can be explained by the radiobiological effect of slowing down the growth curve: fast growing tumors will be slowed down by radiosurgery, but possibly not enough to obtain tumor control, whereas the growth curve of slow growing tumors is bent sufficiently to obtain tumor control [49], [166]. However, our data indicate that relative post-treatment volume changes and tumor halving times do not demonstrate differences between fast and slow growing tumors. We therefore hypothesize that the intrinsic tumor biology of fast growing tumors make them more likely to start growing again several years after radiosurgery, rather than radiosurgery slowing down their growth rate. However, the radiobiological effect of radiosurgery on VS remains unclear from existing literature. There is an ongoing discussion whether the radiosurgical response of VSs results from direct cytotoxic effects to cells, or whether it reflects indirect effects. In the review by Yeung *et al.* [183], the authors discussed three possible mechanisms that could explain the decreased tumor control rates in certain VS tumors: (1) the

Authors and Year of Publication	No. of Patients (Failures)	Follow-up Time (Months)	Volumetric Tumor Measurements	Definition of Treatment Failure	Pre-treatment Growth Rate as Predictor
Varughese *et al.* 2012 [111]	45 (13)	Mean 50	Yes	Post-GKRS VDT > 0 months	No
Marston *et al.* 2017 [49]	68 (9)	Median 43.5	No	Diameter increase > 2 mm	Yes
Niu *et al.* 2014 [165]	58 (3)	Median 27.5	Yes	Expansion with homogeneous enhancement on MRI	Yes
Timmer *et al.* 2011 [41]	67 (5)	Mean 26	No	Secondary treatment needed	No
Larjani *et al.* 2014 [44]	63 (7)	Median 32	Yes	No stabilization or regression of significant growth at 24 months after GKRS treatment	No
Chang *et al.* 2019 [55]	62 (1: 47) (2: 2)	Median 40	Yes	(1) Significant volume increase ($> 20\%$) at last FU (2) Need for salvage treatment	No
Present study	311 (35)	Median 60	Yes	Secondary treatment needed and volumetric model	Yes

Table 3.5 — *Differences in methodological definitions and results among the various studies evaluating the relation between pre-treatment growth rate and GKRS treatment outcome.*

VDT (years)	5-Year hazard rate	10-Year hazard rate
0.5	11.0%	29.6%
1	9.2%	24.7%
2	6.4%	17.1%
3	4.4%	11.9%
4	3.1%	8.2%

Table 3.6 — *Predicted hazard rates for loss of tumor control calculated by the prediction model created in this research. These hazards are calculated including the volumetric failure definition.*

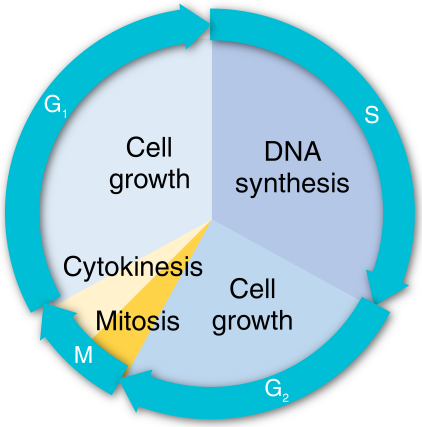


Figure 3.5 — *Visual representation of the cell cycle. There are four distinct phases in a cell cycle: (1) the G_1-phase in which the cell increases in size, (2) the S-phase in which DNA replication occurs, (3) the G_2-phase in which the cell increases further in size, and (4) the M-phase consisting of mitosis and cytokinesis. In this final phase, the cell stops growing and is divided into two daughter cells. There is an additional phase, called the G_0-phase, in which the cell is in a resting state.*

Merlin-induced imbalance in the c-Jun N-terminal kinase pathway and extracellular signal-related kinase pathway, (2) the inadequate angiogenesis and hypoxia, and (3) the radioresistance during cell cycle. A visual representation of the cell cycle is given in Figure 3.5. In general, cell survival data have demonstrated that cells are most sensitive to irradiation during mitosis and in the G_2-phase, less sensitive in the G_1-phase, and least sensitive during the S-phase [184]. This would indicate that radioresistance increases for tumors that relatively lack cell division, i.e. are slow growing. However, we have observed the opposite: fast growing VSs tend to respond less to radiosurgery. This would imply that either fast growing tumors have a superior DNA repair system [183], or the response to radiosurgery reflects indirect radiation effects, such as decreasing tumor vascularity [185].

3.5.5 Consequences of the findings

The results in this chapter clearly indicate that the pre-treatment growth rate influences the volumetric outcome of VS after GKRS. The prediction obtained in this work provides the opportunity to determine the risk of treatment failure for each specific VS, employing the VDT. However, this prediction model needs to be validated with data of other cohorts.

Nevertheless, the findings presented in this chapter raise the question of whether fast growing tumors should be treated differently, i.e. with a higher radiation dose, with microsurgery, or by reducing the radioresistance of these tumors employing radiosensitizers [185]–[188]. A prospective study should be designed to investigate whether an increased radiation dose, or a combination of radiosensitizers with GKRS, for fast growing VS increases the overall tumor control rates, without increasing toxicity. However, Fu *et al.* [182] recently showed that retreatment of VS by GKRS appears to be an effective strategy, suggesting that fast growing tumors may benefit from a second GKRS treatment.

Another possible benefit of being able to predict the chance of tumor control may be the personalization of the standard follow-up protocol: for patients with a slow growing VS, the follow-up interval may be extended, while for fast growing VS, patients may be monitored more closely.

3.5.6 Limitations

One of the most important limitations of this investigation, as well as other studies on this topic, is that there is no clear consensus on the explicit criteria for treatment failure after GKRS. Hence, direct comparison of tumor control rates reported in various studies is problematic, since the definition of treatment failure appears to be inconsistent [47], [49], [166]. Some studies define treatment failure as the requirement for microsurgical resection and do not mention whether a second GKRS treatment is considered as failure [189], [190]. Others, like the radiosurgical team at our center, define tumor control as the absence of radiologically identified progression, which is usually performed by linear assessment [45], [191], [192]. However, even with proven tumor progression, intervention may still be undesirable. Therefore, in addition to proven tumor progression followed by intervention, we have employed a mathematical model for determining treatment failure to simultaneously account for missed small progressions in the clinical setting and undesirable interventions. This model may not be clinically relevant, but it provides an objective measure for determining treatment failure. Because of this strict criterion, the results of our Kaplan-Meier analysis displayed lower tumor control rates than those reported by other studies.

Furthermore, the determination of tumor control is troublesome, because of the transient tumor enlargement phenomenon. Most tumors presenting with this adverse effect reach a maximum volume after a median of 5 months and first signs of regression at a median of 15 months [41], [131], [134], [137]. In this study, we tried to circumvent this issue in defining loss of tumor control by only considering volumetric measurements beyond two years after treatment. However, it is claimed

that the transient swelling can occur as late as 3 – 4 years after radiosurgery [166].

Another limitation of this research is that the data are retrospectively analyzed. This leads, for instance to uncertainties in the volumetric assessment of the tumors, due to differences in MRI scans. For 146 patients, the T1-weighted contrast-enhanced MRI study obtained prior to treatment was unavailable or of poor quality. Employing the T2-weighted MR images for these patients may have introduced uncertainties in the determined tumor volumes. This effect is more critical in small tumors, as relative errors in the volumetric assessments become larger for decreasing volumes [120]. However, we have shown that excluding these small tumors results in statistically improved outcomes.

3.6 Conclusions

In current literature, the influence of the pre-treatment growth rate of VS on the GKRS treatment response can be classified as undetermined, due to the conflicting results reported by various studies. Methodological imperfections can possibly explain these contradicting results. Because of the large number of patients, the long follow-up times of our cohort, and the volumetric tumor assessments both prior to and after GKRS treatment, we had the unique opportunity to accurately investigate the influence of the pre-treatment VS growth rate on the radiosurgical efficacy.

The most important findings of this chapter are as follows. The so-called ”bending-the-growth-curve” effect of Gamma Knife radiosurgery was not found in our data. However, the influence of the pre-treatment growth rate of VS tumors on the long-term GKRS treatment effects with respect to the tumor volume is established. The resulting tumor control rates confirm the high efficacy of GKRS for slow growing VS. The 33%-slowest growing tumors in our unique and large database obtain 5- and 10-year tumor control rates of 98.8% and 91.4%, respectively, even with the inclusion of a very strict and objective treatment failure measure thereby increasing the number of failures. However, the fast growing tumors exhibited significantly lower control rates. Our analyses demonstrate that for the 33%-fastest growing tumors, the calculated 5- and 10-year tumor control rates are 84.6% and 66.4%, respectively. For these cases, different treatment strategies may be considered. Furthermore, by employing the Cox regression, we were able to create a predictive model supporting tumor control. More specifically, by using the volume doubling time as a prediction parameter, the obtained model can be exploited to predict the 5- and 10-year chance of tumor control on an individual patient basis.

Additionally, the results of this research may help in patient counseling and in determining a patient-specific follow-up protocol, where the follow-up frequency of slow growing tumors may be reduced with respect to the frequency needed for fast growing tumors. Especially the fitted Cox regression model can be implemented in the clinical workflow to facilitate physicians in selecting the optimal treatment strategy. By using the volume doubling time of the specific VS tumor,

this model is able to calculate a risk at treatment failure following GKRS.

In this chapter, the impact of a clinically highly interesting tumor-specific risk factor to the Gamma Knife treatment response has been evaluated. The tumor growth rate prior to treatment, calculated by the volume doubling time, has shown to be predictive for the long-term treatment outcome. In the following chapter, the impact of the treatment planning itself on the treatment outcome is examined. This investigation is performed in three different ways. First, the generally accepted clinical planning parameters are evaluated for their impact on the treatment response. Second, we will introduce a novel approach for examining the underlying dose distribution and its effect on the GKRS outcome. Finally, the impact of expert annotations of the tumor on the treatment outcome is investigated.

Chapter 4

Dosimetric parameters

4.1 Introduction

The previous chapter has described one of the tumor-specific parameters that can be indicative for the long-term Gamma Knife treatment response of vestibular schwannomas. It has presented the results obtained in the unique large database from the Gamma Knife center in Tilburg. Using the tumor-specific growth rate as potential risk factor in statistical survival analyses, it has become clear that this characteristic is prognostic for the long-term treatment response. The obtained model enables the calculation of the risk at treatment failure for individual patients with an available pre-treatment MRI scan. This treatment prediction model can be readily implemented in the clinical workflow, since tumor volumes can be determined relatively easy and MRI scans are already obtained for diagnosis and for treatment planning.

In the previous chapter, various global treatment parameters have been included in the statistical analyses. Furthermore, the possible impact of the Gamma Knife treatment parameters and settings cannot be ignored, which was already discussed one chapter earlier in Section 2.2.2. These treatment-related parameters failed to show their significant relation to the treatment outcome. However, the data employed in the previous chapter only included patients with confirmed tumor progression prior to treatment. This patient selection may have caused a bias in the statistical analyses of these parameters.

In this chapter we therefore investigate the impact of the Gamma Knife treatment planning itself on the treatment response. In Section 1.2, an overview of the Gamma Knife modality is given, in which the many parameters and settings that form the final treatment plan are specified. These settings and parameters may be correlated to the individual treatment response and should therefore be investigated extensively.

In Section 2.2.2, an overview of the current state of the art in treatment-related characteristics was given. Furthermore, as can be distilled from Sections 2.2.2 and 2.2.4, the contradicting results in the described studies most likely originate from the significant differences between the implemented methodologies. Numerous publications have investigated the impact of several treatment-related parameters. Hasegawa *et al.* [32] and Lim *et al.* [57] both concluded that slightly higher doses to the tumor margin correlate to increased long-term tumor control

rates. However, many other publications have concluded that the impact of the dose to the tumor margin is not statistically significant in multivariate analyses. Indeed, in a systematic review by Germano *et al.* [60], the authors conclude that within the range of doses used for the treatment of VS, a lower dose had little to none appreciable difference in the progression-free survival, while generally high rates of progression-free survival were reported across a wide range of delivered doses.

In current state of the art, the influence of the treatment procedure itself on the outcome is only investigated using global parameters, such as the prescribed dose to the tumor margin, the number of isocenters, and the beam-on time. A possible reason for only considering such global treatment characteristics is that a Gamma Knife treatment is initially considered to be uniformly planned, conform protocols established by the Gamma Knife community. The dose distribution within the tumor is deemed to be similar to that of other radiosurgical modalities, such as CyberKnife® or linear accelerator (LINAC). However, when comparing Gamma Knife to these other modalities, the dose distribution is determined to be significantly less homogeneous for an elliptic target [193]. This is not a surprise, since Gamma Knife employs a widely varying number of isocenters with varying shapes, sizes, and weights, in contrast to for example LINAC, which uses only a single shot. Indeed, in a publication by Millar *et al.* [110], the authors determined that the biological effectiveness of a given physical prescription dose fluctuates with the variations in treatment parameters between different patients. However, it remains unclear if the actual differences in biological effectiveness influence the treatment outcomes. Nevertheless, it is hypothesized in this chapter that the inhomogeneous dose distribution of Gamma Knife treatments could potentially influence the treatment results. Unfortunately, the inhomogeneity of the dose distribution cannot be expressed in the global parameters commonly employed in literature. This motivates why we explore the predictive value of the spatial dose distribution on the treatment outcome of Gamma Knife radiosurgery on VS.

Another important step in treatment planning that may introduce uncertainties involves the delineation of tumors, because this is the basis for creating the highly accurate radiosurgical treatment planning. Each treatment planning is reviewed using quality indices, such as the coverage and the selectivity. These indices are all calculated using the delineated tumor volume. Therefore, an inaccurate delineation may lead to underexposure of parts of the tumor margin, which in turn can reduce the chance at treatment success. As described in Section 2.5, variations in tumor segmentation are present among operators, although the differences are generally small. Nevertheless, these small differences may be correlated to the Gamma Knife treatment response.

In summary, this chapter focuses on the different treatment planning parameters and their possible influence on the long-term treatment outcome, as described above. The following aspects of the treatment planning will be specifically investigated.

- *Global treatment parameters*: In current state-of-the-art research, global treatment parameters such as prescribed dose to the tumor margin, number of isocenters, and various quality indices are investigated. In this chapter, we will experiment on these factors to evaluate their impact on the treatment outcome in our own unique large database.

- *Inhomogeneous dose distribution*: One of the most interesting aspects of a Gamma Knife treatment is its heterogeneous dose distribution. This may cause significant differences among treatment plans, such as single or multiple hot-spots at varying locations, dose drop-offs inside the tumor, etc. To the best of our knowledge, the impact of these differences on the treatment response has never been investigated. Therefore, we provide the first experiments into analyzing this impact, and explore a novel approach to incorporate the spatial dose-distribution information.

- *Tumor delineation*: Another aspect that has not been considered in the evaluation of the treatment response, is the tumor delineation. To this end, we will explore the accuracy of the tumor delineation and its possible influence on the treatment response.

These aspects are all based on the various parameters and settings involved in the Gamma Knife treatment planning of VSs. However, since the underlying data and methodologies differ significantly among the specific experiments, each aspect is individually discussed in the following sections. First, the influence of global dose parameters, such as the radiation dose to the tumor margin, on the long-term tumor control is investigated in Section 4.2. Next, Section 4.3 examines the influence of the heterogeneous dose distribution with respect to the Gamma Knife treatment response, and presents the results of these methods. Following this, the experiments on the effect of expert annotations on the volumetric response are discussed in Section 4.4. Finally, in Section 4.5, the main outcomes and conclusions are summarized.

4.2 Global treatment plan parameters

In this section, the influence of numerous global treatment-plan parameters on the treatment response is investigated in various experiments. This section is divided into the following subsections. First, a brief background is discussed in Section 4.2.1. Next, the experimental setup is presented in Section 4.2.2, followed by its results in Section 4.2.3. Section 4.2.4 concludes this section with a discussion and conclusions.

4.2.1 Background

In the last decades, the prescribed dose has dropped significantly for the treatment of VSs. Nowadays, most medical centers opt for a dose of 12 – 14 Gy prescribed to the tumor margin. Since small irregularities in the tumor shape can cause high radiation dosage outside the tumor, this so-called *marginal dose* does not cover

the tumor completely. In the early years at the Gamma Knife center in Tilburg, the coverage percentage was selected to be 90%, resulting in a marginal dose of approximately 11 Gy. However, after evaluating the tumor control rates and the follow-up complication cases in a large cohort, the control rates were slightly unfavorable compared to other large studies [47]. After assuming that these differences were directly attributable to the employed dosimetry, the center changed the treatment protocol in 2010 to a marginal dose of 13 Gy. In practice, this meant that the coverage of the prescription isodose volume was raised from 90% to 99%. This dose-level change in the treatment protocol enables evaluation of the impact of the marginal tumor dose to the long-term treatment outcome. Current state of the art shows that its influence is limited. However, as discussed in Sections 2.2.2 and 4.1, the results can be considered inconclusive, since some authors have demonstrated that higher doses lead to increased tumor control rates. Nevertheless, it is hypothesized that minor differences (in the order of magnitude) in marginal tumor dose, constrained by internationally established treatment protocols, do not influence the long-term tumor control rates.

4.2.2 Global treatment parameter experiments

In the experiments that will investigate the impact of global treatment-related parameters on the long-term tumor control rates, the complete database of the Gamma Knife center in Tilburg will be employed. This database is introduced in Section 2.3 and contains 735 unilateral VS patients that were treated between 2002 and 2014.

Several methods are available for evaluating the influence of specific potential risk factors to the long-term treatment outcome. Most methods employed in medical papers are so-called survival analyses, as discussed in Section 3.3.2. Generally, for investigating continuous variables, a Cox proportional hazards regression analysis is performed. Therefore, in this section, each global treatment-related parameter is evaluated using the Cox regression analysis.

Another method for assessing the influence of risk factors to the treatment response is the Kaplan-Meier survival analysis. This method enables the comparison of two or more (sub)cohorts by examining the difference in survival curves. By splitting the complete cohort using different treatment-related parameters, the influence of those specific parameters can be investigated. Since the treatment protocol has changed in 2010, we are able to use the marginal tumor dose of the Tilburg dataset as splitting variable in this analysis.

By reviewing the literature, as described in Section 2.2.4, significant methodological variations can be discerned when comparing studies that discuss long-term tumor control rates. Furthermore, as argued in Section 3.1, radiosurgical efficacy in terms of true tumor control can only be accurately evaluated in patients who have exhibited radiological progression prior to treatment [194]. In order to evaluate the effect of the dosimetry on the long-term tumor control, these issues need to be addressed. We evaluate the impact of these confounding factors on the conducted experiments. Therefore, the influence of employing an objective

Risk factor	p-Value
Age at treatment	0.70
Tumor volume at treatment	<**0.01**
Prescription isodose line	0.46
Coverage	0.16
Selectivity	<**0.01**
Gradient index	0.46
Paddick conformity index	0.10
Number of isocenters	**0.03**
Beam-on time	0.37
Dose covering 99% of tumor volume	**0.05**

Table 4.1 — *Results of the univariate Cox regression analyses. The relevant p-values are indicated in bold.*

failure definition is assessed. Furthermore, the effect is examined of only including patients that presented radiological progression prior to treatment.

4.2.3 Results on global treatment parameter experiments

The following subsections highlight the results of the conducted experiments concerning the global treatment-related parameters. First, the results on the Cox regressions are presented. Next, the outcomes of the Kaplan-Meier analyses are given. Finally, the assessment results of the two methodological confounding factors are depicted.

A. Cox regression

In the univariate Cox regression, the following factors have shown a significant correlation to the long-term tumor control in the large Tilburg database: tumor volume at treatment, selectivity, and the number of isocenters. The p-value for the dose covering 99% of the tumor volume is slightly higher than the required 0.05 for obtaining statistical significance. The results from the univariate analyses can be found in Table 4.1. In a multivariate Cox regression analysis with step-wise forward selection, only the tumor volume at treatment ($p < 0.01$), and the dose covering 99% of the tumor volume ($p = 0.03$) are determined to be significant covariates. These results suggest that, next to the tumor size, the dose to the tumor margin does influence the long-term tumor control.

B. Kaplan-Meier

In the Tilburg dataset, the difference in marginal tumor dose between the two employed protocols is approximately 2 Gy. In contrast with our hypothesis, the above-described results from the Cox regression analyses suggest that the patients treated with the first protocol, i.e. with a lower marginal dose, have a considerable

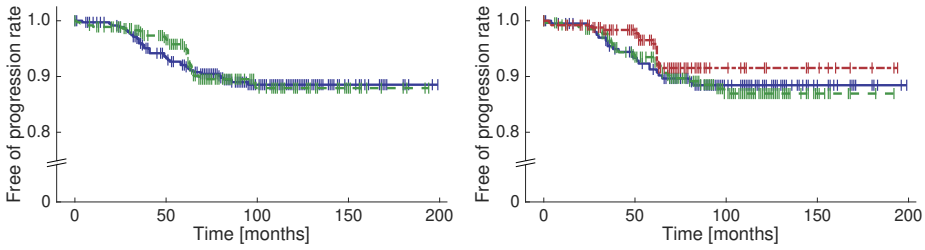


Figure 4.1 — *Left: Resulting Kaplan-Meier curves of the low-dose cohort (blue, solid) and the high-dose cohort (green, dashed). Right: Resulting Kaplan-Meier curves of the low-dose cohort (blue, solid), the average-dose cohort (green, dashed) and the high-dose cohort (red, dot-dashed).*

higher risk at loosing tumor control than those treated with the high-dose protocol. To evaluate whether this is the case, a Kaplan-Meier analysis is performed. By using the median of the marginal tumor dose, i.e. the dose covering 99% of the tumor volume, two cohorts are created.

First, these two cohorts are compared to evaluate whether there are significant differences between them, apart from the treatment planning differences. Specifically the tumor volume at treatment is important in this analysis, since this factor has shown its impact in the previously described results. From Table 4.2, it can be distilled that the only significant differences between both cohorts are related to the treatment planning and to the follow-up duration. However, differences in the number of isocenters and the Paddick conformity indices are not significant. The significant difference in follow-up time can be explained by the fact that the high-dose protocol has been implemented since 2010, while the low-dose protocol was used in the preceding years. The differences in treatment planning parameters between the two cohorts are to be expected, because they are all related to the prescribed dose and employed treatment protocols. The number of patients that underwent secondary treatment for their progressing tumor is also not significantly different between both cohorts ($p = 0.11$). These numbers for the low-dose cohort and the high-dose cohort were 37 and 25, respectively. The time to failure was also comparable between the two cohorts ($p = 0.50$).

Second, the Kaplan-Meier survival tables and curves are generated and compared, using a log-rank test. The resulting curves can be found at the left side of Figure 4.1. The log-rank test determined that there was no significant difference between the two cohorts with respect to their survival curves ($p = 0.69$).

Separation of the two cohorts is performed at the median of the marginal tumor dose in our cohort. However, this value can be arbitrarily chosen, but it should be kept in mind that it has a considerable influence on the outcome. The same analysis can be performed again, but now employing three cohorts, selecting the lowest one-third of the patients as the low-dose cohort, the middle one-third as the average-dose cohort, and the highest one-third as the high-dose cohort.

Characteristic	Low-dose	High-dose	p-Value
Number of failures	37	25	0.12
Time to failure (mos)	41 (32 – 62)	52 (35 – 63)	0.50
Age at treatment (yrs)	57 (48 – 66)	59 (49 – 67)	0.09
Volume at treatment (cm^3)	1.40 (0.60 – 3.75)	1.59 (0.64 – 3.91)	0.47
Follow-up time (mos)	96 (72 – 121)	60 (48 – 84)	<**0.01**
Prescription isodose line (%)	61.0 (57.0 – 66.0)	49.0 (45.0 – 60.0)	<0.01
Coverage (%)	90.0 (89.0 – 91.0)	98.0 (92.0 – 99.0)	<**0.01**
Selectivity (%)	93.0 (87.0 – 96.0)	86.0 (80.0 – 91.0)	<**0.01**
Gradient index	3.17 (2.89 – 3.47)	2.75 (2.61 – 3.06)	<**0.01**
Paddick conformity index	0.83 (0.78 – 0.86)	0.83 (0.77 – 0.86)	0.18
Number of isocenters	16 (11 – 22)	15 (11 – 22)	0.06
Beam-on time (mins)	39.1 (29.8 – 48.4)	46.9 (33.7 – 62.6)	<**0.01**
Dose to 99% of volume (Gy)	11.4 (11.2 – 11.6)	12.6 (12.0 – 13.0)	<**0.01**

Table 4.2 — *Comparison between the cohorts receiving a low dose and receiving a high dose. The value ranges are given as "median (inter-quartile range)". The p-values lower than 0.05 are considered statistically significant, and are indicated in bold.*

The resulting Kaplan-Meier curves can be found at the right side in Figure 4.1. From this graph, it can be distilled that the three cohorts are comparable. Indeed, the log-rank test resulted in a p-value of 0.35. If only the low-dose and high-dose cohorts in this experiment are compared, the log-rank test obtains a p-value of 0.20. These results show that, even though the Cox regression determined that the dose is a significant covariant, the differences between tumor control rates within the protocolized doses are small and statistically not significant.

C. Confounding factors

The results described in the previous subsections show the difficulty in determining the influence of the dose on the long-term treatment outcome. The related conclusions based on these results are highly dependent on the method for analyses and the interpretation of the obtained results. Furthermore, several confounding factors related to methodology can influence the obtained results. In the following paragraphs, the impact of an objective failure definition and the influence on the outcome when only including confirmed progressing tumors are presented.

Factor 1: Objective failure definition

With the introduction of an objective failure definition in this thesis (Section 2.4.1), loss of tumor control can be determined more robustly. In the Tilburg dataset, 39 patients showed a volumetric failure, but did not (yet) require or undergo secondary treatment. Performing the same analyses as in the previous subsections, but now with the inclusion of the volumetric failures, will produce more robust

Risk factor (incl. volumetric failures)	p-Value
Age at treatment	0.64
Tumor volume at treatment	0.05
Prescription isodose line	0.74
Coverage	0.75
Selectivity	0.26
Gradient index	0.97
Paddick conformity index	0.41
Number of isocenters	0.31
Beam-on time	0.80
Dose covering 99% of tumor volume	0.23

Table 4.3 — *Results of the univariate Cox regression analyses with the inclusion of volumetric failures.*

and generic results. The Fisher's exact test, comparing the number of failures in the low-dose and the high-dose tumor cohorts, obtains a p-value of 0.07. This indicates that the number of failures does not differ significantly between the two cohorts, although medical experts may say that a trend is present. Furthermore, the time to failure does not show a statistically significant difference. In the univariate Cox regression analyses, none of the evaluated risk factors appears to be a statistically significant covariate. Even tumor volume at treatment is no longer statistically relevant, although it only has a slightly higher p-value than the required significance level. All resulting p-values can be found in Table 4.3.

Indeed, the Kaplan-Meier curves show comparable results to the case where only secondary treatments are considered as failure. Figure 4.2 presents the curves where the volumetric failure definition is included in the analyses. According to the log-rank tests, these curves do not differ significantly.

Factor 2: Pre-treatment progressing tumors

Since the Gamma Knife center in Tilburg is a tertiary referral center, and generally a wait-and-scan protocol is chosen in the Netherlands for vestibular schwannomas, the database contains a significant amount of patients who showed radiological tumor progression prior to Gamma Knife treatment. As such, the above-described evaluation can be executed for the tumors that were considered to require treatment due to a radiologically progressing tumor. Of the 735 patients in the database, 441 patients were referred due to growing tumors. By splitting the database prior to selecting only these 441 patients, the same dose cut-off is employed. After selecting only progressing tumors, a comparison of the resulting two cohorts highlights that only the tumor volume at treatment is a statistically significant covariate. The other risk factors obtained similar results when compared to the analysis on the

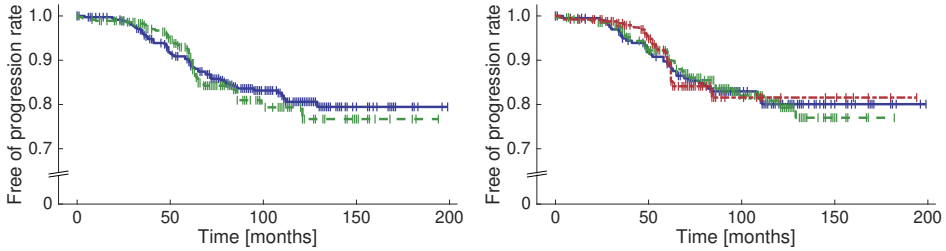


Figure 4.2 — *Kaplan-Meier curves where failure was defined as secondary treatment and volumetric progression. Left: Resulting Kaplan-Meier curves of the low-dose cohort (blue, solid) and the high-dose cohort (green, dashed). Right: Resulting Kaplan-Meier curves of the low-dose cohort (blue, solid), the average-dose cohort (green, dashed) and the high-dose cohort (red, dot-dashed).*

Characteristic	Low-dose	High-dose	p-Value
Number of patients	191	250	
Number of failures	22	15	0.12
Time to failure (mo)	41 (32 – 62)	60 (49 – 62)	0.29
Age at treatment (yr)	59 (52 – 69)	60 (51 – 68)	0.68
Volume at treatment (cm^3)	1.03 (0.53 – 2.45)	1.26 (0.63 – 2.94)	<0.01
Follow-up time (mo)	96 (72 – 121)	60 (48 – 83)	<0.01
Prescription isodose line (%)	61.0 (57.0 – 66.0)	49.0 (45.0 – 57.0)	<0.01
Coverage (%)	90.0 (88.0 – 91.0)	98.0 (95.0 – 99.0)	<0.01
Selectivity (%)	92.0 (86.0 – 95.0)	85.0 (79.0 – 90.0)	<0.01
Gradient index	3.17 (2.92 – 3.52)	2.74 (2.64 – 3.05)	<0.01
Paddick conformity index	0.82 (0.78 – 0.85)	0.82 (0.77 – 0.86)	0.39
Number of isocenters	14 (10 – 19)	13 (9 – 19)	0.58
Beam-on time (min)	36.1 (28.3 – 45.2)	45.8 (31.0 – 59.7)	<0.01
Dose to 99% of volume (Gy)	11.4 (11.2 – 11.6)	12.6 (12.2 – 13.0)	<0.01

Table 4.4 — *Comparison between the cohorts receiving a low dose and receiving a high dose of all patients with radiological tumor progression prior to Gamma Knife treatment. The value ranges are given as "median (inter-quartile range)".The p-values smaller than 0.05 are considered statistically significant.*

complete database. Table 4.4 displays all characteristics and resulting p-values.

In the univariate Cox regression analyses, no significant covariates are found, as depicted in Table 4.5. Even the dose to 99% of the tumor volume was no longer a significant covariate. It is interesting to observe that the tumor volume at treatment is also not a significant covariate in this cohort with only progressing tumors. This can be explained, because larger tumors do not undergo the wait-and-scan protocol. These tumors are referred immediately after diagnosis, due

Risk factor (incl. only progressing tumors)	p-Value
Age at treatment	0.99
Tumor volume at treatment	0.73
Prescription isodose line	0.41
Coverage	0.31
Selectivity	0.46
Gradient index	0.15
Paddick conformity index	0.82
Number of isocenters	0.60
Beam-on time	0.35
Dose covering 99% of tumor volume	0.65

Table 4.5 — *Results of the univariate Cox regression analyses on a cohort where only progressing tumors are included. None of the resulting p-values indicate statistical significance.*

to their size. Therefore, this cohort is not a good representation of the VS patient population.

4.2.4 Discussion and conclusions on global parameters

In this section, the unique large database of the Gamma Knife center at the ETZ in Tilburg is exploited for evaluating the influence of global treatment parameters on the long-term tumor control. First, we employed Cox regression analyses for this investigation. The obtained results show that the tumor volume at treatment is a significant risk factor and that the dose covering 99% of the tumor volume becomes statistically relevant in the multivariate analysis. It is interesting that this dose is related to the long-term treatment response, particularly because literature shows inconclusive results. The odds ratio, i.e. the ratio between those obtaining loss of tumor control and those that do not, equals 0.63 for the dose. As presented in Section 3.4, this results in a decreasing risk of loss of tumor control with factor 0.63 for a tumor receiving a specific marginal tumor dose compared to a tumor with a 1-Gy higher dose, which is quite significant. Thus, when increasing dose, the tumor control rates will also expand accordingly. However, increasing the dose may lead to a significant rise in toxicities, and is therefore unwanted.

The Kaplan-Meier analyses have demonstrated the opposite: splitting the complete cohort using the marginal tumor dose does not result in a statistically significant difference in tumor control rates among the created sub-cohorts. Although the previous paragraph indicated that the dose may influence the long-term tumor control rates, it can be deduced that the effect is negligible within the limited protocolized dose range, according to the Kaplan-Meier results. Furthermore, by investigating two major confounding factors, the obtained results show a loss in statistical significance of all analyzed risk factors. It can be therefore concluded that the effect of these globally determined treatment parameters have no influence on

the long-term tumor control, as long as they remain within protocolized ranges.

However, since Gamma Knife treatments are not homogeneous in their dose distribution, the analysis of only global parameters may be too restrictive to determine the actual impact of this treatment on the tumor response. Therefore, in the next section, the heterogeneous character of the dose distribution will be examined.

4.3 Dose distribution

In this section, the impact of the heterogeneous dose distribution on the treatment response will be investigated. This section is divided in the following subsections. First, a brief background is given in Section 4.3.1. Second, the conducted experiments are explained in Section 4.3.2 and their corresponding results are presented in Section 4.3.3. Finally, in Section 4.3.4, a discussion on the obtained results and conclusions on these experiments are given.

4.3.1 Dose distribution background

A single Gamma Knife treatment consists of the execution of multiple isocenters at different positions within the tumor, building up a specific marginal tumor dose level. Therefore, the resulting plan is heterogeneous in its dose distribution within the tumor. This may cause significant differences among treatment plans, such as single or multiple hot-spots at varying locations, dose drop-offs inside the tumor, etc. Indeed, Yu *et al.* [193] proved that the Gamma Knife has a highly heterogeneous dose distribution compared to other modalities. They employed dose-volume histograms and specific homogeneity index (HI) values for this comparison. Furthermore, each individual plan is tailored to the specific tumor. As such, each plan is unique in its dose distribution. Therefore, the differences between treatment plans with respect to the heterogeneous dose distributions may impact the response. Up to now, only global treatment-related parameters have been investigated, as discussed in the previous section. However, the inhomogeneity of the dose distribution cannot be expressed in the global parameters employed in literature. This motivates why the predictive value of the dose distribution on the treatment outcome of GKRS on VSs is investigated.

4.3.2 Dose distribution experimental setup

The steps for realizing the proposed experiments are shown in Figure 4.3. First, descriptions of the pre-processing steps applied to the data are provided. Next, the methods for extracting dose-distribution features are presented. These methods are described in more detail below. The final step of the approach consists of classification and prediction of the treatment outcome. Classification is performed by support vector machine (SVM) and validation of these steps is realized with leave-one-out (LOOCV) and 10-fold cross-validation methods. The classification performance is measured as accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the receiver operating characteristic curve (AUC).

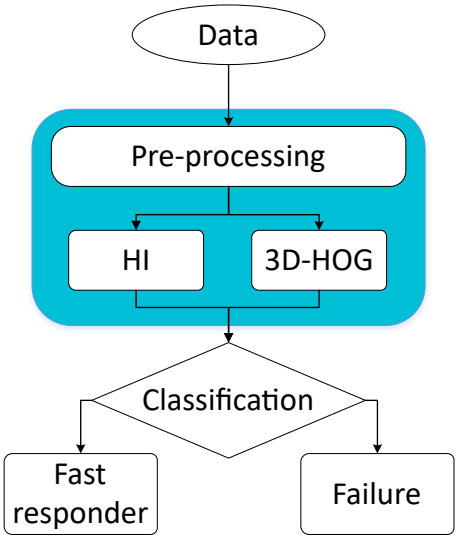


Figure 4.3 — *Flow diagram of the proposed approach for inhomogeneous dose distribution. For the analysis of the dose distribution, both homogeneity indices (HI) and three-dimensional histograms of oriented gradients (3D-HOG) are calculated.*

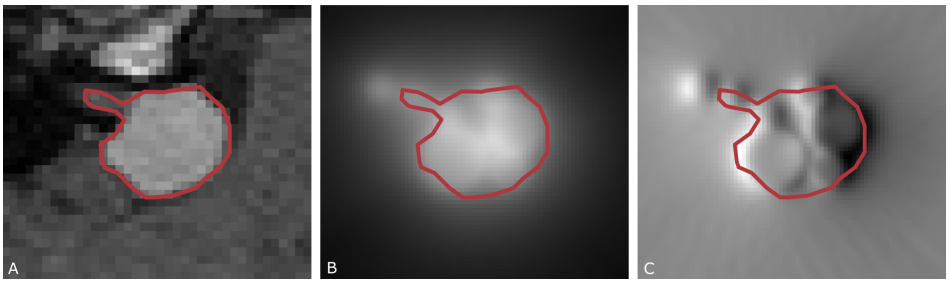


Figure 4.4 — *Image examples of the data. The contouring determined by the neurosurgeon during treatment planning is superimposed in red. (A) T1-weighted, contrast-enhanced MRI of a vestibular schwannoma, where the image part containing the tumor is delineated. (B) Calculated dose distribution, where the dose intensities are proportional to the gray-level intensities. (C) Gradient of the dose distribution in the x-direction. The strength of the gradient is represented by the gray-level intensities, where fully white is a strong positive gradient, and fully black a strong negative gradient. The different isocenters used in a Gamma Knife treatment are clearly visible in this image.*

A. Pre-processing the dose data

As described in Section 1.2, each treatment planning requires a T1-weighted, contrast-enhanced MRI scan. Using this scan, the VS tumor is segmented by the neurosurgeon. An example image of such an MRI scan, including tumor segmentation, is visualized in Part A of Figure 4.4. The segmentation consists of a set of x- and y-coordinates per MRI slice in the axial MRI plane, and are generated with sub-pixel accuracy.

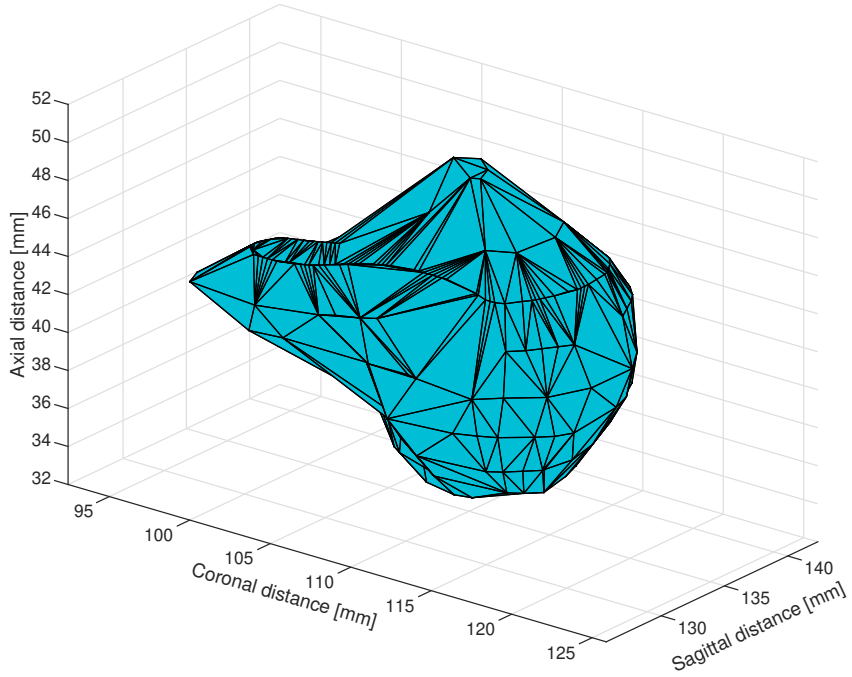


Figure 4.5 — *Example of a three-dimensional interpolation of VS contours.*

After treatment planning, the dose distribution details are calculated on a $0.5 \times 0.5 \times 0.5$ cubic millimeter grid, representing the stereotactic coordinate system of the Gamma Knife treatment machine. An example of such a dose distribution can be seen in Figure 4.4, Part B. In this image, the pixel intensity values are related to the dose intensities. Compared to the fine grid of the dose distribution within the tumor, the segmentation contours have a coarse axial resolution. Hence, in order to determine the dose distribution within the tumor, the segmentation contours need to be mapped to the dose distribution space. To achieve this, the following pre-processing steps have been employed. First, both spaces are mapped to the same image coordinate system. Second, a 3D tumor shape is created using the MATLAB AlphaShape tool, by interpolating the contours to fit the axial resolution of the dose distribution. An example of such a 3D tumor shape is presented in Figure 4.5. The value for parameter α used by the MATLAB tool is minimized per tumor, such that the generated 3D tumor surface does not contain any holes while maintaining a high conformity to the provided segmentation. In the final step, the dose distribution voxels inside the interpolated tumor volume are extracted.

B. Homogeneity indices

In order to measure the homogeneity of a dose distribution, Kataria *et al.* [195] have proposed four equations, calculating the homogeneity indices (HIs). The HIs are based on the dose-volume histograms (DVHs). These DVHs are calculated for

each patient on the basis of the input percentage. Employing these computations, the following four HI values are calculated by the respective parameters HI_i and the $DVH(x)$:

$$HI_1 = \frac{DVH(5)}{DVH(95)}, \tag{4.1}$$

$$HI_2 = \frac{DVH(1) - DVH(98)}{D_p} \cdot 100\%, \tag{4.2}$$

$$HI_3 = \frac{DVH(5) - DVH(95)}{D_p} \cdot 100\%, \tag{4.3}$$

$$HI_4 = \frac{DVH(1)}{D_p}. \tag{4.4}$$

In the above equations, $DVH(x)$ denotes the minimum dose that $x\%$ of the target volume receives, and D_p is the prescribed dose to the target.

C. 3D Histogram of oriented gradients

The previously described HIs do not consider spatial information, since these are calculated using the dose-volume histograms. As such, the locations of hot- and cold-spots are not taken into account. To incorporate the spatial information on the dose distribution, three-dimensional histograms of oriented gradients (3D-HOGs) are employed. These 3D-HOGs are calculated on each of the dose distributions, and are based on the work by Dalal and Triggs in 2D [196]. The computation is implemented by calculating the gradients of the dose distribution in x-, y-, and z-directions and performing orientation binning.

The gradients are computed by employing centered one-dimensional point-derivative masks inside a bounding box of $80 \times 80 \times 64$ voxels, encompassing the VS tumor. A visual example of such a gradient measurement of the dose distribution can be found in Part C of Figure 4.4. In this image, the heterogeneous character of the dose distribution is clearly visible. Using the calculated gradients in the x-, y-, and z-directions, a gradient vector can be constructed for each dose distribution voxel. This gradient vector $\vec{a}$ is defined for each dose distribution voxel $D(i)$ by:

$$\vec{a}\,(i) = \begin{bmatrix} \frac{\partial}{\partial x} D(i) \\ \frac{\partial}{\partial y} D(i) \\ \frac{\partial}{\partial z} D(i) \end{bmatrix}, \tag{4.5}$$

where i is the dose distribution voxel index.

Orientation binning is performed in cells, where each cell contains $16 \times 16 \times 16$ voxels, resulting in $5 \times 5 \times 4$ cells for each tumor. Each bin of the histogram can be interpreted as a unit vector $\vec{b}_n$ within an (x, y, z) coordinate system. We have implemented two different experiments, where the amount of bin vectors $\vec{b}_n$ per cell equals either 6 or 26. This results in a total of 600 and 2,600 bin vectors,

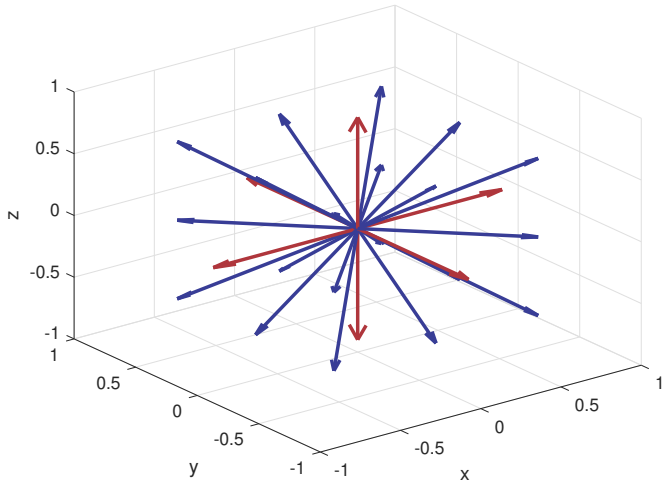


Figure 4.6 — *Illustration of the bin vectors employed in the 3D-HOG calculations. The bin vectors for a 6-bin HOG are depicted in red, while the additional 20 vectors for the 26-bin HOG are highlighted in blue, and are derived from the 6 red vectors.*

respectively, for the complete bounding box around the tumor. When using 6 bin vectors, every bin vector is oriented along an axis in a 3D space in either the positive or negative direction, i.e. $\pm x$, $\pm y$, and $\pm z$ (see Figure 4.6 in red color). When employing 26 bin vectors, the additional 20 vectors are constructed from the base 6 vectors (i.e. the vectors in the $+$ and $-$ directions of the x-, y-, and z-axes) with a separation angle of 45°. These additional vectors are derived by linear weighting of the nearest neighboring three red vectors. These 20 vectors form a unit sphere as they all have radius unity, and can be found in Figure 4.6 in blue.

Next, the angle θ_n between gradient vector $\vec{a}\,(i)$ and every bin vector $\vec{b}_n$ is computed. In the 2D case presented in the paper by Dalal and Triggs [196], every gradient pixel votes for the two closest bin vectors, i.e. the two vectors $\vec{b}_{k,l}$ associated with the smallest angle to $\vec{a}$. Here, we employ three-dimensional data, thus each gradient voxel should vote for the three bin vectors that have the smallest θ_n with respect to the gradient vector $\vec{a}\,(i)$. These votes are calculated by:

$$v_k = |\vec{a}\,(i)\,| \, \frac{\theta_l + \theta_m}{2\,(\theta_k + \theta_l + \theta_m)}, \tag{4.6}$$

$$v_l = |\vec{a}\,(i)\,| \, \frac{\theta_k + \theta_m}{2\,(\theta_k + \theta_l + \theta_m)}, \tag{4.7}$$

$$v_m = |\vec{a}\,(i)\,| \, \frac{\theta_k + \theta_l}{2\,(\theta_k + \theta_l + \theta_m)}, \tag{4.8}$$

where k, l, and m are the indices of the bin vectors with the smallest angle θ to $\vec{a}\,(i)$. The closer the gradient vector $\vec{a}\,(i)$ is to a bin vector, the higher its vote

for that specific bin becomes. For each bin vector $\vec{b}_n$ within a cell of $16 \times 16 \times 16$ gradient voxels, the sum of all its votes v_n is calculated. Subsequently, each cell is represented with an array of size 6 or 26, depending on the number of employed bins, containing the sum of voting values of its vector bins. Finally, for each dose distribution, the arrays containing the bin vector values of each individual cell are combined to construct an array of 600 or 2,600 voting values. This array is implemented in the machine learning stage as input feature.

4.3.3 Results on dose distribution experiments

In this section, first the data used in this research are described, after which the classification results of the individual features are presented. All implementations are realized in MATLAB (MathWorks inc., Natick, Massachusetts, USA). Training of the models is performed by the Classification Learner Application in MATLAB using default settings. For the 3D-HOG features, principal component analysis (PCA) is applied to limit the number of features by using the most relevant Eigenvalues and discarding the other ones.

A. Dataset

The dataset in these experiments involves clinical dose distribution data of 40 selected VS tumors. The VS segmentation contours were drawn by the neurosurgeon during treatment planning, and are thus available for each tumor for which the volumetric response to Gamma Knife is known. The treatment resulted in a failure on 20 tumors, and on the remaining 20 tumors, Gamma Knife was considered successful as these tumors showed a significant volume reduction within the first year after treatment (fast responders). Treatment outcome was determined by evaluating the volumetric response, where an increase in tumor volume is considered a failure. This evaluation was performed in a clinical setting by a medical team of specialists. In contrast to this, treatment success is difficult to define. The response of VS to Gamma Knife can only be determined after years of follow-up, where treatment failures can still occur after several years [97]. Therefore, we selected tumors that displayed a significant decrease in volume within a relative short time-period following treatment. The relative tumor volumes are plotted over time in Figure 4.7, where time was defined at treatment moment and at follow-up visits. The dose distribution is calculated by the treatment planning software (GammaPlan, Elekta AB, Stockholm, Sweden) and is mapped to a 3D space with a resolution of 0.5 mm $\times$ 0.5 mm $\times$ 0.5 mm, with every element containing a high-precision representation (16 bits) of the delivered dose.

B. Homogeneity indices

Employing the DVH per tumor, the HI features and their means and variances for the two patient cohorts are calculated. The resulting statistics are depicted in Table 4.6. To evaluate differences between the two cohorts, the results for each HI are tested using a paired *t*-test. The resulting p-values are also listed in Table 4.6. There is no significant statistical difference between the two cohorts, as indicated

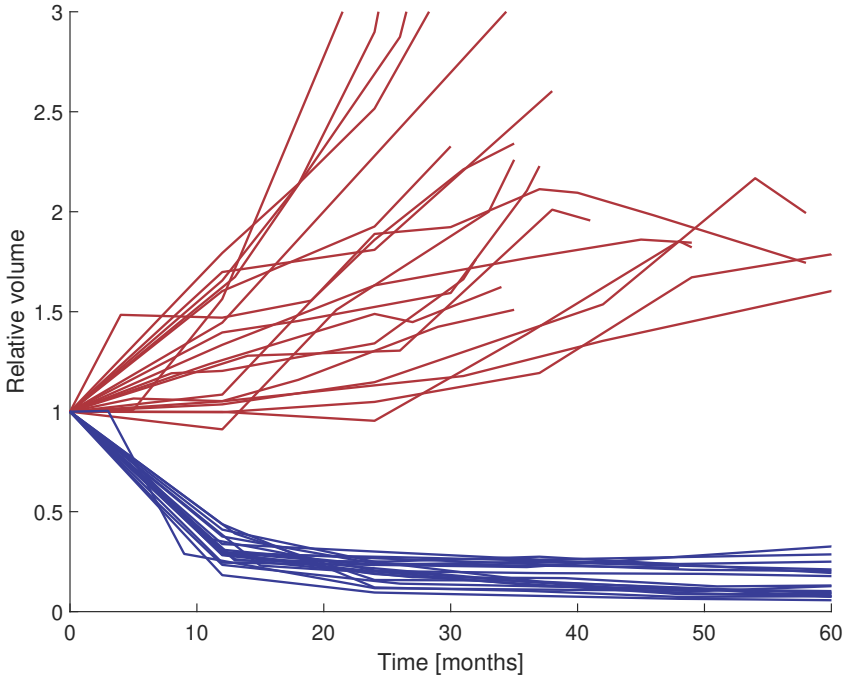


Figure 4.7 — *Plot of the relative volumes of the tumor following Gamma Knife radiosurgery. The red curves refer to the volume responses of the VS tumors for which the treatment failed. In blue color, the volume changes of the rapid responding VSs are highlighted.*

by a p-value smaller than 0.05. Thus, it can be concluded that there is no statistical difference in HI values between the successfully treated tumors and the tumors for which treatment failed. Indeed, employing these features in a machine learning algorithm yielded unreliable results (see Table 4.7). One particular trained model obtained zero scores for all performance indices. This happens if the model assigns labels randomly. The SVM model cannot find a reasonable decision boundary in the data, since the classes are inseparable and the model predicts everything into the majority class. Because we have 2 cohorts of 20 tumors, leaving one out will result in an incorrect prediction, since the model predicts the opposite, majority class. This results in all-zero scores. Furthermore, the highest performing SVM trained on these features obtained an ACC of 52.5%, showing that these features cannot distinguish treatment failure from treatment success.

C. 3D Histogram of oriented gradients

Employing a bounding box of $80 \times 80 \times 64$ voxels, divided into $5 \times 5 \times 4$ cells, we obtain a number of 3D-HOG features that is equal to 100 times the number of bins. Since the number of features is high, PCA is employed to reduce this number. The classification results of these 3D-HOG features show that overall,

Homogeneity index	Successful		Unsuccessful		p-Value
	Mean	Std.	Mean	Std.	
HI_1	1.94	0.72	1.80	0.27	0.39
HI_2	97.8	31.4	93.6	27.1	0.59
HI_3	73.2	26.6	69.7	19.5	0.59
HI_4	1.80	0.28	1.77	0.25	0.61

Table 4.6 — *Statistical description of the resulting homogeneity index (HI) values. For each HI feature, the mean and standard deviation (std.) are presented, including the resulting p-values from the paired t-tests comparing the two cohorts.*

SVM type	Validation	ACC [%]	TPR [%]	TNR [%]	AUC
Linear	10-fold CV	40.0	70.0	10.0	0.32
	LOOCV	12.5	5.0	20.0	0.05
Quadratic	10-fold CV	52.5	55.0	50.0	0.53
	LOOCV	37.5	40.0	35.0	0.41
Cubic	10-fold CV	52.5	55.0	50.0	0.40
	LOOCV	40.0	40.0	40.0	0.28
Fine Gaussian	10-fold CV	50.0	50.0	50.0	0.59
	LOOCV	42.5	40.0	45.0	0.51
Medium Gaussian	10-fold CV	45.0	40.0	50.0	0.38
	LOOCV	17.5	5.0	30.0	0.03
Coarse Gaussian	10-fold CV	45.0	5.0	85.0	0.39
	LOOCV	0.0	0.0	0.0	0.00

Table 4.7 — *Classification performance for homogeneity indices, measured in accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the ROC curve (AUC). Results are obtained from 6 support vector machine (SVM) models. Validation was performed by 10-fold cross-validation and leave-one-out cross-validation (LOOCV).*

linear SVM obtains the highest ACC values, regardless of the amount of principal components or bins. The highest ACC (77.5%) is obtained by employing 26 bins, 20 (largest) principal components and 10-fold cross-validation. The resulting TPR, TNR, and AUC values are 80%, 75%, and 0.79, respectively. Results of the other SVMs obtained by employing 26 bins and 20 principal components are presented in Table 4.8. The obtained SVMs by employing 6 bins show similar results.

4.3.4 Discussion and conclusions on dose distribution experiments

The objective of these experiments was to investigate the possible influence of the GKRS dose distribution on the treatment response for vestibular schwannoma. For the binary classification problem in this research, treatment success and treatment failure need to be defined. The purpose of a GKRS treatment on VS is to

SVM type	Validation	ACC [%]	TPR [%]	TNR [%]	AUC
Linear	10-fold CV	77.5	80.0	75.0	0.79
	LOOCV	72.5	75.0	70.0	0.75
Quadratic	10-fold CV	67.5	75.0	60.0	0.77
	LOOCV	67.5	70.0	65.0	0.71
Cubic	10-fold CV	62.5	75.0	50.0	0.79
	LOOCV	65.0	75.0	55.0	0.73
Fine Gaussian	10-fold CV	55.0	45.0	65.0	0.59
	LOOCV	0.0	0.0	0.0	0.00
Medium Gaussian	10-fold CV	65.0	75.0	55.0	0.77
	LOOCV	62.5	75.0	50.0	0.72
Coarse Gaussian	10-fold CV	75.0	75.0	75.0	0.77
	LOOCV	0.0	0.0	0.0	0.00

Table 4.8 — *Classification performance for 3D-HOG features, measured in accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the ROC curve (AUC). Results are obtained from 6 support vector machine (SVM) models. Validation was performed by 10-fold cross-validation and leave-one-out cross-validation (LOOCV).*

stop tumor expansion. From this purpose, it is easy to define treatment failure. However, continuous or even recurring tumor growth can occur as late as 10 years following treatment [32]. As such, treatment success is not simply 'no failure', even after several years of follow-up. To this end, we have selected tumors that showed a significant volume reduction in the first year following GKRS. This definition of treatment success has enabled us to create two very dissimilar cohorts of each 20 patients, so that the experiments were technically well-defined. However, this definition causes a limitation on the actual evaluation of the dose-distribution influence, as it provides a bias in the data. Furthermore, since the treatment planning involves many settings that result in significant variations in the dose distribution, the number of patients included in these experiments may be too limited. Hence, the influence of the dose distribution needs to be investigated further on larger datasets, with a broader definition of treatment success.

Nevertheless, there seems to be a correlation between the actual heterogeneous dose distribution and the treatment outcome. The details in the 3D-HOG features suggest that there are measurable differences that could potentially influence the treatment efficacy, even though treatment plannings are considered uniform. These findings may provide a basis for refining towards personalized treatments and prediction of treatment efficacy, if positively validated on larger datasets.

4.4 Tumor delineation

Another important step in the treatment planning that may influence the GKRS outcome of VSs is found in the tumor delineation by the neurosurgeon. Therefore,

in this section, we will attempt to establish an explanation for tumor response by exploring the possible relation between the accuracy of tumor delineation by physicians and the tumor response derived from an imaging database. In short, we will investigate whether the accuracy of the tumor delineation plays a role in the tumor response. This section is split up in the following subsections. First, a background in assessing tumor delineation quality is provided in Section 4.4.1. Next, the experiments are introduced in Section 4.4.2, after which the results are presented in Section 4.4.3. Finally, Section 4.4.4 will discuss the obtained results and provide conclusions.

4.4.1 Background in tumor delineation

One of the first steps in creating a conformal treatment plan is the delineation of the tumor on the MRI scans. This segmentation is needed for calculating various treatment quality indices, such as the conformity and selectivity. Since the final segmentation of the tumor is a human-defined element which is employed for calculating the treatment quality indices, it is highly susceptible for inter-observer variations and inaccuracies, as already discussed in Section 2.5. Such deviations from the true tumor outline may lead to underexposure of parts of the tumor, which in turn could reduce the chance at treatment success. This is especially critical in Gamma Knife treatments because of the steep dose drop-off. Consequently, tumor tissue that was not included in the segmentation may receive significantly lower radiation doses than protocolized.

For evaluating delineations, metrics for image segmentation quality can be applied and modified if needed. To assess segmentation quality, many methods and algorithms have been introduced. Fenster *et al.* [197] grouped the necessities of medical image segmentation evaluation into accuracy, precision, and efficiency, where the latter is related to the practical use of the evaluated algorithm. This becomes important when segmentation should be performed in real time and hence, it is a measure of time consumption. For accuracy, the authors evaluated two quality aspects: the boundary and the size. The precision category evaluates the repeatability of a technique: variations in delineated tumors can be caused by subjective observer interactions with the segmentation method.

To evaluate the boundary and size, the metrics known as accuracy and precision are obvious, simple methods which can be used to assess the performance of a segmentation if the ground truth is available [198]. Furthermore, sensitivity and specificity are valuable, pixel-based metrics, which can be used to assess the quality of a segmentation. The sensitivity measures the percentage of true positives that are correctly segmented, whereas the specificity measures the percentage of true negatives that are correctly segmented.

In spite of the low complexity of the pixel-based metrics mentioned above, the main drawback of these methods is that they are highly correlated with the segment size. Metrics that are size invariant are the Jaccard index [199] and the Dice coefficient [200], which quantify the similarity between the segmentation and the ground truth, by measuring the area of each segmentation and their spatial

overlap. Taha *et al.* [201] showed that Dice and Jaccard are related. Their research investigated different metrics for evaluating 3D medical image segmentation, including the previously mentioned Dice and Jaccard indices, but also a number of other options. Furthermore, they used the notion that medical segmentation can be seen as fuzzy, meaning that voxels have a grading of membership in the unity interval. This is the case when the underlying segmentation is the result of averaging different delineations of the same structure annotated by different operators. This is useful for obtaining binary representations which can be evaluated as ground truth. Alternatively, sweet-spot training can also be applied for estimating a ground truth, even when annotations vary significantly between multiple experts [139]. It is comparable to the Jaccard index, where multiple annotations are compared instead of only two.

Shi *et al.* [202] described various types of segmentation errors, based on four basic error types: (1) added regions, (2) added background, (3) inside holes, and (4) border holes. The described segmentation errors are evaluated using the following similarity measures: (1) quantity, i.e. the number of segmented objects, (2) area accuracy, (3) contour similarity, and (4) the content, i.e. the existence of inside holes and boundary holes in the segmented region.

With respect to the vestibular schwannoma and the Gamma Knife treatment, the evaluation of tumor delineation accuracy has not yet been investigated. Because the treatment planning is reviewed by multiple specialists prior to executing the treatment, each segmentation can be considered highly accurate. Nevertheless, the small variations at the tumor boundary may explain the differences in the treatment response. Since a ground-truth segmentation is not available in the case of VS treated with GKRS, metrics such as precision, sensitivity, and specificity are not suited for the evaluation of the treatment delineation in this work. Here, a comparison is made between two different annotations using metrics based on the first two categories proposed by Fenster *et al.* [197], and on the alignment measures as proposed by Taha *et al.* [201]. Furthermore, since MRI slices can have varying thickness, a novel metric is proposed that considers differences at a specific normalized axial height of the tumor.

4.4.2 Tumor delineation experimental setup

In this section, a description of the proposed experiments is provided. To evaluate the differences between tumor delineations, the following metrics are examined: (1) variation in number of annotated slices, (2) deviations in segmentation area per slice, and (3) Jaccard indices, which provide a measure for delineation similarity. The Jaccard index is selected because it can directly compare two delineations. The differences will be tested for significance by using (1) Student's t-tests or rank-sum tests, and (2) visual evaluation of a notched box-plot, if applicable. In the following subsections, the employed dataset is presented, followed by the detailed descriptions of each utilized similarity metric.

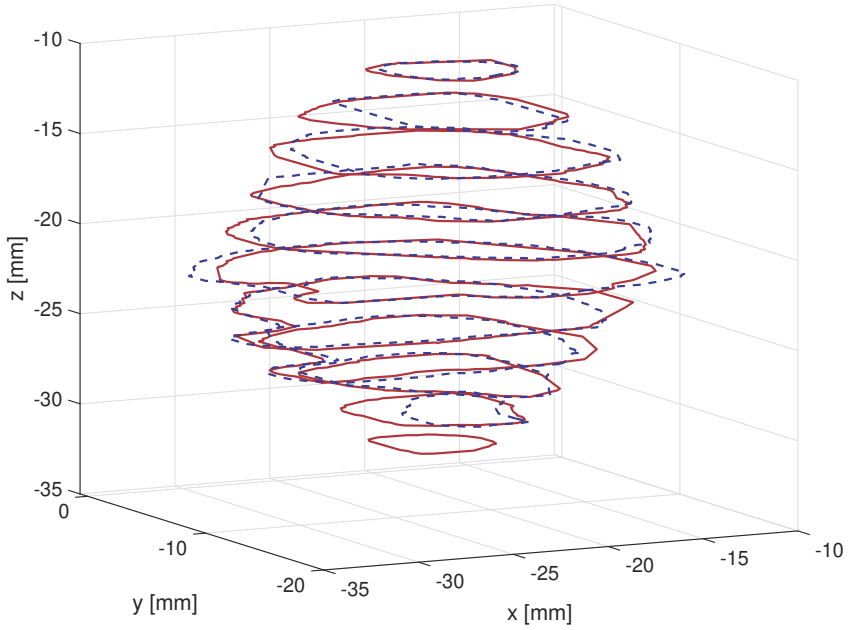


Figure 4.8 — *Example of the treatment delineation (red solid line) and the second delineation by the VS specialist (blue dashed line) in the planes at specific axial (z-axis) distances. The values on the x-, y-, and z-axes are given by the patient coordinate system defined by the MRI scanner.*

A. Dataset

The tumor delineation experiments focus on evaluating annotations of VSs, which are used for prescribing the protocolized dose with high accuracy to the tumor. Using the same dataset as described in Section 4.3.3, two groups of VS tumors with a significantly different response to the Gamma Knife treatment are available. Since the true tumor segmentation (ground truth) is unknown, one of the neurosurgeons who is specialized in treating VS tumors was tasked to create a second segmentation of the VS tumors in that dataset. He was blinded to the annotation created at time of treatment and had no knowledge of the GKRS response, in order to prevent bias. Furthermore, the amount of time required for the tumor segmentation was unrestricted, such that the delineation could be optimized. As a visual example, a pair of annotations is shown in Figure 4.8. The delineations are drawn on parallel, equidistant T1-weighted, contrast-enhanced MRI scans in axial directions. Furthermore, T1-weighted and T2-weighted MRI scans were available as resource for creating the segmentation. The resulting dataset contains 40 delineations created during treatment planning and another 40 created several years after the treatment was performed. A total of 401 MRI slices were employed for segmentation.

B. Volumes

The tumor volumes of both the treatment delineation and the revisited delineation were calculated using the treatment planning software (GammaPlan version 10, Elekta AB, Stockholm, Sweden). Both volumes are subtracted for each patient. A large difference in volume represents a less accurate delineation, with respect to the revisited segmentation. However, a small volume difference cannot directly represent a high accuracy, because two delineations from two completely different objects can have the same volume. Therefore, the assumption is made that all delineations are closely aligned to the actual tumor. As such, the volume difference can provide an estimate for the delineation performance.

C. Number of slices

The number of slices can differ between both delineations. This happens if one of the operators considers some of the voxels in MRI slices just above the top or just below the bottom of the tumor as part of the tumor, while another operator does not, like at the bottom part of the tumor in Figure 4.8. This provides information about the differences, targeting the top and bottom parts of the tumor, between the two annotations. By comparing the difference in the number of slices between the two patient cohorts, we can verify if the top or the bottom part of the tumor is the sensitive part that may influence the radiosurgical response.

D. Slice area

After finding the differences between the complete 3D tumor delineations, each tumor annotation is compared on a slice-by-slice basis. To this end, the annotation slices are converted into high-resolution binary image masks (pixels within the tumor contour are assigned unity value). Next, the slice area is calculated by counting the unity pixels. Furthermore, a novel metric is implemented in which the area differences are computed for each slice at a specific, normalized axial height. This height is determined by employing the following equation:

$$H_{\text{norm}}(j) = \frac{z(j) - z_{\text{min}}}{z_{\text{max}} - z_{\text{min}}}, \tag{4.9}$$

where H_{norm} denotes the normalized height position, j is the slice number, $z(j)$ is the j-th slice position in millimeters with respect to the patient coordinate system defined by the treatment planning software, and z_{min} and z_{max} are the minimum and maximum z-values per patient among both delineations, respectively. The area difference is determined by subtracting the area of the treatment delineation from the area of the revisited delineation. For the two cohorts, the summed absolute pixel (SAP) difference and the summed relative area (SRA) difference at a specific normalized height are computed and plotted. The SAP is calculated as $\sum_j |S_{\text{treat}}(j) - S_{\text{revisited}}(j)|$, where j is the slice index and S_{treat} and $S_{\text{revisited}}$ are the masks of the treatment delineation and the revisited delineation, respectively. The calculation for the SRA difference is similar, but now with respect to the individual surface areas of the delineations. This metric assists in investigating whether there

Metric		Failure	Success	p-Value
Volume [mm^3]	mean vol_{treat}	2538.05	3687.75	0.08
	mean $vol_{\text{revisited}}$	2490.80	3436.10	
	mean vol_{diff}	47.25	251.65	
Total slice difference	bottom	3	9	-
	top	6	6	
No. of patients with slice diff.	bottom	3/20	5/20	-
	top	6/20	6/20	
Jaccard index	mean $\pm$ std	0.792 ± 0.252	0.749 ± 0.260	<0.01

Table 4.9 — *Comparison of all metrics between the two delineation sets.*

is a crucial part in the tumor delineation that should be accurately delineated.

E. Slice similarity

Similar to the comparison of the slice area, the slice alignment is compared using the Jaccard index. This index is calculated by exploiting the same binary image masks as in the previous metric. From these image masks, the union and intersection of the two different masks per slice are computed and employed in the formula for the Jaccard index, which is given by:

$$J(j) = \frac{|S_{\text{treat}}(j) \cap S_{\text{revisited}}(j)|}{|S_{\text{treat}}(j) \cup S_{\text{revisited}}(j)|}, \tag{4.10}$$

where J is the Jaccard index for slice j. This formula results in a Jaccard index value within the unity interval, where zero means no intersection at all and unity represents two completely overlapping shapes. If either mask is empty, there is zero overlap and hence the resulting Jaccard index will be zero as well.

4.4.3 Tumor delineation results

In this section, the results obtained from the experiments are presented. These results are highlighted for each of the individual metrics in separate subsections.

A. Tumor volume differences

The results of the volume measurements are listed in Table 4.9. The volume of the treatment delineation is denoted as vol_{treat}, whereas the volume of the revisited delineation is denoted as $vol_{\text{revisited}}$. The volume difference vol_{diff} is calculated by subtracting vol_{treat} from $vol_{\text{revisited}}$. From the results, it can be observed that the mean volume difference for the successfully treated cohort is larger than the mean difference for the failed treatment cohort. However, this difference is not statistically significant according to the *t*-test (p = 0.08). This can also be distilled from the box-plot depicted in Figure 4.9, where the notches of both plots overlap. Nevertheless, there may be a trend present due to the p-value being close to the

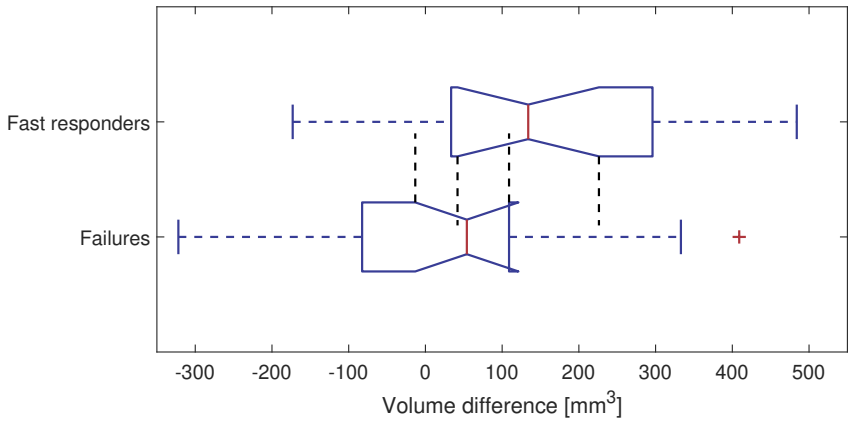


Figure 4.9 — *Boxplots of the tumor volume differences. The vertical black dashed lines indicate notch range borders to improve the visual inspection of the boxplots. Here, the notches clearly overlap. It can therefore be concluded that the differences in tumor volumes between the two sets are not statistically significant.*

significance threshold of p = 0.05.

B. Number of segmentation slices

From the results given in Table 4.9, it is observed that the number of slices in the successful cohort, i.e. fast responders, and the failure cohort are quite similar. Most of the possible differences are in fact equal to zero, as can be seen from the number of patients with differing slice numbers. Compared to the revisited delineation, it is interesting that the slice number differences at the bottom of the tumors are always positive in the fast-responding cohort, whereas they are negative in the failed cohort. Because there are only a few cases that show this difference, the effect of such a difference on the treatment outcome can be most likely neglected. This statement also holds for the top parts of the tumors, where the differences are typically small. It is worthwhile to note that in some cases, at the bottom part of the tumor, the difference in number of slices may sometimes be larger than a single slice. This discrepancy is typically caused by a difference in interpretation by the various neurosurgeons. It is sometimes unclear where the tumor boundaries are situated.

C. Slice area

As a pre-processing step required for the slice-area metric and the slice-similarity metric (discussed in the next subsection), the tumor delineations are converted into high-resolution binary image masks with a resolution of 2,048 $\times$ 2,048 pixels. An example of a pair of tumor delineation masks is shown in Figure 4.10.

Furthermore, the height position of each slice is normalized using Eq. (4.10). After adding each slice-area difference at the normalized height position, the left

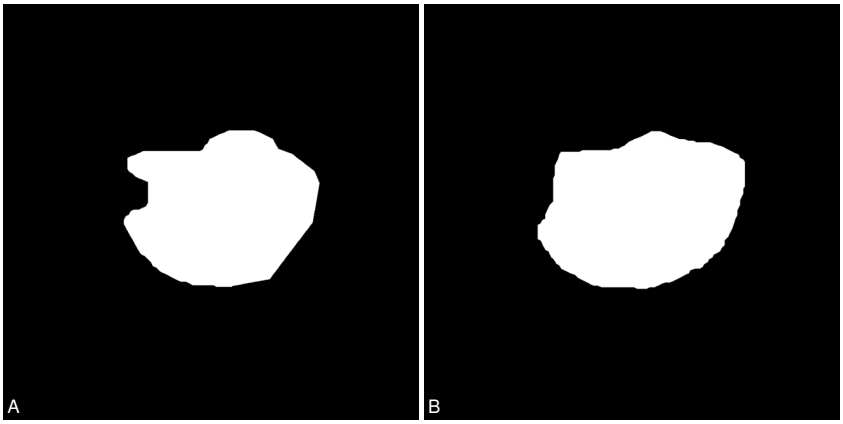


Figure 4.10 — *Example of two binary image masks. (A) Mask from the treatment delineation. (B) Mask of the same MRI slice, created by the VS specialist.*

graph in Figure 4.11 is obtained. It shows the SAP differences between the treatment delineation and the revisited delineation for all patients in the two cohorts. From this plot, we observe that the SAP difference varies across the complete normalized height, and that the SAP differences are larger for the successfully treated cohort. At the right of Figure 4.11, the SRA curves are presented. In this graph, significant differences appear at the top and bottom parts of the tumors, while in the rest of the tumor these differences are small. In this plot it can be noticed also that the SRA differences for the successfully treated cohort are overall slightly larger. This can also be measured from the average difference in the number of pixels per slice, which are 12,000 $\pm$ 28,926 and 44,859 $\pm$ 118,797 for the failure and successful cohort, respectively. With respect to the slice area itself, these average difference results become 4.4%$\pm$ 29.4% and 11.7%$\pm$ 31.3%. These variations are significantly different when tested using a *t*-test (p = 0.02). Indeed, the boxplots in Figures 4.12 and 4.13 show no overlap for the SAP and SRA values, respectively, thereby visually supporting the statistical significance found in the *t*-tests.

D. Slice similarity

Table 4.9 presents the summarized statistical results of the Jaccard indices for both cohorts. The mean Jaccard value of the failure cohort is higher than the mean value of the successfully treated cohort. This means that the similarity of the treatment delineation is higher in the failure cohort, compared to the revisited delineation. In other words, the treatment delineation of the failure cohort has a higher accuracy. The distribution of the indices shows a slightly higher skewness for the failure cohort when compared to the successfully treated cohort, being -2.04 and -1.86, respectively. The boxplots shown in Figure 4.14 depict the distributions of the Jaccard indices for the two cohorts. It can be derived from the boxplots

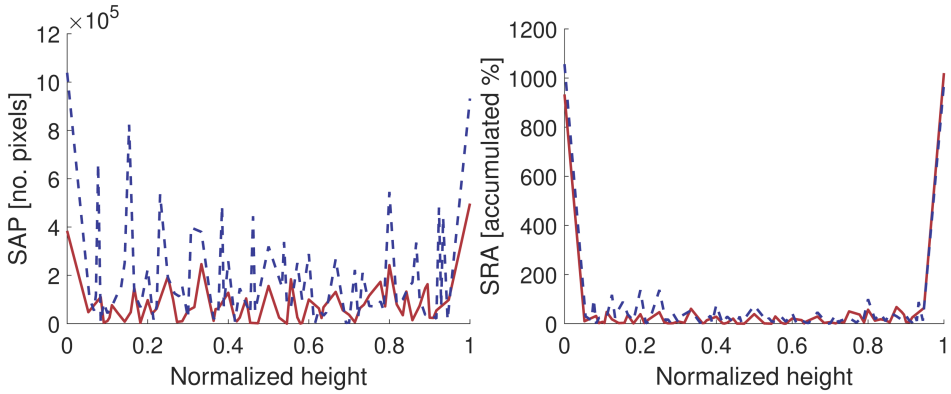


Figure 4.11 — *Results of the slice area metrics. Left: SAP differences of the treatment delineation and the delineation of the specialist. Right: the SRA differences of both delineations. In both graphs, the red solid line denotes the differences in the failure cohort, whereas the blue dashed line depicts the differences in the fast-responding cohort.*

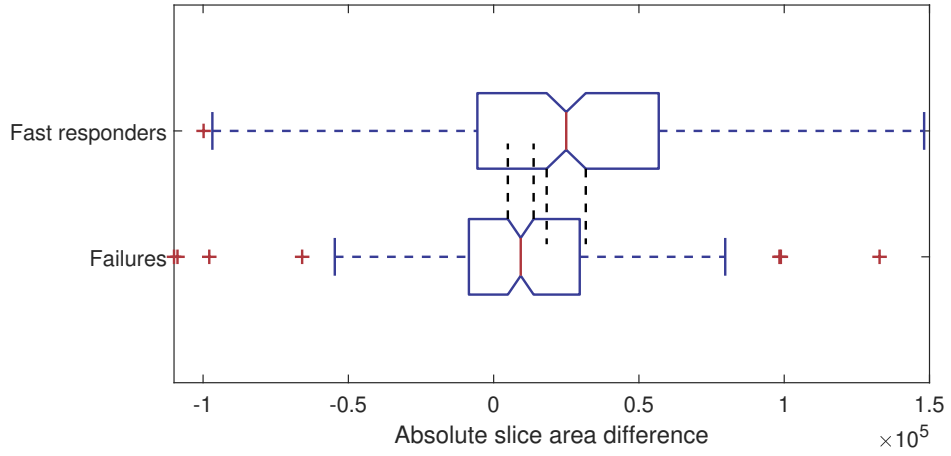


Figure 4.12 — *Boxplots of the relative area differences. The vertical black dashed lines emphasize the notch range boundaries to improve the visual inspection of the boxplot. It becomes clear that there is no overlap in the notches, suggesting that the differences in slice areas between the two delineation sets are statistically significant.*

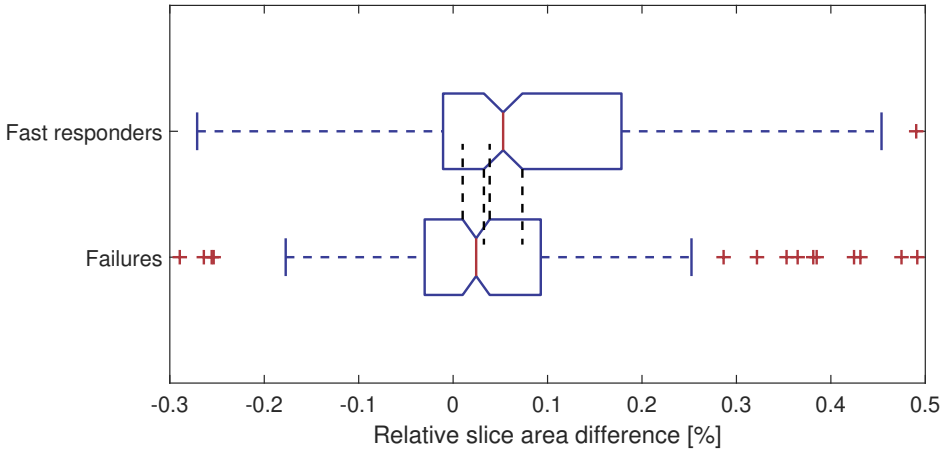


Figure 4.13 — *Boxplots of the relative area differences. The vertical black dashed lines emphasize the notch range boundaries to improve the visual inspection of the boxplot. It is visible that there is no overlap in the notches, suggesting that the relative differences in slice areas between the two delineation sets are statistically significant.*

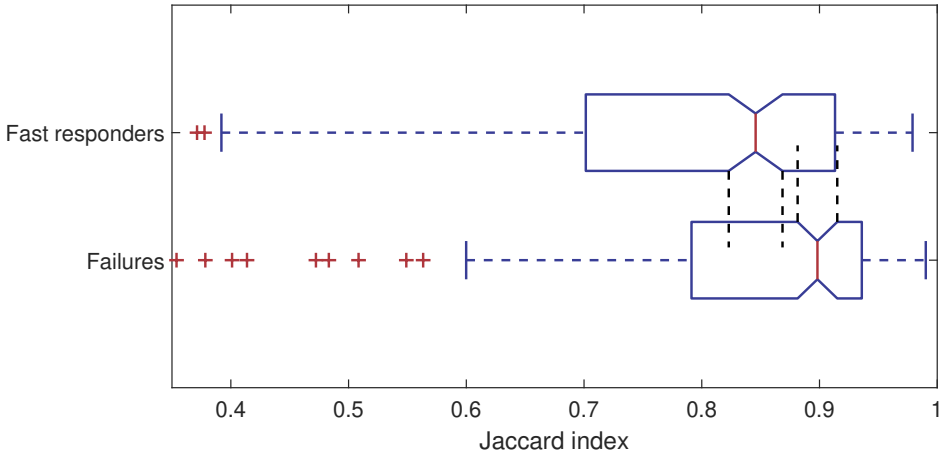


Figure 4.14 — *Boxplots of the slice similarity differences, calculated by the Jaccard indices. The vertical black dashed lines emphasize the notch range boundaries to improve the visual inspection of the boxplot. Here, the notches do not overlap, thereby implying that the differences in slice similarities between the two delineation sets are statistically significant.*

that the notches do not overlap. Therefore, the difference in Jaccard indices can be considered statistically significant. This is also the conclusion after the rank-sum test, which results in a p-value of less than 0.01, thereby confirming the strong dissimilarity between both cohorts.

4.4.4 Discussion and conclusions on tumor delineation experiments

In this section, the correlation between the GKRS treatment response of VS tumors and the tumor delineation accuracy during treatment planning have been investigated. This is performed by implementing well-known metrics for assessing the quality of medical image segmentation and by introducing novel SAP and SRA difference metrics. We have compared the treatment delineation with the annotation of a neurosurgeon who is specialized in treating VS tumors, which were created retrospectively, using volumes, number of slices, slice areas, and slice similarities. The obtained results show no evidence to the hypothesis that a less accurate delineation leads to an increased failure rate. The results illustrate that the delineation of the fast-responding group is less accurate compared to the delineation of the VS specialist, rather than the delineation of the failure cohort, which is surprising. This suggests that when there is more disagreement between the two tumor delineations, the more likely it is to achieve treatment success within this small dataset, where treatment success is considered to be a rapid decrease in tumor volume within a limited time frame. This leads to the conjecture that the type of tumors that are more difficult to delineate are more likely to show a positive treatment response.

Because of the limited framework of the current experiment, e.g. a quite limited dataset, biased selection of subjects for simple binary classification, and a lack of multi-variate analysis, the drawn conclusion can only be considered as the results of a technical experiment, rather than a justified relevant medical conclusion. However, the obtained results suggest that a tumor that is apparently more difficult to delineate, responds better to the GKRS treatment. This is an interesting topic for further research, where the shape and texture of the tumor on the MRI scans, which may represent the underlying biology of the tissue, should be incorporated.

4.5 Conclusions

Currently, many research articles have discussed the influence of the Gamma Knife treatment parameters on the radiosurgical response of vestibular schwannomas. Many authors have concluded that the different dose-related factors do not impact the volumetric response, although some investigations show contradicting results. In our unique large database from a single institution, we have investigated the treatment parameters on the long-term tumor control. The performed Cox regression analysis reveals that the dose to the tumor does significantly correlate to the long-term tumor control, with a resulting risk factor of 0.63 ($p = 0.03$), implying that lowering the dose will result in higher risks at treatment failure. However, analyzing the variation in long-term tumor control between multiple sub-cohorts, split by using the dose to the tumor margin, did not reach statistical significance. Therefore, it can be concluded that even though there may be an influence, it is limited within the boundaries of the treatment protocols.

Since Gamma Knife radiosurgery involves a combination of multiple radiation

shots at different shapes and weights, global parameters may oversimplify the actual underlying dose distribution. Therefore, we have introduced a novel method for evaluating the resulting heterogeneous dose distribution and its impact on the treatment response. Calculated homogeneity indices presented no statistical differences within our data. Our novel method, employing three-dimensional histogram of oriented gradients (3D-HOG) to describe spatial differences between treatment plans, resulted in a classification accuracy, true positive rate, true negative rate, and area under the curve of 77.5%, 80.0%, 75.0%, and 0.79, respectively. These values are based on exploiting linear support vector machines for classification. Although these metric results are attractive, they need to be more extensively analyzed on larger datasets to validate their impact on the treatment response.

Finally, we have investigated another important aspect that is generally not considered: the accuracy of the tumor delineation during treatment planning and its possible influence on the radiosurgical response. Since the dose is prescribed to the tumor margin, inaccuracies in the tumor annotations may lead to underexposure of the tumor margins. As such, we have conducted experiments to determine whether the differences between annotations may have influenced the treatment response. To this end, we have compared the treatment delineation with the annotation by a VS specialist, created retrospectively, using volumes, number of slices, slice areas, and slice similarities. By calculating the Jaccard indices, it is illustrated that the delineations of the fast-responding group have more variation compared to the delineations of the VS specialist. In contrast, the delineations of the failure cohort show less variations. This suggests that tumors that are more difficult to segment, are expected to obtain a positive treatment outcome.

In this thesis, it is suggested that the underlying intrinsic tumor biology is the main cause for the variations in the Gamma Knife treatment response. In this point of view, the previously described results on the impact of tumor delineations are interesting. Therefore, the following chapter will further research this suggestion, where the influence of the tumor shape and appearance on the MRI scans are investigated by means of quantitative shape and texture features. We will examine the shape of the tumor to evaluate whether this parameter enables the prediction of the treatment response. Furthermore, the first experiments on the MRI tumor texture are conducted, in order to explore whether differences in the calculated quantitative tumor texture features facilitate treatment response prediction. If possible, this will ultimately lead to the ability of predicting the vestibular schwannoma response to Gamma Knife radiosurgery prior to treatment, aiding physicians and their patients in selecting the optimal treatment strategy on an individual basis.

Chapter 5

Shape and texture of the vestibular schwannoma

5.1 Introduction

The previous chapter explored the impact of various treatment-related parameters on the response to Gamma Knife radiosurgery of vestibular schwannoma. In current state of the art, only global treatment parameters have been investigated, and their impact on the treatment response is reported to be limited. Indeed, using the unique Tilburg database of VS patients, we have determined that the correlation is minimal and significantly related to the employed methodology. Furthermore, we have introduced a novel method for evaluating the heterogeneous character of the planned dose and determined that the resulting metrics show correlation to the treatment outcome. Finally, the previous chapter has also investigated the impact of the tumor delineation accuracy on the treatment response. The obtained results highlight that inter-operator differences are significantly larger in the fast responding group compared to the group where treatment failed. This leads to the conjecture that the type of tumors that are more difficult to delineate, are more likely to obtain a positive treatment response. Consequently, it is hypothesized that the shape and appearance of the tumor on MRI scans may have impact on the treatment response.

In this chapter we therefore examine the predictive value of the tumor shape and textural appearance on MRI. It is speculated that these parameters are related to the underlying tumor biology. Since the MRI appearance of VS tumors is highly variable, calculated MRI features may reflect specific differences in tumor biology, which in turn can explain the variations in treatment response, as discussed in Chapter 1. Some examples of the VS appearance on MRI are found in Figure 5.1.

MRI pattern scoring and diffusion parameters

Currently, most of the studies involving MRI characteristics of VSs employ scoring of the contrast-enhanced MRI patterns, e.g. homogeneous versus heterogeneous and cystic versus non-cystic tumor classification. Frisch *et al.* [46] explored the outcomes after stereotactic radiosurgical treatment of predominantly cystic VS, and determined that these tumors tend to have a larger size reduction compared to solid tumors. Bowden *et al.* [50] confirmed in their study that tumor volume regression was most evident in patients with cystic tumors.

Other studies have also employed apparent diffusion coefficients (ADC), in

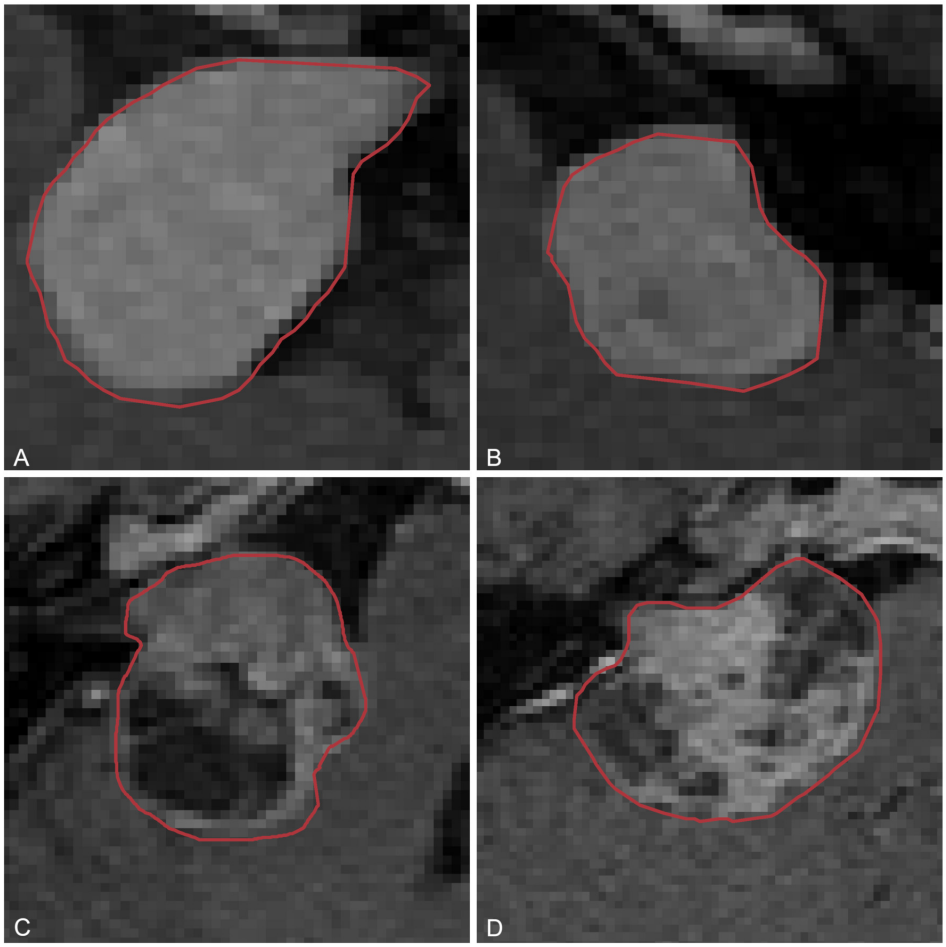


Figure 5.1 — *Examples of the MRI appearance of vestibular schwannoma. From these four slices, it can be verified that shape and texture can differ significantly.*

addition to scoring of MRI patterns. These coefficients can reflect differences in tumor biology and are calculated using MRI with diffusion-weighted imaging. High ADC values usually indicate sparse cellularity, necrosis, or cystic features, because it is a measure of the magnitude of diffusion of water molecules within tissue. Camargo *et al.* [51] determined that higher minimum ADC values are associated with non-responding tumors. Using an ADC cut-off value of 800 $\mu m^2/s$, they distinguished non-responders from responders in 18 out of 20 cases, with sensitivity and specificity of 77.8% and 100%, respectively. Furthermore, the authors determined a significant correlation between the minimum ADC value and the percentage of tumor-size reduction. They also explored other qualitative image features of the tumors, such as homogeneous versus heterogeneous and cystic versus solid tumors. However, these qualitative factors did not show any predictive value of the radiation response. Wu *et al.* [53] determined that the mean

of all maximum pre-radiosurgical ADC values of VSs was significantly higher for those with tumor regression or stabilization at last follow-up, compared to those with progression. They determined an ADC cut-off value of 1274 $\mu m^2/s$ for the pre-radiosurgical ADC value to generate the optimal combination of sensitivity (69.2%) and specificity (70%). They also found that enhancing patterns, dichotomized into homogeneous and heterogeneous enhancement, at the time of GKRS did not significantly correlate with the volume reduction ratio at last follow-up. However, the authors did discover that tumors with cysts, making up at least one-third of the whole tumor volume, were more likely to regress or stabilize and also had greater volume-reduction ratios at last follow-up, when compared to non-cystic tumors.

Radiomics in medical imaging

Although diffusion-weighted imaging adds functional information to the largely anatomical data gathered by the conventional MRI sequences, differences in tumor biology can also be determined by these conventional MRI scans. Texture features obtained from these MRI scans have shown good results in the medical field for computer-aided diagnosis (CAD) and the creation of prediction models. Wibmer *et al.* [78] presented a method for detecting cancerous tissue in prostates, applying Haralick features on MRI images. They determined that, next to ADC values, five Haralick-based texture features were significantly different between cancerous and non-cancerous tissue. Chaddad *et al.* [79] proposed a novel model to characterize glioblastoma multiforme tissue phenotypes using MRI and gray-level co-occurrence matrices in three anatomical planes, reaching an accuracy, sensitivity, and specificity of 88.14%, 85.37%, and 96.1%, respectively. Zhou *et al.* [80] employed mean intensity and gray-level non-uniformity (GLN) based on run-length matrices (RLM) of contrast-enhanced MRI scans, to predict histological grading of hepatocellular carcinomas. Their model, using only the mean intensity value, resulted in an optimal sensitivity of 76% and specificity of 100%, whereas the model employing the GLN in four different directions led to comparable or even higher rates.

Furthermore, in the field of medical image analysis, several experiments have been conducted in regard to relating shape to tumor classification. Soltanian-Zadeh *et al.* [74] compared multi-wavelet, wavelet, Haralick, and shape descriptors for micro-calcification classification in mammograms. They found that their shape descriptors were superior to the wavelet and Haralick features. Alvarenga *et al.* [75] used so-called *morphometric parameters* obtained from the normalized radial length, i.e. the normalized Euclidean distances between the centroid and all points on the boundary, and trained a multi-layer perceptron classifier to obtain results in terms of sensitivity and specificity of around 90% in classifying breast tumors. Boujelben *et al.* [76] used this research as a basis for their own experiments, and added the index angle and convexity descriptors to their k-nearest neighbors analyses. They found improved results compared to the work of Alvarenga *et al*. Furthermore, Czarnek *et al.* employed algorithmic three-dimensional analysis of

tumor shape in MRI to improve prognosis of survival in glioblastoma. Employing five features that quantify the extent of irregularity in tumor shape in two and three dimensions, they concluded that the tumor shape is statistically prognostic of survival for patients with glioblastoma multiforme. In a machine learning system, Zacharaki *et al.* [77] implemented tumor shape and intensity characteristics, as well as rotation-invariant texture features, to distinguish different types of brain tumors. They obtained accuracy, sensitivity, and specificity scores of 85%, 87%, and 79%, respectively, for discriminating metastases from gliomas, and 88%, 85%, and 96% for discerning high-grade from low-grade neoplasms.

Radiomics for vestibular schwannomas

However, for VS tumors, only one study has investigated quantitative tumor texture features from conventional MRI in relation to the Gamma Knife treatment response. Speckter *et al.* [59] determined in a cohort of 14 progressors and 9 regressors that first-order statistical texture features, such as the mean and standard deviation, do not show significant differences between both treatment response groups. Nevertheless, using a separator value, which is based on the lower quartile of the kurtosis value obtained from the T2-weighted MRI scans of the progressors cohort, they achieved a sensitivity to predict progression of 71% and a specificity of 78% within the complete group, resulting in a positive predictive value of 86% and a negative predictive value of 59%. However, their classification labels of regressor and progressor are based on the response within the first 18 months following GKRS treatment. As such, their model does not differentiate between transient and permanent progression, which is crucial for predicting the actual treatment response and the corresponding possible consequences. Nevertheless, their results do show the potential of employing the tumor appearance on MRI for treatment response prediction.

Therefore, in this chapter, the differences in tumor appearance on the conventional MRI scans are investigated. The correlations between the GKRS treatment outcome and the quantitative shape and MRI texture characteristics are explored, with the purpose to predict the effectiveness of GKRS treatment for each individual VS patient. If successful, the benefit for patients is that they can be well-informed for choosing the best treatment option, together with their treating physicians. Additionally, it can also potentially provide a basis for an individualized follow-up protocol, reducing the overall number of follow-up hospital visits and MRI scans. As a refinement of the general problem statement, we list the following aspects.

- *Data selection*: For the training of a prediction model, clear and objective classification labels and data are needed. Therefore, two distinct patient cohorts are selected to evaluate the impact of the tumor shape and appearance on the treatment response.
- *Features*: Numerous *radiomic* features, quantitatively describing tumor shape and tumor texture, are evaluated in a machine learning environment for

their impact on the treatment response.

- *Response prediction*: Experiments are performed to determine which shape and texture descriptors are most suited for treatment response prediction. Using these features, the possibility to train a model that *a-priori* predicts the treatment response is investigated. If possible, this model will lead to an enhanced treatment selection strategy and ultimately to an improved overall treatment efficacy.

This chapter is outlined as follows. First, in Section 5.2, the employed methods are discussed, followed by the experimental setup in Section 5.3. Next, Sections 5.4 and 5.5 elaborate on the obtained results using shape descriptors and texture features, respectively. In Section 5.6, the obtained results are discussed and limitations of these experiments are highlighted. Finally, several conclusions to these experiments are given in Section 5.7.

5.2 Methods for shape and texture feature extraction

5.2.1 Shape descriptors

In pattern recognition and computer vision, many contour-based shape descriptors are used. These descriptors can be divided into two groups: two-dimensional (2D) shape descriptors, and three-dimensional (3D) shape descriptors. In the following subsections, the shape features employed in the experiments of this chapter are described.

A. Two-dimensional descriptors

First, several 2D shape descriptors are considered. These are calculated on each individual contour drawn on a single MRI slice. The formulas used for computing the shape features are as follows.

Consider the contour as a non-self-intersecting polygon defined by N ordered points (x_0, y_0), (x_1, y_1), $\cdots$, (x_{N-1}, y_{N_1}), where x_n and y_n are the x- and y-coordinates in the MRI slice. Each contour has a centroid with coordinates (C_x, C_y), which can be calculated using the following equations:

$$C_x = \frac{1}{6A} \sum_{n=0}^{N-1} (x_n + x_{n+1}) (x_n y_{n+1} - x_{n+1} y_n), \text{ and} \tag{5.1}$$

$$C_y = \frac{1}{6A} \sum_{n=0}^{N-1} (y_n + y_{n+1}) (x_n y_{n+1} - x_{n+1} y_n), \tag{5.2}$$

where A is the area of the contour (Eq. (5.6)), $x_N = x_0$, and $y_N = y_0$. With the centroid known, the radial length RL and the normalized radial length RL_{norm} of each point n on the contour can be calculated using the following equations,

respectively:

$$RL\,(n) = \sqrt{(x_n - C_x)^2 + (y_n - C_y)^2}, \text{ and} \tag{5.3}$$

$$RL_{\text{norm}}\,(n) = \frac{RL\,(n)}{\max\,(RL)}. \tag{5.4}$$

With these four equations, the following 19 2D shape descriptors can be computed:

- *Perimeter length* This is specified by:

$$P = \sum_{n=0}^{N-1} \sqrt{(x_{n+1} - x_n)^2 + (y_{n+1} - y_n)^2}, \tag{5.5}$$

  where $x_N = x_0$, and $y_N = y_0$.

- *Area size* This parameter is computed as:

$$A = \frac{1}{2}\left|\sum_{n=0}^{N-1} (x_n y_{n+1} - x_{n+1} y_n)\right|, \tag{5.6}$$

  where $x_N = x_0$, and $y_N = y_0$.

- *Mean normalized radial length* This feature is calculated by:

$$\mu_{RL_{\text{norm}}} = \frac{1}{N}\sum_{n=0}^{N-1} RL_{\text{norm}}\,(n)\,. \tag{5.7}$$

- *Area ratio* The area ratio computes the percentage of the tumor outside the circular region defined by the mean normalized radial length, by using the equation of [75] as follows:

$$AR = \frac{1}{N \cdot \mu_{RL_{\text{norm}}}} \sum_{n=0}^{N-1} \left(RL_{\text{norm}}\,(n) - \mu_{RL_{\text{norm}}}\right), \tag{5.8}$$

  where $RL_{\text{norm}}\,(n) - \mu_{RL_{\text{norm}}} = 0$ for $\forall\, RL_{\text{norm}}\,(n) \leq \mu_{RL_{\text{norm}}}$.

- *Roughness index* The roughness index is a measure for the average distance between neighboring contour points over the entire contour. This is computed using the equation of [75] as follows:

$$RI = \frac{1}{N}\sum_{n=0}^{N-1} \left|RL_{\text{norm}}\,(n) - RL_{\text{norm}}\,(n+1)\right|, \tag{5.9}$$

  where $RL_{\text{norm}}\,(N) = RL_{\text{norm}}\,(0)$.

- *Normalized shape compactness* [203] This shape characteristic is determined using:

$$P2A_{\text{norm}} = \frac{P^2}{4\pi A}. \tag{5.10}$$

- *Haralick's circularity measure* [204] This descriptor is computed by:

$$HC = \frac{\mu_{RL_{\text{norm}}}}{\sigma_{RL}}, \tag{5.11}$$

where σ_{RL} is the standard deviation of all radial lengths $RL\,(n)$.

- *Danielssons's shape factor* [205] This factor is derived by computing:

$$G = \frac{A}{9\pi\mu^2_{RL_{\text{norm}}}}. \tag{5.12}$$

- *Bribiesca's normalized discrete compactness* [206] This parameter is calculated using:

$$C_{D,\text{norm}} = \frac{C_D - C_{D,\text{min}}}{C_{D,\text{max}} - C_{D,\text{min}}}, \tag{5.13}$$

where $C_D = (4A - P)/2$ is the perimeter of contact, and the bottom and top limits of the perimeter of contact are defined by $C_{D,\text{min}} = A - 1$ and $C_{D,\text{max}} = (4A - 4\sqrt{A})/2$, respectively.

- *Eccentricity* This feature is computed by:

$$\epsilon = \frac{\left(\mu_{2,0} - \mu_{0,2}\right)^2 - 4\mu^2_{1,1}}{\left(\mu_{2,0} + \mu_{0,2}\right)^2}, \tag{5.14}$$

where $\mu_{1,1}$, $\mu_{2,0}$, and $\mu_{0,2}$ are second-order central moments.

- *Convexity* [76] This variable is derived by:

$$CVX = \frac{P}{P_{\text{cvx}}}, \tag{5.15}$$

where P_{cvx} is the perimeter of the convex hull of the contour points.

- *Fourier descriptor coefficients* [207] First, a complex number $u\,(n) = x_n + iy_n$ for $n = 0, 1, \ldots, N-1$ is defined, with (x_n,y_n) the contour coordinates. For translation invariance, the difference of the n-th contour point and the centroid is taken, i.e. $\overline{u}\,(n) = x_n - C_x + i\,(y_n - C_y)$. Next, the Fourier transform of $\overline{u}\,(n)$ is calculated. For scaling invariance, the magnitude of $F\,[k]$ is divided by the magnitude of the first Fourier value, i.e. $|F\,[0]|$, to obtain $\overline{F}\,[k]$. Finally, the second to the fifth Fourier coefficient are taken as shape descriptors, i.e. $\overline{F}\,[1], \ldots, \overline{F}\,[4]$.

- *Contour sequence moments features* For calculating the contour sequence moments (CSMs) m_r and central CSMs μ_r, the following equations from [208] are needed, respectively:

$$m_r = \frac{1}{N}\sum_{n=0}^{N-1} RL^r(n), \text{ and} \tag{5.16}$$

$$\mu_r = \frac{1}{N}\sum_{n=0}^{N-1} \left(RL(n) - m_1\right)^r. \tag{5.17}$$

These functions can be made translation-, rotation-, and scale-invariant by normalization, which leads to:

$$\overline{m}_r = \frac{m_r}{\mu_2^{r/2}}, \text{ and} \tag{5.18}$$

$$\overline{\mu}_r = \frac{\mu_r}{\mu_2^{r/2}}. \tag{5.19}$$

While these can be used directly as features for shape classification, less noise-sensitive results can be obtained from the following features:

$$CSM_1 = \frac{\sqrt{\mu_2}}{m_1}, \tag{5.20}$$

$$CSM_2 = \frac{\mu_3}{\mu_2^{3/2}}, \tag{5.21}$$

$$CSM_3 = \frac{\mu_4}{\mu_2^2}, \tag{5.22}$$

$$CSM_4 = \frac{\mu_5}{\mu_2^{5/2}}. \tag{5.23}$$

The first three CSMs can be viewed as the normalized amplitude variation, the coefficient of skewness, and the coefficient of kurtosis, respectively.

The above-specified shape descriptors can be also employed as feature per tumor. This is done by calculating the mean, standard deviation, skewness, and mean absolute deviation over all contours per tumor.

B. Three-dimensional descriptors

Since the extracted contours are 2D, they need to be transformed to 3D shapes in order to determine the 3D shape descriptors. To this end, MATLAB is used to create 3D alpha shapes. Employing these shapes, the following six 3D shape descriptors are computed:

- *Approximated volume* The approximate volume of the 3D shape is calculated by MATLAB's AlphaShape tool. This volume is denoted by V.

- *Surface area* The surface area is calculated by MATLAB using the AlphaShape tool. This surface is denoted as A_{surf}.
- *Surface-to-volume ratio* This ratio is calculated using

$$SV = \frac{A_{\text{surf}}}{V}. \tag{5.24}$$

- *Compactness* This feature is computed by

$$Comp = 36\pi \frac{V^2}{A_{\text{surf}}^3}. \tag{5.25}$$

- *Sphericity* This feature is a measure of how spherical, or rounded the shape is. It is defined as the surface area of a sphere with the same volume as the tumor shape, divided by the surface area of the tumor shape. This is calculated as follows:

$$\psi = \frac{\pi^{1/3} \left(6V\right)^{2/3}}{A_{\text{surf}}}. \tag{5.26}$$

- *Spherical disproportion* This characteristic indicates how close a 3D shape is to a sphere with radius R and equal volume. This discrepancy can be computed as follows:

$$\psi_D = \frac{A_{\text{surf}}}{4\pi R^2}, \tag{5.27}$$

where $R = (3V/4\pi)^{\frac{1}{3}}$.

5.2.2 Texture features

In radiomics, numerous texture features are employed. In the experiments of this chapter, we have opted for using well-known and successful features. These include first-order statistics and second-order statistics, based on gray-level co-occurrence matrices (GLCMs) and run-length matrices (RLMs). These features are calculated on the MRI intensities of the largest VS tumor slice. In the following subsections, the employed features are discussed.

A. First-order statistics (FOS)

For each selected MRI slice, the following FOS of the gray-level intensities within the tumor contour are calculated: Mean, standard deviation, and median.

We have selected these specific features, because they present a highly condensed representation of the gray-level intensities within the MRI scans. Since the appearance of the tumors on MRI are either homogeneous or heterogeneous, these three metrics may provide enough information on the difference among the VS tumors to enable treatment outcome prediction. Homogeneous tumors will have similar mean and median values, accompanied with a low standard deviation, while heterogeneous tumors will have different mean and median values and higher standard deviation values.

B. Gray-level co-occurrence matrices (GLCMs)

In the experiments described in this chapter, several second-order statistical metrics based on GLCMs are exploited [86]. The GLCMs are calculated for each selected tumor slice, using the method described in Section 2.7.1. In summary, let $\mathbf{P}_{\theta,N_\ell,d}$ be the GLCM constructed at an angle θ and a neighboring pixel distance d of an image quantized to N_ℓ levels. Element (i,j) of $\mathbf{P}_{\theta,N_\ell,d}$ denotes the number of times that a co-occurrence between two pixels exists, where the pixel of interest with gray-level i occurs together with another pixel with gray-level j at the specified distance d, number of quantization levels N_ℓ, and angle θ. Based on these matrices and the research by Parmar *et al.* [209], the following four features related to GLCMs are computed:

- *Contrast*: $Cn = \sum_{i,j} |i-j|^2 \mathbf{P}_{\theta,N_\ell,d}(i,j)$,
- *Correlation*: $Co = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)\mathbf{P}_{\theta,N_\ell,d}(i,j)}{\sigma_i \sigma_j}$,
- *Energy*: $E = \sum_{i,j} \mathbf{P}^2_{\theta,N_\ell,d}(i,j)$,
- *Homogeneity*: $H = \sum_{i,j} \mathbf{P}_{\theta,N_\ell,d}(i,j)/(1+|i-j|)$.

Here, i and j are the row and column indices of the GLCM, respectively, μ_i and μ_j are the averages of row i and column j, respectively, and σ_i and σ_j are the standard deviations of row i and column j, respectively. The parameter choices for θ, N_ℓ, and d for the GLCMs are selected within ranges that are viable for each specific parameter within the data. This empirical choice was made after visual inspection of the MRI images and statistical analyses on the sample data.

C. Run-length matrices (RLMs)

Another set of second-order statistical metrics incorporated in the experiments is based on the gray-level RLMs [87]. Again, these matrices are calculated per selected tumor slice. Let $\mathbf{R}_{\theta,N_\ell}$ be the RLM constructed of an image quantized to N_ℓ levels at an angle θ. Then, $\mathbf{R}_{\theta,N_\ell}(i,j)$ denotes the number of times that a run of length j occurs having gray-level intensity i in the direction θ. Figure 5.2 depicts a visual representation for calculating the RLMs. Based on these RLMs, the following features are computed:

- *Short-run emphasis*: $SRE = \sum_{i,j} j^{-2}\mathbf{R}_{\theta,N_\ell}(i,j)/S_{\text{RLM}}$,
- *Long-run emphasis*: $LRE = \sum_{i,j} j^{+2}\mathbf{R}_{\theta,N_\ell}(i,j)/S_{\text{RLM}}$,
- *Gray-level non-uniformity*: $GLN = \sum_i \left(\sum_j \mathbf{R}^2_{\theta,N_\ell}(i,j)\right)/S_{\text{RLM}}$,
- *Run-length non-uniformity*: $RLN = \sum_j \left(\sum_i \mathbf{R}^2_{\theta,N_\ell}(i,j)\right)/S_{\text{RLM}}$,

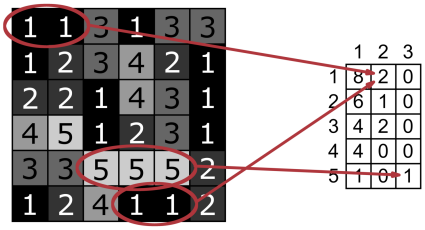


Figure 5.2 — *Graphical representation of the calculations of the run-length matrices (RLMs). For the RLM, the number of equal-valued connected pixels in a specific direction (e.g. horizontal or vertical) is counted. The pixel values depend on the number of quantization levels, and for different levels distinct RLMs can be calculated. In this figure, some example run-lengths in the horizontal direction are highlighted. These run-lengths consist of connected pixels with values "1" and "5" of size 2 and 3, respectively. Since there are two run-lengths with pixel value "1" of size 2, the corresponding position in the resulting matrix becomes "2". The same can be done for the run-length with the value "5". This results in a "1" on the corresponding position in the matrix. For each individual RLM a single feature vector is calculated incorporating the above-described statistics, which serves then as input for training.*

- *Run percentage*: $RP = S_{\text{RLM}}/n$,

where S_{RLM} is the sum of all elements of matrix $\mathbf{R}_{\theta,N_\ell}$. Again, i and j are the row and column indices of the RLM, respectively, and n is the total number of pixels of the tumor in the selected MRI slice. The parameter choices for θ and N_ℓ for the RLMs are selected within ranges that are viable for each specific parameter within the data. This empirical choice was made after visual inspection of the MRI images and statistical analyses on the sample data.

5.3 Experimental setup

This section provides a description of the experimental setup. This setup is visualized in a flow diagram in Figure 5.3. The individual blocks of this diagram are discussed in detail in the following subsections. First, Section 5.3.1 discusses the employed data. Second, in Section 5.3.2, the shape feature selection procedure is presented. Third, for the tumor texture analysis, pre-processing of the MRI data is needed. This step is highlighted in Section 5.3.3. This section is concluded in Section 5.3.4, where the employed classification and validation strategies are presented.

5.3.1 Data

In this chapter, the first experiments in evaluating the prognostic value of the VS tumor shape and its MRI appearance on the Gamma Knife treatment response are presented. In the previous chapter, a limited dataset of 40 selected tumors is employed. These VSs were selected for their extreme response to treatment, resulting in either a continued volume progression for 20 tumors, or in a significant decrease

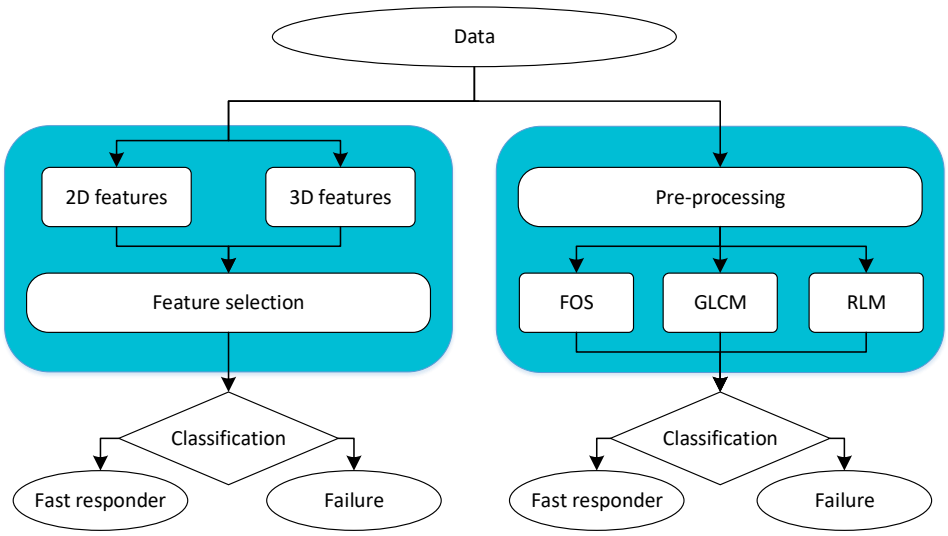


Figure 5.3 — *Flow diagram of the proposed approach, of which two separate experiments are visualized. The left part depicts the shape experiments, while the right part illustrates the texture experiments. For the shape feature extractor, both two- and three-dimensional characteristics are calculated. The texture feature extraction methods applied in this research are first-order statistics (FOS), and second-order statistics based on the gray-level co-occurrence matrix (GLCM) and run length matrix (RLM). Classification is performed by support vector machines and decision trees for both experiments.*

in tumor volume within a relative short time-period for the other 20 tumors. Because of their distinct difference in treatment response, as seen in Figure 4.7, we are able to evaluate the impact of the tumor shape and the tumor appearance on the treatment response. The contour of each VS tumor, as drawn by the neurosurgeon during treatment planning, is extracted from the treatment-planning station. From these contours, the tumor shape descriptors can be calculated. Furthermore, each MRI obtained at the day of treatment was collected. For enabling the tumor texture feature calculations, the MRI tumor voxels, i.e. the voxels within the annotated tumor contours, are extracted from the treatment MRI scans. These include the T1-weighted (T1), T2-weighted (T2), and T1-weighted contrast-enhanced (T1CE) MRI scans.

5.3.2 Selection of shape features

In a machine learning approach, the quality of the resulting model is influenced by the number of input features. It needs enough informative characteristics to enable a clear distinction between the different classes. However, too many input features can cause overfitting. This leads to a model that corresponds too closely to the data, and as such may fail to reliably fit future observations. Therefore, feature selection is required. To this end, shape descriptors are visually evaluated, using the following steps.

- *Scatter plots*: The values per shape for each feature are visualized in a scatter plot to assess the inter-class variations.
- *Feature distribution*: Histograms are created per shape descriptor to evaluate the overlap of the per-class constructed histograms.
- *Boxplots*: The per-class distribution of the individual shape feature values are captured in boxplots to visually check if the notches overlap.

If these plots suggest a significant difference between both data classes, i.e. fast responders or failures, the related feature is included in the machine learning approach. Furthermore, unpaired Student's *t*-tests are employed to evaluate the statistical significance of differences between both cohorts for each shape descriptor. Every feature with a resulting p-value lower than 0.05 is considered for inclusion.

Finally, for training the machine learning classifiers, the following four experiments are conducted.

- *Single contour, 2D shape descriptors*: Each individual contour with a feature vector consisting of the included 2D shape descriptors, is employed in the machine learning approach.
- *Complete tumor, 2D shape descriptors*: For each individual tumor, the mean, standard deviation, skewness, and mean absolute deviation of the per-contour-calculated 2D shape descriptors are computed.
- *Complete tumor, 3D shape descriptors*: Each individual tumor with a feature vector consisting of the included 3D shape descriptors, is utilized in the machine learning approach.
- *Complete tumor, 2D and 3D shape descriptors*: For each individual tumor, the 2D and 3D feature vectors are combined and implemented in the machine learning approach.

5.3.3 Pre-processing MRI data for texture features

Unlike other medical imaging techniques such as computed tomography, MRI data are expressed in arbitrary units, which differ between MRI machines, sequences, studies, and subjects. To support comparison of results, MRI intensities need to be normalized. In this chapter, the normalization step is performed by subtracting the minimum voxel intensity within the complete MRI, followed by dividing all gray-level intensities by the maximum MRI intensity. This results in voxel values within the unity range. After normalization, the MRI slice with the largest tumor area is selected for calculating the texture features.

5.3.4 Classification and validation strategies

In the experiments of this chapter, both support vector machine (SVM) and decision tree (DT) learning algorithms are utilized for classification and prediction of

the treatment response. Section 2.6 presents an in-depth description of these algorithms. For training the prediction models, the classification learner application of MATLAB is exploited. This application offers multiple model types per classifier for training. For the experiments, the following model types are exploited for the DT classification: simple tree, medium tree, and complex tree. The resulting trees have node depths of 4, 20, and 100, respectively. For the SVM classification, we employ the following six model types: linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian.

To examine the predictive accuracy of the fitted models, leave-one-out cross-validation (LOOCV) and 10-fold cross-validation are employed, as discussed in Section 2.6.3. The performance of each classifier is assessed by accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the receiver operating characteristic curve (AUC). In case of having SVM models or DT models with the similar performance, the simplest SVM model or DT model is chosen. Hyper-parameter optimization was not applied, i.e. all settings in the application are left on default.

5.4 Results on shape descriptors

This section presents the results of the treatment outcome prediction strategy, based on shape descriptors. First, results on the feature selection methods are given, followed by the classification performance of the SVM- and DT approaches employing the included shape features.

5.4.1 Results on selected shape features

Visual inspections of the scatter plots, histograms, and boxplots suggest that the descriptors are unable to clearly separate the two classes. From the scatter plots it becomes clear that inter-class variations are small, while intra-class variations are large. Some resulting scatter plots can be found in Figure 5.4. Based on the histograms, it is observed that the data is skewed and histograms overlap for a significant part. Using the boxplots, it is discerned that in most cases the notches overlap completely, except for the roughness index descriptor. Some of the boxplots can be found in Figure 5.5.

In order to statistically assess the differences in each shape descriptor between the two classes, unpaired t-tests are conducted. The results are depicted in Table 5.1. The p-values obtained from these tests reveal that P, A, RI, and the four CSM shape descriptors are significantly different between the two included classes. For this reason, these shape descriptors are selected for the machine learning approach. For the 3D shape descriptors, the t-tests display no statistical significance of the differences between the two classes, as can be seen in Table 5.1. Nevertheless, all 3D shape features are considered to be possible predictors, because of previously obtained results in the advancing field of radiomics. This approach results in feature vectors consisting of seven 2D shape descriptors per

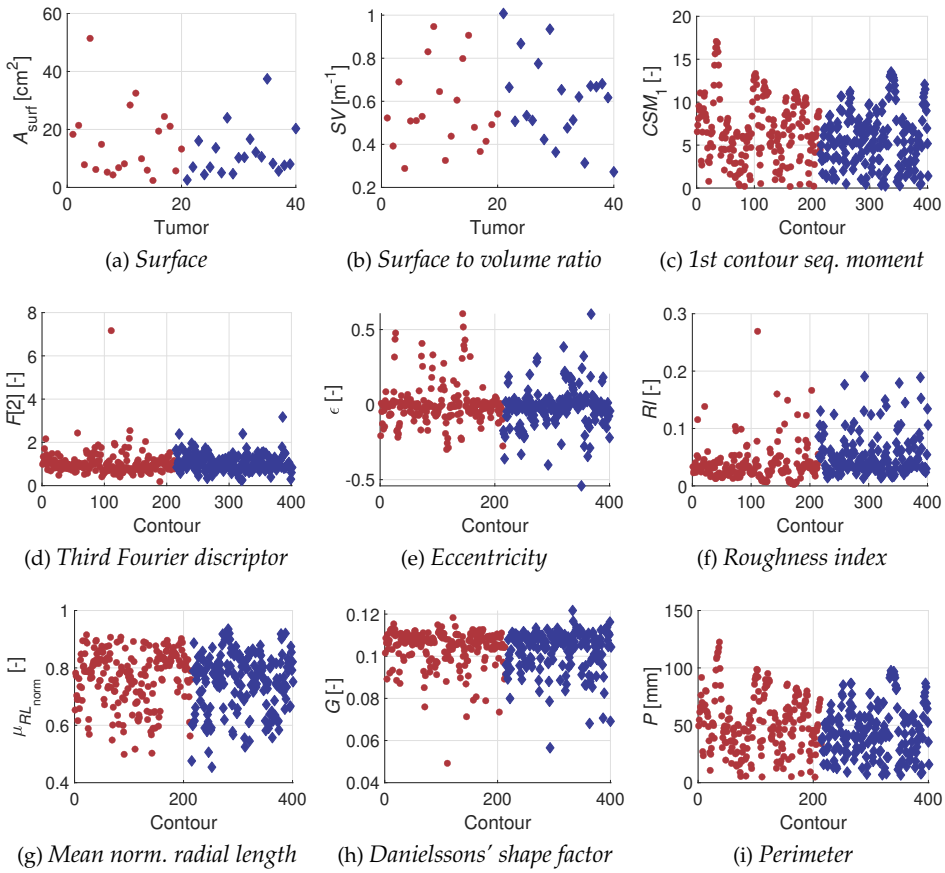


Figure 5.4 — *Some of the resulting scatter plots for visual inspection, showing the feature value per tumor or contour (depending on shape descriptor) for the fast responders (red dots) and the failures (blue diamonds). None of the scatter plots show clear differences between the two cohorts.*

contour, and twenty-eight 2D and six 3D shape descriptors per tumor.

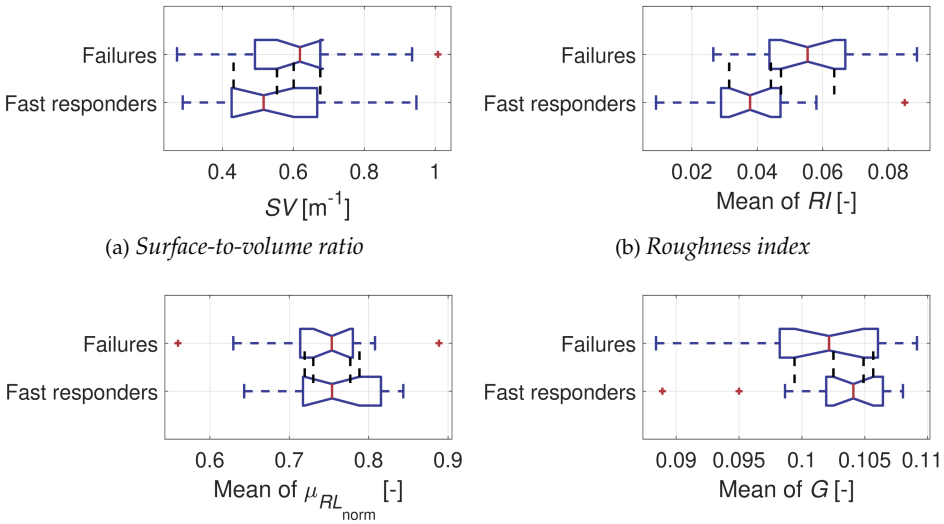
5.4.2 Classification results on shape features

The performance scores of each one of the best SVMs in predicting whether GKRS on VS leads to a failure can be found in Table 5.2. This table shows that the SVM trained on the per-tumor averaged contour data obtains the best-performing model, with an accuracy of 67.5%. The highest TNR (70%) and AUC (0.70) are also obtained with this SVM. Models trained on the other feature vectors obtain comparable values for the accuracy, TPR, and AUC. However, for the SVM trained on the 2D individual contour shape descriptors, the specificity is significantly lower (57.8%).

The results obtained by the best-performing DTs are presented in Table 5.3.

	Shape descriptor	Parameter	p-Value
2D	Perimeter	P	**<0.01**
	Area	A	**<0.01**
	Mean normalized radial length	$\mu_{RL_{\text{norm}}}$	0.13
	Area ratio	AR	0.33
	Roughness index	RI	**<0.01**
	Normalized shape compactness	$P2A_{\text{norm}}$	0.33
	Haralick's circularity measure	HC	0.49
	Danielssons' shape factor	G	0.21
	Normalized discrete compactness	$C_{D,\text{norm}}$	0.39
	Eccentricity	ϵ	0.21
	Convexity	CVX	0.50
	First Fourier coefficient	$\overline{F}[1]$	0.23
	Second Fourier coefficient	$\overline{F}[2]$	0.55
	Third Fourier coefficient	$\overline{F}[3]$	0.27
	Fourth Fourier coefficient	$\overline{F}[4]$	0.35
	First contour sequence moment	CSM_1	**<0.01**
	Second contour sequence moment	CSM_2	**<0.01**
	Third contour sequence moment	CSM_3	**0.01**
	Fourth contour sequence moment	CSM_4	**0.03**
3D	Volume	V	0.33
	Surface area	A_{surf}	0.26
	Surface-to-volume ratio	SV	0.49
	Compactness	$Comp$	0.89
	Sphericity	ψ	0.74
	Spherical disproportion	ψ_D	0.57

Table 5.1 — *Results of the unpaired t-tests on each individual shape descriptor, comparing the two included classes, i.e. fast responder and treatment failure. From these tests only the following 2D shape descriptors reached statistical significance (in bold): perimeter, area, roughness index, and the four contour sequence moments.*



(a) *Surface-to-volume ratio*

(b) *Roughness index*

(c) *Mean normalized radial length*

(d) *Danielssons' shape factor*

Figure 5.5 — *Selection of the resulting boxplots for visual inspection. Only the roughness index (RI) presents with non-overlapping notches, thereby suggesting a significant difference between both cohorts.*

Data	Descriptors	ACC [%]	TPR [%]	TNR [%]	AUC
Per contour	2D	63.6	68.7	57.8	0.70
Compl. tumor	2D	67.5	65.0	70.0	0.70
Compl. tumor	3D	65.0	60.0	70.0	0.60
Compl. tumor	2D + 3D	65.0	65.0	65.0	0.69

Table 5.2 — *Performance scores of the best support vector machines. Here, the experiments are divided based on the input data, i.e. either individual contours, or complete tumor using all image slices, and the implemented features: 2D shape descriptors, 3D shape descriptors, or the combination of 2D and 3D shape descriptors. The performance is measured in accuracy (ACC), sensitivity (or true positive rate (TPR)), specificity (or true negative rate (TNR)), and area under the receiver operating characteristic curve (AUC).*

Data	Descriptors	ACC [%]	TPR [%]	TNR [%]	AUC
Per contour	2D	62.6	82.7	39.6	0.62
Compl. tumor	2D	55.0	55.0	55.0	0.59
Compl. tumor	3D	42.5	40.0	45.0	0.36
Compl. tumor	2D + 3D	55.0	70.0	40.0	0.63

Table 5.3 — *Performance scores of the best decision trees. Here, the experiments are divided based on the input data, i.e. either individual contours, or complete tumor using all image slices, and the implemented features: 2D shape descriptors, 3D shape descriptors, or the combination of 2D and 3D shape descriptors. The performance is measured in accuracy (ACC), sensitivity (or true positive rate (TPR)), specificity (or true negative rate (TNR)), and area under the receiver operating characteristic curve (AUC).*

These are all obtained using a coarse DT strategy. The validation results show that the DT trained on the individual 2D shape descriptors reaches the highest accuracy (62.6%) and sensitivity (82.7%). However, the resulting specificity values are inferior to the results obtained with SVM. The DT model trained on 3D shape descriptors alone obtained the worst results and is even inferior to random selection.

From all best classifiers, the SVM trained on the tumor data with averaged 2D shape descriptors has attained the maximum performance in predicting whether the Gamma Knife treatment would lead to a failure or to a fast volume reduction, achieving accuracy, sensitivity, specificity, and AUC values of 67.5%, 65%, 70%, and 0.70, respectively.

5.5 Results on tumor texture

This section discusses the obtained results using tumor texture features. First, the selected parameter values of the GLCM and RLM calculations are presented, followed by the performance results of the texture features. All implementations are realized in MATLAB, using the RLM implementation based on the algorithm by Wei [210].

5.5.1 Parameter selection

For the GLCM and RLM matrices, only neighboring pixels in the horizontal and vertical directions are considered, resulting in θ equal to 0° and 90°. The diagonal angles are not considered in this research, since visual inspection of the data did not present any particular diagonal structures. The number of levels N_ℓ in which the gray-scale image intensities are quantized for feature extraction, is chosen as a power of two, with a maximum number of levels set at 64. This choice is based on the expectation that very subtle intensity changes are not relevant. The maximum neighboring pixel distance d in the GLCM method is set to 4. This value is based on the average horizontal and vertical sizes of the tumors in the dataset, which is found to be 9 pixels. Combining all these parameter choices with the applied

Parameter	Symbol	Values
Angle	θ	0°, 90°
Levels	N_ℓ	4, 8, 16, 32, 64
Distance	d	1, 2, 3, 4

Table 5.4 — *Parameters of the GLCM and RLM texture feature extractors.*

Classifier	Validation	ACC [%]	TPR [%]	TNR [%]	AUC
DT, coarse tree	10-fold	65.0	75.0	55.0	0.68
DT, coarse tree	LOOCV	67.5	62.0	70.0	0.85
SVM, fine Gaussian	10-fold	85.0	85.0	85.0	0.85
SVM, fine Gaussian	LOOCV	85.0	85.0	85.0	0.85

Table 5.5 — *Best classification performance scores for FOS texture features, measured in accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and area under the receiver operating characteristic curve (AUC). The best results are obtained from exploring three different decision tree (DT) models and six support vector machine (SVM) models (Section 5.3.4). For DT, the coarse tree obtained the best scores, and for SVM the fine Gaussian model. Validation was performed by 10-fold cross-validation and leave-one-out (LOOCV) cross-validation, both obtaining comparable results.*

classifiers, i.e. DT and SVM, and validation methods (10-fold CV and LOOCV), results in a total of 204 separate experiments in the MATLAB classification learner application: 4 experiments for the FOS features, 160 experiments for the GLCM features, and 40 experiments for the RLM features. All different parameter settings for the GLCMs and RLMs are summarized in Table 5.4.

5.5.2 Classification performance on texture features

The results of the best performing model for each combination of classification and validation method for the FOS features are presented in Table 5.5. It shows that the type of validation has little impact on the achieved results. However, the difference between machine learning techniques is significant. SVM achieves higher accuracy than DT, being 85.0% and 67.5%, respectively. The highest sensitivity (85.0%) and specificity (85.0%) are also obtained from SVM. Since there are only three different features used in training, the medium and complex trees did not improve the results obtained from training a simple tree, where only maximally 4 splits are implemented, compared to 20 and 100 for the other two methods, respectively. For the SVM training, the best results are achieved by applying a fine Gaussian kernel. Cubic SVM reaches comparable accuracy results (77.5%) with the same sensitivity, but the specificity is reduced to 70.0%. Hence, the fine Gaussian method is selected as the best performing classifier.

Employing GLCM features for classification resulted in Figure 5.6. This bar

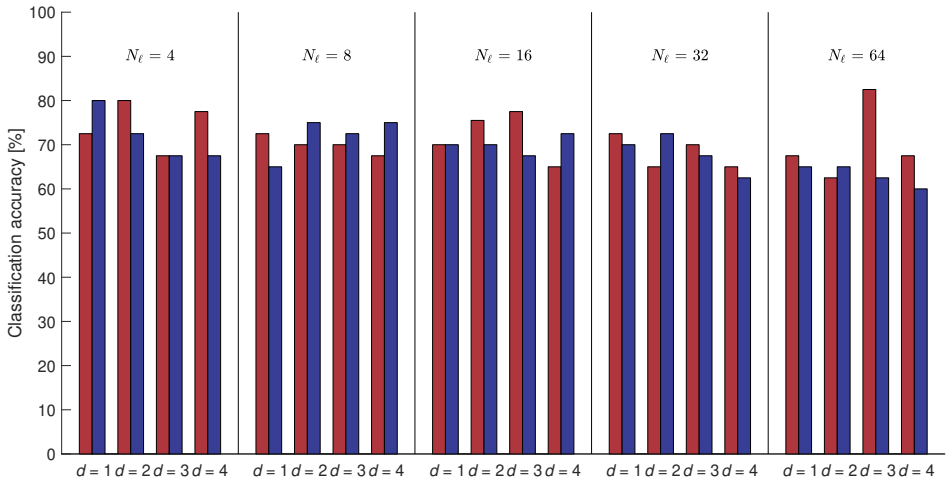


Figure 5.6 — *Bar chart of the best-performing classification models trained on GLCM features. The red and blue bars denote the results trained on GLCMs constructed with inter-pixel angle $\theta = 0°$ and $\theta = 90°$, respectively. Furthermore, the bars are grouped per inter-pixel distance (d) and number of quantization levels (N_ℓ).*

chart depicts the best results for each combination of feature parameters. After evaluating the results of the 204 classification and validation experiments, it appears that there is no distinguishable overall difference between the training methods, as was the case with the FOS features, highlighted in Table 5.5. However, models trained by the coarse Gaussian method and validated by LOOCV consistently result in a failed classification or too low accuracy results, reaching at most 47.5%. Most of the other training results achieve accuracy values of less than 70%. The highest obtained accuracy for GLCM features is 82.5% for a simple tree, trained with parameter combination $d = 3$, $N_\ell = 64$, and $\theta = 0°$, using LOOCV cross-validation. The resulting model obtains a sensitivity and specificity of 90.0% and 75.0%, respectively. This result stands out as the overall accuracy results seem to decline for increasing N_ℓ and d, as can be seen in Figure 5.6. Overall, quadratic and cubic SVMs show promising results, especially for a low number of quantization levels and inter-pixel distances, resulting in accuracy values of about 80%.

The results for models trained by the RLM features are presented in Figure 5.7. From this figure it can be distilled that SVMs obtain overall the best accuracy results, with the only exception for the feature parameter combination of $N_\ell = 64$ and $\theta = 0°$. When evaluating the specific SVM methods, the Gaussian models have the best classification capability. The best performing RLM texture feature vectors achieve an accuracy of 77.5%, obtained with feature parameters $N_\ell = 4$ and $\theta = 90°$, $N_\ell = 8$ and $\theta = 90°$, and $N_\ell = 16$ and $\theta = 90°$. The corresponding sensitivity and specificity values are 95.0% and 65%, respectively, for the first parameter vector, 80.0% and 75%, respectively, for the second parameter vector,

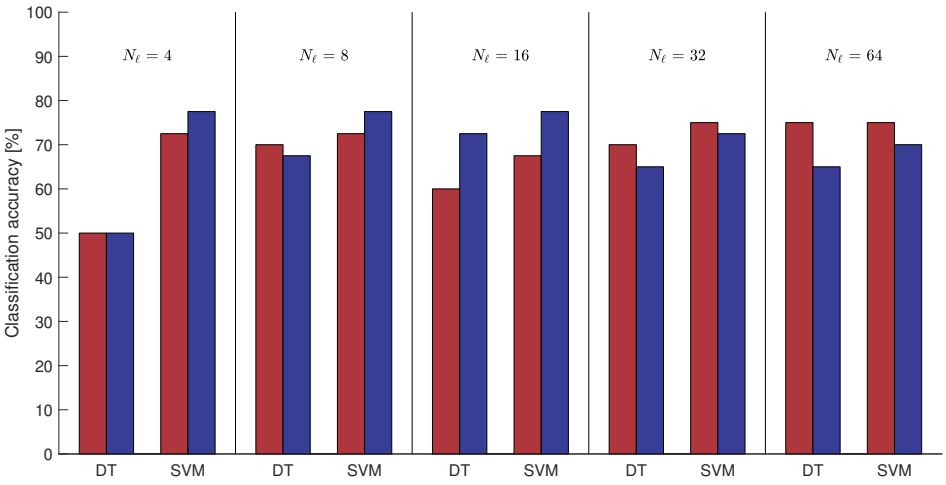


Figure 5.7 — *Bar chart of the best-performing classification models trained on RLM features. The red and blue bars denote the results trained on RLMs constructed with run-length angle $\theta = 0^\circ$ and $\theta = 90^\circ$, respectively. Furthermore, the bars are grouped per machine learning algorithm (decision tree (DT) and support vector machine (SVM)), and number of quantization levels (N_ℓ).*

and 75.0% and 80.0%, respectively, for the third parameter vector. Although the sensitivity is high, especially for the first parameter vector, the specificity is lower than using the models trained on FOS features.

When comparing the FOS results from Table 5.5 with the second-order statistical features from Figures 5.6 and 5.7, the overall observation is that FOS with SVM outperforms the GLCM and RLM features.

5.6 Discussion

The experiments described in this chapter have been designed to investigate the possible predictive value of image features obtained from conventional clinical MR images for the GKRS treatment of VS tumors. In these experiments, the tumor shape and tumor texture on MRI have been evaluated.

A. Shape as predictor

Experiments on the employed data have demonstrated that shape is a weak predictor for considering that GKRS on VS will result in a failure. The best trained classifier achieves an accuracy and AUC of 65.0% and 0.71, respectively.

In research concerning the GKRS outcome of VS, the only shape-related parameter that has been investigated is the tumor volume. In a previous study by Klijn *et al.* [47], tumor volume was appointed as a possible predictor of tumor control after GKRS on VS. They employed statistical analyses on records of 420 patients to show that tumor volume was significantly correlated to tumor control ($p < 0.01$). These results are concurred by other authors, but are also contradicted by alternative studies, as discussed in Section 2.2.3. In this chapter, using machine

learning classifiers, the conducted experiments have demonstrated that tumor volume is not a good predictor. Separation of the different outcomes, i.e. fast responder and failure, proved to be limited with the tumor volume as input feature. Furthermore, employing other 2D and 3D shape features in various combinations do not improve the classification performance in terms of accuracy and AUC.

In other oncological fields, such as breast cancer, authors determined that shape of the tumors can classify different lesions, supporting the creation of computer-aided diagnosis tools [74]–[76]. Even in brain cancer research, shape can be employed for improving the survival prognosis [211]. Nevertheless, our data has revealed that the commonly employed shape features are not suitable for predicting the GKRS treatment response on VS. Both almost perfect spherical tumors and lobulated tumors are found in each of the two classes, thereby underlining that the overall differences in shape between the two selected classes is limited.

B. Tumor texture as predictive feature

Conversely, our experiments on tumor texture information do yield predictive value of the Gamma Knife treatment responses in the employed data. More specifically, first-order statistics on MRI gray-level intensities obtain high accuracy, sensitivity, specificity and AUC values of 85.0%, 85.0%, 85.0%, and 0.85, respectively.

Concurring with our findings, Speckter *et al.* [59] also determined that the histogram of tumor gray-level intensities provided information that may enable the *a-priori* prediction of the Gamma Knife response of VS tumors. However, they achieved lower sensitivity and specificity values in their employed dataset. Nevertheless, it can be concluded that there is information in the VS tumor texture that may enable the Gamma Knife treatment response prediction.

C. Limitations

In order to investigate the predictive value of quantitative image features on the Gamma Knife treatment response, two equally sized tumor cohorts were selected. The included patients have presented an extreme treatment response: either a continued tumor progression, or a significant volume reduction in the first year following treatment. This selection has enabled technically well-defined experiments. However, it may have caused a bias in the employed data. Furthermore, the clinical definition for treatment success is stopping tumor expansion and not necessarily reducing the size of the tumor. However, in large tumors that are already causing pressure on nearby brain structures, the ability to predict a significant volume reduction within a short time period following GKRS would be highly beneficial for the treatment selection process. Nevertheless, there is a discrepancy between the definition employed in this chapter, i.e. fast responder, and the clinically desired long-term tumor control. The applied definition of treatment success enabled us to create two very dissimilar cohorts of each 20 patients. However, this definition of treatment success causes a limitation on the actual predictive value of the implemented features. The best choice for texture features needs to be investigated further, employing extensive analyses on larger datasets, includ-

ing a broader definition of treatment success. This will be a logical next step in confirming and refining the predictive value of MRI texture features of VSs on the Gamma Knife treatment response.

Because of the above discussion, the experiments concentrated on a patient cohort with fast-reducing VSs after GKRS. As a consequence, the obtained results found in this research thereby confirm the conclusions drawn by Frisch *et al.* [46] and Bowden *et al.* [50], that cystic tumors tend to have better volume reductions. These tumors have lower image intensities on T1-weighted, contrast-enhanced MRI scans, which we found to be of predictive value in our dataset. However, in our case, the experiments did not focus on cystic tumors specifically, and our patient cohort was not analyzed on this characteristic.

5.7 Conclusions

Prognostic factors of tumor control after GKRS for VS are largely unknown, so it remains difficult to *a-priori* predict the GKRS treatment response of an individual VS patient. Such prediction is of crucial importance for each specific patient, in order to choose the most-suited treatment modality, i.e. microsurgery or stereotactic radiosurgery. Furthermore, a likelihood estimation of tumor control after GKRS can potentially provide a basis for an individualized follow-up protocol, reducing the overall number of follow-up visits and MRI scans.

The experiments on evaluating the predictive value of tumor shape on the Gamma Knife treatment response, result in the conclusion that shape appears to be a weak predictor. Both SVM- and DT-trained models indicate that classifying the treatment response of a VS, based on the calculated shape descriptors, does not provide a significant improvement over random classification. This shortcoming is explained by the observation that the inter-class differences, i.e. variations between the failure and the fast-responder cohorts, are limited, whereas intra-class differences are considerable. Hence, the variations between both classes are too restricted, thereby reducing the prognostic value of these shape descriptors significantly.

However, experiments involving the tumor texture have shown that popular second-order statistical metrics, like GLCM and RLM, are suitable for describing texture and predicting the Gamma Knife treatment response. Nevertheless, these metrics are slightly outperformed by simple first-order statistics, like mean, standard deviation and median, obtaining an accuracy, sensitivity, and specificity of 85.0%. Nevertheless, the best choice for texture description can only be made after performing more extensive analyses on larger datasets. In any case, these experiments provide useful texture measures for successful prediction of GKRS treatment outcome for VS and invokes further research on patient-specific evaluation for VS treatment options.

The latter positive conclusion fuels the demand for further research on this topic. Therefore, in the following chapter, we will further investigate the predictive

value of MRI tumor texture on another, clinically highly relevant Gamma Knife treatment response of vestibular schwannomas. Using the results obtained in this chapter, the next chapter will examine their correlation to the short-term adverse effect of radiation-induced transient tumor enlargement. This temporary swelling of the tumor occurs in a broad range of all VS patients within the first two to three years after treatment. It can cause a temporary increase in cranial nerve morbidities and, in case of large VSs, even life-threatening morbidities. It is therefore extremely beneficial to enable the prediction of this adverse effect prior to treatment, such that severe patient problems can be avoided.

Chapter 6

Transient tumor enlargement

6.1 Introduction

The previous chapter has introduced the first experiments into evaluating the predictability of the treatment response, based on tumor-specific features obtained from the treatment MRI scans. Using a limited dataset, with 20 tumors that showed a significant volume reduction within the first year after treatment, and 20 tumors for which the treatment did not result in a halted tumor progression, the previous chapter has highlighted that tumor shape descriptors are not suitable for treatment outcome prediction. Furthermore, by employing specific image-related tumor texture features, we were able to obtain accuracy, sensitivity, and specificity values of 85.0%, 85.0%, and 85.0%, respectively.

Nevertheless, the previous chapter has also presented one significant concern in determining the predictive value of the tumor texture features: the limitation of the employed dataset. Although the rapid and significant volume reduction may be clinically interesting, there remain some methodological concerns. As discussed in Section 2.4, clear and objective definitions of the treatment response are needed. Despite these concerns, the previous chapter has provided interesting results, concluding that tumor texture features may have predictive value on the Gamma Knife treatment response.

Therefore, in this chapter, clear and objective definitions are employed to create a dataset that is both clinically highly interesting as well as technically well defined. From a clinical point of view, one of the major contraindications for Gamma Knife treatment of VS tumors is the tumor size. As discussed in Section 1.1, the discussion concerning the best treatment strategy is difficult to answer, especially for large VS tumors. Most medical centers consider microsurgical resection as the optimal treatment strategy for these large tumors, because it effectively averts the compression of surrounding critical brain structures, such as the brain stem, the cerebellum and the neighboring cranial nerves. Since the risks involved in microsurgery can be contra-indicative for this strategy, less invasive treatments such as radiosurgery and radiotherapy have been considered increasingly in the last decade. These strategies have obtained good results for large VSs and achieved acceptable radiation-induced morbidities [17]–[22], [24], [61].

Nevertheless, radiosurgical treatments of large VSs remain controversial due to the possible transient tumor enlargement (TTE). This radiation-induced swelling

of the tumor, also known as pseudo-progression, occurs in a broad range of 11 – 74% of all VS patients in the two to three years following treatment and can cause a temporary increase in cranial nerve morbidities [33], [52], [123]–[136]. For large VSs, where the tumor already exhibits a mass effect on the brain stem, this post-radiation effect may cause severe, and in some cases, life-threatening morbidities. This adverse effect necessitates salvage treatment, further increasing the risk of surgical complications.

As TTE is one of the major contra-indicators for radiosurgical treatment of large VSs, it would be extremely beneficial if this effect can be *a-priori* predicted. This will enable physicians to select the most optimal treatment strategy on an individual basis. However, it remains unclear why some patients exhibit TTE, while others do not show TTE but exhibit an arbitrary volumetric response. Several investigations into the correlation of tumor- and treatment-related characteristics to this effect have been reported [33], [52], [123]–[136], [212]. However, their results remain inconclusive. Treatment-related characteristics, such as marginal radiation dose and maximum tumor dose, were found not to correlate with TTE occurrence in all but one study [133]. Some papers describe that tumor volume is significantly different between VSs presenting TTE and those that do not [126], [133], [136], while others could not find this correlation [33], [123]–[125], [130], [132], [134], [135]. Also tumor appearance on MRI, classifying a VS tumor as cystic or solid, has been investigated. Shirato *et al.* [212] determined that cystic tumors are more likely to exhibit TTE. However, others did not find this correlation [124], [126], [131], [134], or even observed that cystic tumors are less likely to exhibit TTE [52].

The assumed biological effect of radiosurgery on VS cells is a combination of acute inflammation and vascular occlusion [67], [68]. Because of this and the previously described contradicting results, it is hypothesized that differences in tumor biology may be the cause of TTE in a subset of patients. Ideally, a biopsy is performed to analyze tumor tissue. However, this is an undesirable procedure, since post-biopsy hemorrhage is one the most frequently encountered complications, which can cause even death due to the close proximity of the VS location to the brain stem. The more readily available source of biological information is through imaging techniques, such as magnetic resonance imaging (MRI). These scans are already obtained for diagnostics and may contain information on the biological tumor features.

A literature study was carried out to find out as to how far image analysis techniques were explored for determining features describing biological tumor properties. In a review by Gillies *et al.* [69], the authors reported on the potential power of medical image analysis using radiomics, to facilitate improved clinical decision making. Indeed, numerous studies describe the ability of employing computer-aided diagnosis using medical imaging for classifying disease and treatment response. Yang *et al.* [81] evaluated tumor-derived MRI-texture features for discriminating molecular subtypes of glioblastomas and the corresponding 12-month survival status. Their study obtained area under the receiver operating characteristic (AUC) values of 0.70 to 0.82 for the specific subtypes, and 0.69

for the 12-month survival status. Moreover, specifically for radiosurgical treatment responses, several authors evaluated the possibility to distinguish true tumor growth from radionecrosis in primary malignant brain tumors and brain metastases. Utilizing computer-extracted texture features, Tiwari *et al.* [82] were able to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI. Their method obtained AUC values of 0.79 on fluid-attenuated inversion-recovery (FLAIR) MRI images, both for primary malignant brain tumors and for brain metastases, thereby outperforming the diagnosis made by the medical experts. Zhang *et al.* [83] evaluated 285 texture features calculated on four different MRI sequences, to find a predictive model distinguishing radionecrosis from true tumor progression following radiosurgery on brain metastases. They obtained an AUC value of 0.73 using so-called delta feature values, that represent the change in feature values over time. Peng *et al.* [84] obtained an AUC value of 0.79 on distinguishing radionecrosis from tumor progression, using 10-fold cross-validation of their prediction model. Wang *et al.* [85] demonstrated that multimodality MRI imaging and radiomics analysis have the potential to identify early treatment response of malignant gliomas treated with concurrent radiosurgery and bevacizumab.

These studies all show the potential of distinguishing different radiosurgical treatment responses in malignant brain tumors. However, the ability to predict such a treatment response *prior* to treatment is crucial in the case of large VS tumors, because this can lead to a well-informed treatment selection based on quantitative analysis. Since clinical- or treatment-related parameters have not shown any prognostic value, it is hypothesized that quantitatively analyzing the tumor appearance on readily available MRI scans can facilitate pre-treatment prediction of the TTE effect.

Therefore, the objective of this chapter is to explore whether TTE after radiosurgery, specifically Gamma Knife radiosurgery (GKRS) on VS, can be predicted from the measured MRI tumor texture characteristics. As a refinement of this general problem statement, we therefore list the following research challenges.

- *Patient-inclusion conditions*: In order to train a prediction model, MRI data of a sufficient number of patients need to be available. There are several trade-offs that influence the amount of patients that can be included. First, a strict definition of transient tumor enlargement is required for preventing misclassifications. Second, the adverse reaction to a transient swelling is highly relevant in patients with large VS tumors, since in these cases intervention is conceivable. Finally, patients should have at least a $2-3$ years follow-up MRIs available to determine if the swelling was indeed transient.

- *Data balancing*: Another important factor concerning the MRI input data is the number of patients available in the two distinct cohorts. Since these numbers may not be evenly distributed, the machine learning approach should be able to handle the impact of this imbalance.

- *Tumor volumes*: Up to this point, the MRI-scan resolution of the VS tumors has been unrelated to the tumor size. Therefore, larger tumors contain more tumor voxels and may show more detail in tumor texture. We will evaluate the impact of tumor volume on the prediction results.
- *Machine learning*: We will analyze several texture features and apply machine learning as a technique for classifying the MRI observations.
- *TTE prediction*: We aim at *a-priori* prediction of transient tumor enlargement using machine learning. Moreover, this developed model should preferably allow an evaluation on an individual basis, where this model can be possibly implemented in a clinical decision-support system, aiding physicians and patients in selecting the optimal care path.

This chapter is outlined as follows. First, Section 6.2 elaborates on the available data and the pre-processing steps needed to normalize the MRI data. Next, the experimental setup is described for evaluating the predictability of transient tumor enlargement in Section 6.3. Then, the results of the experiments are presented in Section 6.4, after which these are discussed in Section 6.5. Finally, Section 6.6 concludes this chapter.

6.2 Data and pre-processing

This section presents a description of the available data and discusses the proposed approach for MRI normalization.

6.2.1 Data

The data employed in this research consist of prospectively collected patient- and treatment-related information and clinical MRI image data. The included patients have been selected based on a volumetric threshold derived from the Koos grade. This grade is used in a clinical setting to classify the VS tumor size. This classification is based on the tumor stage, where Grade I is selected for tumors only present in the auditory canal and Grade IV for tumors displacing the brain stem. Koos Grade IV is in medical terms considered a large tumor. For these large tumors, the adverse effects of TTE can cause severe complications because of the already caused displacement of critical brain structures and cranial nerves. Koos Grade IV tumors have a corresponding average tumor volume of 4.17 ± 2.75 cubic centimeter (cm^3) [137]. In this chapter, we select a lower bound of 1.42 cm^3 as the minimum inclusion threshold. Furthermore, as TTE will occur between 6 – 18 months after treatment, all included patients had at least an available MRI scan at 6 months following treatment and were followed-up for at least 18 months. These follow-up scans were employed for calculating tumor volume changes, to determine whether a TTE has occurred or not. This resulted in the inclusion of 99 patients.

The obtained patient- and treatment-related information include (1) age at treatment, (2) tumor volume at treatment (gross target volume, GTV), (3) radiation

dose to 99% of the GTV, (4) coverage (ratio between GTV within prescription isodose volume GTV_{PIV} and GTV), (5) selectivity (ratio between GTV_{PIV} and PIV), (6) gradient index (ratio between volume enclosed by half the prescription isodose and PIV), (7) Paddick conformity index (coverage multiplied by selectivity), (8) number of isocenters, and (9) the beam-on time. We will employ Student's *t*-tests to evaluate differences in these patient- and treatment-related characteristics between patients that suffered from TTE and those that did not.

The clinical MRI data employed in this research for texture analysis consisted of the MRI scans that were already acquired for treatment planning. These included T1-weighted, T2-weighted, and contrast-enhanced T1-weighted (T1CE) MRI scans and were obtained on the day of treatment. Ideally, for discriminating TTE from non-TTE, a histopathological evaluation of tumor tissue is employed. However, in current clinical practice, surgical intervention is highly unwanted due to the significantly increased inherent risks. This is why the medical team at our center opted for GKRS in the first place and also has the protocol to only intervene when the tumor expansion becomes life-threatening. In all other cases, watchful waiting is preferred for the first 2 – 3 years following GKRS. As such, the presence or absence of TTE needs to be determined from the MRI data obtained at follow-up visits. To this end, tumor volumes were calculated on each available follow-up MRI, by segmenting the tumor using the treatment planning software (GammaPlan Version 11, Elekta AB, Stockholm, Sweden). Several publications report that the maximum TTE is observed between 6 – 15 months after treatment, followed by volumetric reduction [33], [49], [127], [132], [134]. For this reason, the TTE effect is defined as in Section 2.4.3. This definition states that a volumetric increase of at least 10% within the first 12 months after treatment, followed by volumetric reduction to at least the tumor volume at treatment. Examples of a treatment and follow-up MRI scans of a VS tumor that exhibited TTE are shown in Figure 6.1. If tumor expansion was less than 10% during the first 2 years, the VS was considered to be stable or shrinking and consequently classified as non-TTE. Using these definitions for TTE and non-TTE, 38 out of the included 99 patients experienced a TTE after GKRS treatment. The remaining 61 patients were classified as non-TTE.

The treatment MRI scans, including the tumor delineations created by the neurosurgeon on the day of treatment, were extracted from the database of the Gamma Knife treatment system. The data from which image features are extracted consist of the MRI volume elements (voxels) within the tumor delineations.

6.2.2 Pre-processing

Whereas data from other medical imaging modalities are measured in absolute units, MRI data provides relative values. These values can differ between MRI machines, but also between patients and even between two scans of the same patient. To support comparison between subjects, the MRI intensities need to be normalized. To this end, we employ a multi-landmark intensity normalization (MLIN),

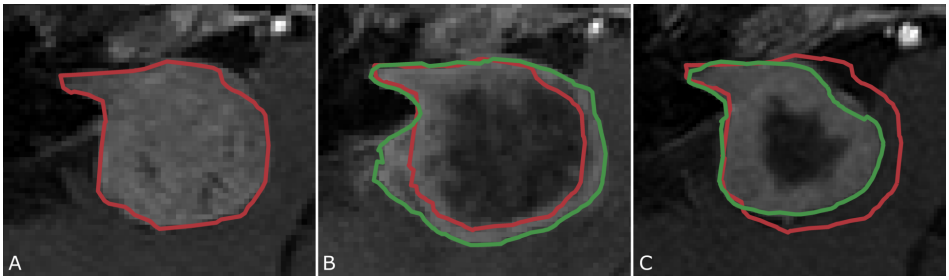


Figure 6.1 — *T1-weighted, contrast-enhanced magnetic resonance images of a vestibular schwannoma that exhibited transient tumor enlargement after Gamma Knife radiosurgery. In each part of the figure, the red delineation depicts the tumor at time of treatment. Part A: tumor at time of treatment, with a volume of 12.8 cm^3. Part B: in green, the tumor 6 months after treatment, with a volume of 17.7 cm^3. Part C: in green, the tumor 24 months after treatment, with a volume of 8.7 cm^3.*

which is based on the work by Madabhushi and Udupa [213]. This method aims to find a generalized intensity scale, such that MRI scanning parameters have limited influence on the image analysis techniques. It performs a nonlinear normalization utilizing tissue-specific landmarks. For T1 and T1CE MRI scans, the utilized landmarks are the brain stem and the fiducial markers. For the T2 scans, the selected landmarks are the brain stem, the fiducial markers, and the cerebrospinal fluid. Examples of the landmarks are given in Figure 6.2. These landmarks are manually selected, and the median value is calculated for each selected landmark region. Using a training set, a generalized scale is estimated. Next, each specific landmark value is mapped onto that generalized scale, resulting in a piece-wise linear transformation function for each individual MRI. Furthermore, in order to cope with artifacts that lead to misleadingly high intensity values, a histogram-percentile-based cut-off for the high image intensities is employed. These percentiles were empirically chosen as 99.8%, 99.8%, and 99.999%[1] for the T1, T1CE, and T2 MRI scans, respectively.

6.3 Experimental setup

In order to evaluate the possibility to predict the transient tumor enlargement effect of VS tumors following Gamma Knife treatment, a number of experiments is conducted. These experiments involve the evaluation of the impact on the TTE of the MRI tumor texture. Figure 6.3 depicts the flow diagram of the proposed approach. This section first discusses the training data selection in Section 6.3.1. Second, in Section 6.3.2, elaborates on the employed MRI image feature extractors. Finally, the classification and validation method is described in Section 6.3.3.

[1]This percentile value is chosen to include almost all voxel values. Outliers on the T2 MRI are rare and lie well beyond the normal values. Furthermore, the intensity values have almost always the highest values for the fiducial markers on T2 MRIs.

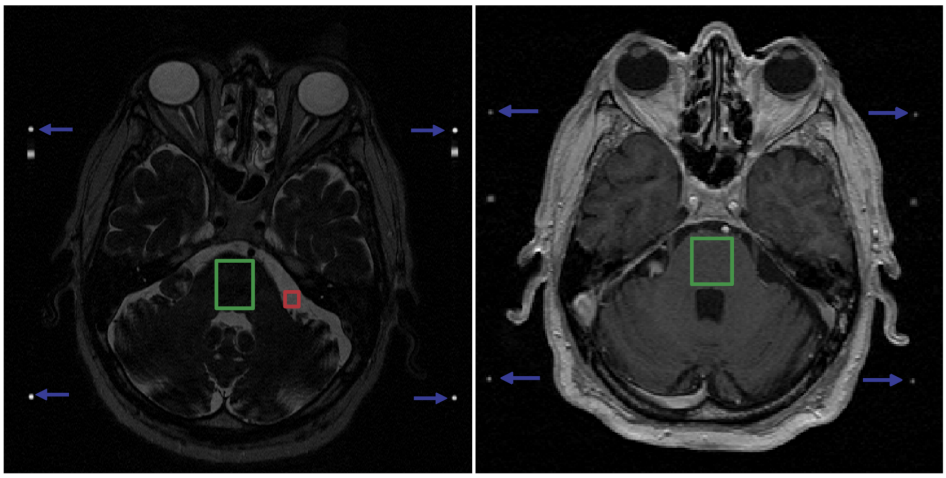


Figure 6.2 — *Examples of the landmarks used in the multi-landmark intensity normalization method for the T2-weighted magnetic resonance imaging (MRI) scans (left), and the T1-weighted, contrast-enhanced MRI scans (right). For the T1-weighted MRI scans, the same landmarks are used as shown in the right image. Specific areas are highlighted for the cerebrospinal fluid (red), brain stem (green), and the fiducial markers (blue arrows).*

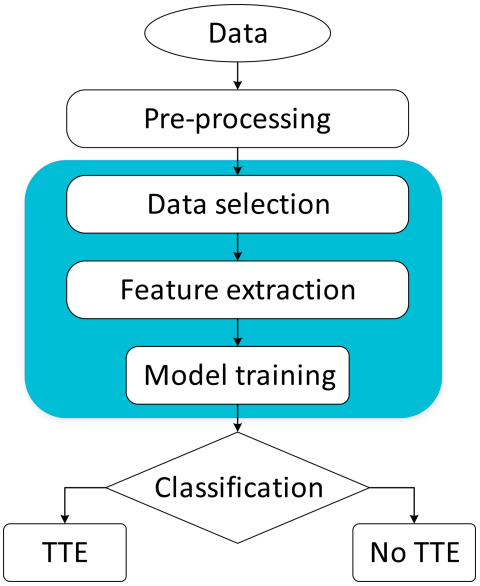


Figure 6.3 — *Flow diagram of the proposed transient tumor enlargement prediction approach.*

6.3.1 Pre-selection of training data

The conducted experiments in this chapter are based on possible confounding factors present in the data. As discussed in Section 6.1, the included cohort contains several challenging aspects. Therefore, training of the models is based on different data inclusion criteria. These criteria are discussed below.

Imbalanced dataset

The first criterion is derived from the small imbalance in the available data, where the majority class contains 61 patients (non-TTE) and the minority class 38 patients (TTE). Due to this imbalance, training can lead to a model that is skewed towards the majority class. This way, a classification algorithm can obtain a reasonable accuracy, at the cost of a low specificity. To evaluate whether this imbalance impacts the results, training is performed in two different ways. First, all available data points are employed in training the SVM models. Second, a balanced training set is used, in which each cohort is equally sized. Balancing the data is performed by random subsampling of the majority class. To account for possible data biases, a further validation loop is employed. In this loop, models are trained using n resampled subsets of the majority class. The resulting model is the average of these n models, indicating the combined model performance.

MRI sequence

The second data inclusion criterion is based on the MRI sequence. In this chapter, we have T1-, T1CE-, and T2-weighted MRI data available. In the classification approach, we have evaluated each individual MRI sequence, as well as the combination of all three sequences.

Tumor size

The third data inclusion criterion is based on the tumor volume. The data from which the MRI image features are extracted consist of the MRI voxels within the tumor delineations. Due to the employed scanning method and parameters, each tumor is scanned using the same voxel dimensions. Thus, MRI scans of large tumors contain more tumor voxels than scans of small tumors. If the number of tumor voxels increases, the amount of texture information also grows. Therefore, we have also explored the impact of the tumor volume, by imposing various volume thresholds for specific selection of the data (volume filtering). The selected volume thresholds are 2, 3, 4, 5, 6, and 7 cm^3, since higher thresholds result in too few number of patients in the minority class.

6.3.2 Feature extraction

In this section, the employed feature extraction methods are presented. These include first-order statistics and second-order statistics, based on Minkowski functionals (MFs) and gray-level co-occurrence matrices (GLCMs).

For each tumor and individual MRI sequence, the following first-order statistics (FOS) are computed from the tumor MRI voxels: mean, standard deviation,

skewness, kurtosis, and a 16-bin histogram. Next, we calculate the Minkowski functionals (MFs) as defined by Hadwiger [164], again for each individual MRI sequence. In mathematical morphology, these functionals represent geometric measurements of shapes. These shapes are obtained by transforming gray-scale images to binary images using a threshold value. Varying this threshold value will result in multiple instances from which MFs can be computed. These variations allow for extracting texture information. From the binarized data, thresholded at level T, the following elementary geometric shape objects are extracted: (1) number of cubes N_c, (2) number of open faces N_f, (3) number of open edges N_e, and (4) number of open vertices N_v [162]. These objects are employed in the calculation of the following four functionals: foreground volume M_0^T, surface area M_1^T, curvature M_2^T, and Euler number M_3^T. These functionals are specified by

$$M_0^T = \quad N_c \,, \tag{6.1}$$

$$M_1^T = -6N_c + 2N_f \,, \tag{6.2}$$

$$M_2^T = +3N_c - 2N_f + N_e \,, \tag{6.3}$$

$$M_3^T = -N_c + N_f - N_e + N_v \,. \tag{6.4}$$

The MFs are highly scale-dependent. Since the VS tumors in our dataset have a ratio between minimum and maximum of 13:1, the MFs need to be normalized with respect to the tumor volume. This is performed by dividing the functionals by the maximum tumor volume in the dataset.

Finally, for each individual MRI sequence, the gray-level co-occurrence matrices (GLCMs) are computed. For GLCM $\mathbf{P}_{\theta,d,N_\ell}$, each matrix element $\mathbf{P}_{\theta,d,N_\ell}(i,j)$ denotes the number of times a pixel with intensity i occurs together with a pixel of intensity j, at angle θ, distance d, and quantization level N_ℓ. Each element is normalized with respect to the total number of elements in the GLCM. From these matrices, the following four features are calculated: entropy (H), contrast (Cn), energy (E), and correlation (Co) [86]. These features are specified by:

$$H = \sum_{i,j} -\mathbf{P}_{\theta,d,N_\ell}(i,j) \log\left(\mathbf{P}_{\theta,d,N_\ell}(i,j)\right), \tag{6.5}$$

$$Cn = \sum_{i,j} |i-j|^2 \, \mathbf{P}_{\theta,d,N_\ell}(i,j), \tag{6.6}$$

$$E = \sum_{i,j} +\mathbf{P}^2_{\theta,d,N_\ell}(i,j), \tag{6.7}$$

$$Co = \sum_{i,j} \frac{(i-\mu_i)\,(j-\mu_j)\,\mathbf{P}_{\theta,d,N_\ell}(i,j)}{\sigma_i \sigma_j}. \tag{6.8}$$

Here, i and j are the row and column indices of each GLCM element, respectively. Parameters μ_i and σ_i denote the mean and standard deviation of row i and μ_j and σ_j denote the mean and standard deviation of column j, respectively.

Characteristic	Mean	Inter-quartile range	Range
Age [years]	58	47 – 66	24 – 84
Tumor volume [cm^3]	6.54	3.10 – 6.04	1.44 – 18.72
Dose 99% of tumor vol. [Gy]	12.36	11.80 – 13.00	11.10 – 13.20
Coverage [%]	95.74	91.00 – 99.00	86.00 – 100.00
Selectivity	0.89	0.85 – 0.90	0.71 – 0.99
Gradient index	2.74	2.58 – 2.82	2.45 – 3.60
Paddick conformity index	0.84	0.84 – 0.89	0.17 – 0.93
Number of isocenters	24	17 – 31	1 – 53
Beam-on time [min]	60.27	42.18 – 75.04	22.80 – 144.80

Table 6.1 — *Patient- and treatment-related characteristics for the complete patient cohort.*

6.3.3 Classification and validation

The final step in the experiments is to train a classifier for binary prediction of TTE. The implemented machine learning method in this chapter is support vector machines (SVMs), since it has proven to be effective in binary classification problems without requiring large amounts of data. As our dataset is relatively small and containing only 99 tumors, advanced algorithms such as neural networks are not well suited for classification. The considered SVM types include linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian, which are all implemented in MATLAB (MathWorks inc., Natick, Massachusetts, USA). Validation is performed by 10-fold cross-validation. The performance metrics for determining the optimal model are the sensitivity and specificity. In case the multiple models perform equally well, preference is given to a higher specificity: false positives can be related to erroneous prediction of TTE occurrence, which necessitates salvage treatment in a patient. Alternatively, false negatives are related to erroneous prediction of the absence of TTE, resulting in the selection for microsurgical treatment. Because the first situation has a larger impact on the well-being of the patient, higher emphasis is placed on specificity.

6.4 Results

In this section, first a description is given of the statistical analyses of the patient- and treatment-related characteristics. Next, the feature extraction parameters and results are presented. Finally, the classifier performances are discussed with regards to balancing of the training data, the employed features, and the tumor volume filtering.

6.4.1 Statistical analyses

For the statistical analyses, all patient- and treatment-related characteristics of the included patients were obtained from a prospectively collected database. A

summary of the resulting characteristics can be found in Table 6.1.

First, Student's *t*-tests are employed for evaluating differences in patient- and treatment-related characteristics between patients suffering from TTE and those that do not show TTE. These tests are also performed after implementing the additional volume thresholds. The resulting p-values are presented in Table 6.2. None of the tests obtained statistical significance ($p < 0.05$), showing that the patient- and treatment-related characteristics have no prognostic value for the occurrence of TTE. This is fully in agreement with the found literature on the subject.

6.4.2 Classification performance

This section presents the implemented feature parameters and results obtained per feature extraction method. For each extractor, we evaluate the impact of the volume thresholding, as discussed above. First, the FOS results are elaborated. Next, the results of the MFs are presented, and finally the GLCM-based results are given.

A. First-order statistics (FOS)

The calculated FOS of the MRI scans are the mean, standard deviation, skewness, and kurtosis. Furthermore, a 16-bin histogram is included, resulting in a total of 20 features per MR image sequence. In Table 6.3, the performance of both training strategies, including all available data and including balanced data, of the best FOS-based models are presented for the various volume thresholds. For the FOS-based features, the model that is trained with balanced training data achieves a sensitivity and specificity of 0.72 and 0.40, respectively. Excluding tumors smaller than 7 cm^3 from the training data improves this performance slightly to values of 0.66 and 0.58, respectively. However, these results do not provide a significant improvement over random classification.

Additionally, training the SVM models on all available data results in models that are skewed towards the majority class. This is clearly visible in Table 6.3, where the values for the specificity are all significantly below 0.50 in all but one of the best-performing models. The model of exception obtained sensitivity and specificity values of 0.73 and 0.52, respectively. It can therefore be concluded that, regardless of the balancing step, these experiments clearly show that FOS-based features are not well-suited for predicting TTE.

B. Minkowski functionals

The Minkowski functionals provide an alternative feature description of the data and are computed as a function of the binarization threshold T. These computations can be performed for all available discrete levels, though the functionals show high correlation between subsequent thresholds when the difference between thresholds is small. Therefore, we employ 9 threshold levels equally spaced within the unity interval, ranging from 0.1 up to 0.9. The resulting 36 MF features

Volume threshold (>)	-	2 cm^3	3 cm^3	4 cm^3	5 cm^3	6 cm^3	7 cm^3
Number of patients (TTE – non-TTE)	38 – 61	34 – 58	31 – 45	25 – 41	24 – 37	19 – 32	17 – 26
Age	0.315	0.514	0.696	0.604	0.614	0.643	0.149
Tumor volume	0.527	0.513	0.142	0.332	0.191	0.254	0.121
Dose to 99% of tumor vol.	0.152	0.145	0.094	0.126	0.204	0.202	0.301
Coverage	0.581	0.739	0.590	0.681	0.672	0.993	0.782
Selectivity	0.909	0.908	0.919	0.980	0.761	0.901	0.739
Gradient index	0.383	0.443	0.248	0.280	0.225	0.378	0.595
Paddick conformity index	0.961	0.989	0.954	0.932	0.774	0.740	0.757
Number of isocenters	0.792	0.645	0.786	0.499	0.687	0.768	0.819
Beam-on time	0.548	0.630	0.550	0.853	0.504	0.611	0.990

Table 6.2 — *Resulting p-values of Student's t-tests for general clinical parameters and for increasing volume thresholds. In the second row, the number of patients after each volume threshold is given. None of the p-values reach statistical significance.*

Volume	Balanced		Full	
Threshold (>)	Sensitivity	Specificity	Sensitivity	Specificity
-	0.72	0.44	0.84	0.34
2 cm^3	0.48	0.63	0.87	0.29
3 cm^3	0.47	0.70	0.73	0.52
4 cm^3	0.63	0.50	0.83	0.40
5 cm^3	0.46	0.65	0.95	0.25
6 cm^3	0.67	0.55	0.94	0.32
7 cm^3	0.66	0.58	1.00	0.35

Table 6.3 — *Performance scores of SVM models trained with FOS features for various volume thresholds and two data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).*

are computed per MRI sequence. For training the SVM models, we employ each functional M_i^T for $i = 0, \ldots, 3$ individually, as well as all four combined. The performance metrics of the best MF-based models are given in Table 6.4. These results are obtained from various different models, where the applied MF features are varying per case. These experiments already indicate that the MF features are not very selective in finding the discriminating information, but the information can be retrieved by specific combinations of them. Furthermore, for each volume threshold, different SVM strategies may perform best. This may enable the search for an ensemble approach of using MF features with specific SVM combinations to enhance performance. This search is beyond the scope of our experiments. Instead, we focus on the performance of the best model employing MF features, combined with a balanced training set, which results in sensitivity and specificity values of 0.69 and 0.53, respectively. Implementation of the volumetric threshold defined at 7 cm^3 slightly increases these metrics to 0.64 and 0.61, respectively. The impact of the imbalance in the dataset is less influential for the MF-trained models, compared to the FOS-trained models, although some models still are skewed towards the majority class. The highest-performing MF-based model trained on all available data obtains sensitivity and specificity values of 0.80 and 0.60, respectively. These values are found with a minimum volumetric inclusion criterion of 4 cm^3.

C. Gray-level co-occurrence matrices (GLCMs)

Generally, GLCMs are evaluated for the four unique two-dimensional (2D) directions, chosen as $\theta \in \{0°, 45°, 90°, 135°\}$. Specific subsets may be chosen based on existing clinical knowledge. However, such clinical information is not available, due to the fact that influencing factors of TTE are unknown. Since 3D MRI scans are available, we employ the 3D extension of the GLCM directions. Each direction is separated by a 45° rotational offset on the cardinal planes, resulting in 13 unique GLCM directions per MRI scan. The second parameter of the GLCM, distance d, is evaluated for the integer values $1, 2, \ldots, 6$. The upper bound of 6 is

Volume	Balanced		Full	
Threshold (>)	Sensitivity	Specificity	Sensitivity	Specificity
-	0.69	0.53	0.82	0.50
2 cm^3	0.74	0.50	0.80	0.50
3 cm^3	0.63	0.59	0.93	0.42
4 cm^3	0.61	0.60	0.80	0.60
5 cm^3	0.63	0.56	0.73	0.63
6 cm^3	0.60	0.67	0.87	0.47
7 cm^3	0.65	0.65	0.69	0.64

Table 6.4 — *Performance scores of SVM models trained on MF features for various volume thresholds and two data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).*

Volume	Balanced		Full	
Threshold (>)	Sensitivity	Specificity	Sensitivity	Specificity
-	0.69	0.75	0.82	0.69
2 cm^3	0.64	0.76	0.76	0.65
3 cm^3	0.68	0.73	0.84	0.61
4 cm^3	0.70	0.75	0.88	0.64
5 cm^3	0.67	0.75	0.89	0.67
6 cm^3	0.71	0.79	0.77	0.89
7 cm^3	0.79	0.75	0.85	0.75

Table 6.5 — *Performance scores of SVM models trained on GLCM-based features for various volume thresholds and two data selection methods. Training data is either a balanced subset (Balanced) or the entire set (Full).*

chosen to align with half the size of the smallest tumor dimension in the dataset, so that the distance is clearly embedded within the size of larger tumors. The third parameter, i.e. the maximum number of quantization levels N_ℓ, affects the fine details retained in the input image. This parameter is evaluated for values with powers of 2: $2^2, 2^3, \ldots, 2^6$. Implementing all parameter combinations yields a total of 390 unique GLCMs per MRI sequence.

The GLCM features employed in training a single SVM model are composed of the entropy, contrast, energy, and correlation, calculated from a single GLCM. Given the number of GLCMs per MRI sequence and the number of SVM types, a total of 9,360 GLCM-based models are trained. Table 6.5 shows the results of the top-performing models for each data inclusion setting.

Initial tests with GLCM-based features were performed with a balanced training set. Without volumetric exclusion of the data, sensitivity and specificity values of 0.69 and 0.75 are obtained, respectively. Applying volumetric thresholds on

the data improves the model performance. A sensitivity and specificity of 0.79 and 0.75 are obtained, respectively, when implementing the maximum volume threshold.

Next, the effect of data balancing is explored. Utilizing the full dataset, in contrast to a balanced subset, increases the number of included samples by more than 20%. The effect of these additional data yields a performance improvement, increasing sensitivity and specificity to 0.82 and 0.69, respectively. For a minimum volume inclusion criterion of 6 cm^3, the highest sensitivity and specificity values of 0.77 and 0.89 are obtained, respectively. From these results, it can be concluded that GLCM features contain the most predictive information of TTE, compared to FOS- and MF-based features. Application of all training data compared to a balanced training set, only slightly improves the performance of these GLCM-based models.

However, the largest effect on the performance for these models is imposing more strict volumetric data thresholds. Table 6.5 shows this influence on the highest obtained sensitivity and specificity. A different metric employed to evaluate model performance is the area under the curve (AUC) of the receiver operating characteristic (ROC). In Figure 6.4, the ROC curves of the best GLCM-based models for the various volume thresholds are depicted. Here, all thresholds obtain similar performance. The models obtain AUC values of approximately 0.90 – 0.95. These results are validated by performing bootstrapping[2], resulting in confidence intervals of approximately 0.80 up to 0.99.

Furthermore, the models obtaining these results show large variations in their parameters. Among the seven best models, one for each volume threshold, all three image modalities perform best at least once. Additionally, all quantization levels show the same effect, where each level is implemented at least once in the highest-performing models. Because the data mainly differ in volume between these models, the large variations between parameters indicate the presence of information at various levels. A combination of models and features may prove to further enhance the results.

6.5 Discussion

The research in this chapter was performed in order to find predictive features of transient tumor enlargement (TTE) after GKRS treatment of VS. More specifically, several experiments were conducted to examine the possible prediction of this adverse effect after GKRS. This possibility is relevant because if this effect can be *a-priori* predicted, a different treatment strategy may be considered.

First, this discussion starts with a summary of the obtained results. After this, specific limitations of this investigation are addressed in a broader perspective.

[2]Bootstrapping is a statistical procedure that resamples a single dataset to create numerous simulated samples. This allows for constructing confidence intervals.

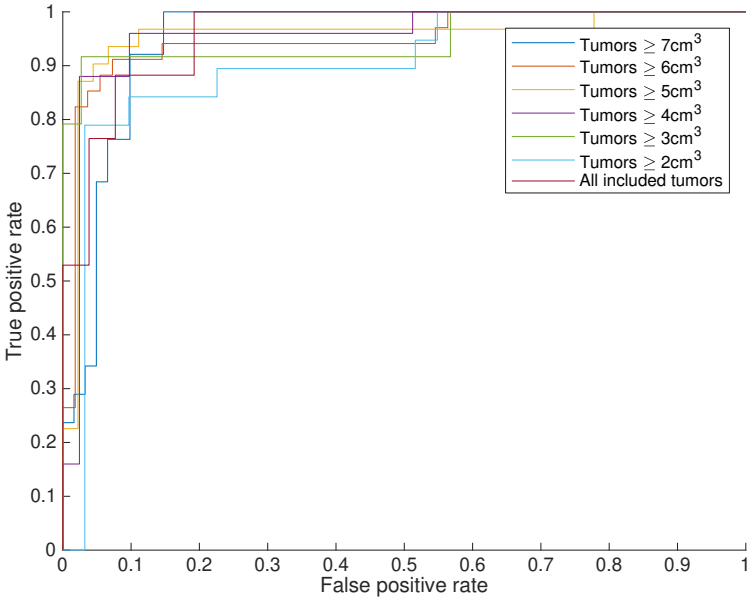


Figure 6.4 — *Receiver operating characteristic curves of the best performing SVM classification, based on GLCM features, per volume threshold setting. These results are obtained with a balanced training strategy (figure best viewed in color).*

Furthermore, the robustness of our findings are discussed.

6.5.1 Obtained results

Previous studies investigated this problem from a clinical point-of-view [33], [52], [123]–[136], [212]. These studies did not find decisive correlations. Moreover, several studies contradicted the results previously found in other studies. Therefore, at present it remains unknown if prediction of TTE is possible. Our study is able to achieve a classification sensitivity and specificity of 0.82 and 0.69, respectively. When employing volume thresholding, we have obtained improved performances for increasing volumes. For tumors larger than 6 cm^3, a sensitivity and specificity of 0.77 and 0.89 are realized, respectively. These results have been obtained by employing features from individual gray-level co-occurrence matrices (GLCMs) and represent the highest scoring models. Additionally, multiple models based on individual GLCMs have achieved promising classification results. Combining features from these individual GLCMs may improve the presented results and enable prediction of TTE with even higher accuracy, sensitivity, and specificity. Furthermore, we have determined that features calculated from different MR sequences also show promising results. Thus, besides combining features calculated from individual GLCMs, the results can also be improved when using the combination of features from different MR sequences.

6.5.2 Feature extraction methods

The three feature extraction methods implemented in this study are selected both on technical and on clinical aspects. Technically speaking, the implemented features and classification method have a proven track record in other healthcare image analysis applications, for instance in oncology. Both GLCM and Minkowski functionals attempt to measure local changes in gray-level texture within the MRI image, thereby addressing heterogeneous properties of the tissue. Furthermore, SVM has shown to be effective in binary classification problems without requiring large amounts of data. We are aware that these techniques are at present outperformed by machine learning using convolutional neural networks (CNNs). However, since this work is the first to explore this data and the related research questions, the dataset was inherently limited at the beginning, which prevented straightforward application of this new machine learning technology. In this view, the current work can serve as a good baseline benchmark. Furthermore, we consider that starting with such an advanced technique, clinical application would only be accepted if the learning network would provide what is actually learned from the data. Since our exploration has indicated important features to be used as a reference, we have learned what is important in the images and this knowledge can be further exploited in developing so-called explainable artificial intelligence.

Coming back on the employed feature extraction techniques, but now from a clinical point of view, we remark that they are based on the supposition of the neurosurgeons that perform the GKRS treatment of VS tumors in the Gamma Knife center. They surmise that enhancing tumors with inhomogeneous texture properties show different behavior than the homogeneously enhancing tumors. More specifically, inhomogeneity in the form of dark streaks and dark areas within the enhancing lesion are considered to be the most informative visual properties. Thus, we have selected the three described feature extractors, since these can adequately quantify such forms of heterogeneity. However, the results in this chapter may further improve by investigating other texture features employed in radiomics analysis of medical images, which is a point of further research. Some examples of such alternative features are gray-level size zone matrices [88], or frequency-based methods like Gabor wavelet analysis [214].

6.5.3 Retrospective character

A significant confounder in this research is its retrospective character. One of the disadvantages of the retrospectively analyzed data is that MR image intensities can vary between subjects, because MR protocols and scanners may have changed in the course of time. Despite our attempt to minimize the impact of the inter-subject MR intensity variations by implementing an advanced normalization method, these variations may still be present in the prediction approach, albeit at reduced level.

Another confounder in this work is the applied definition for TTE. As stated by Marston *et al.*, TTE is difficult to differentiate from true tumor growth [49] Ideally, a histopathological examination of tumor tissue obtained from resection

is employed for determining the ground-truth labeling. However, in the case of a VS, surgical intervention is only warranted if the mass effect of the tumor causes life-threatening issues. In all other cases, the transient swelling is accepted and carefully followed-up. Thus, only the volumetric data obtained from the follow-up MRI scans can be used for determining TTE. Nevertheless, this may have caused an incorrect labeling of the data, creating uncertainty in the final classification results.

6.5.4 Inter- and intra-observer variations

Moreover, inter- and intra-observer variations in measuring tumor volumes make the determination of true volume changes difficult, as discussed in Section 2.5. During treatment planning, the VS tumor is segmented by the treating neurosurgeon, using a semi-automatic contouring tool incorporated in the treatment planning software. In the course of time, a total of 6 different neurosurgeons treated VS tumors at the Gamma Knife center. Furthermore, the follow-up MRI scans were segmented using the same tool by one neurosurgeon and one researcher. An in-house evaluation of the inter- and intra-observer variations demonstrated that these variations decrease for increasing volumes. For volumes larger than 1 cm^3, this variation reduces to less than 10%. This, together with the confounding ground-truth labeling, motivates why we have only included patients who presented obvious TTE and obvious non-TTE. This strict selection has been implemented to create two distinct cohorts. However, this definition may have caused a selection bias that has influenced the obtained results.

Furthermore, the inter- and intra-observer variations also influence the amount of voxels included in the feature extraction algorithms. However, due to the employed method for tumor segmentation, the variations in contouring are found in the so-called partial-volume effects of the MRI scans. These variations are considered to have a limited impact, because they constitute less than 10% of the total amount of voxels and because features are calculated globally. Nevertheless, it could have influenced the calculated features and the obtained results.

6.5.5 Robustness of the results

The above-described confounding factors may have influenced the obtained results and the robustness of it. Currently, the Gamma Knife center in Tilburg is the only center in the Netherlands treating this type of brain tumor using GKRS. As such, we assume that we have selected a good cross-section of all VS patients in the Netherlands. Thus, the results found in this chapter are most likely applicable to other Gamma Knife centers as well. However, the obtained results need to be validated, preferably in a joint multi-center setting. This would ensure that these confounding factors are reduced, thereby improving the robustness of the obtained results. Furthermore, a prospective study could be designed to cope with the previously described problems. Nonetheless, the results achieved in this study strongly suggest the possibility of TTE-prediction for individual treatment selection, making an implementation of this in the clinical workflow conceivable.

6.6 Conclusions

At present, small-to-medium size vestibular schwannomas (VSs) are generally treated using Gamma Knife radiosurgery (GKRS), since the treatment goal for these tumors has shifted from complete removal with inherent risks for the cranial nerve functions to less invasive techniques such as GKRS. However, for large VS tumors, microsurgical excision remains the preferred treatment strategy. Since the risk involved in microsurgery can be contra-indicative for this strategy, less invasive treatments such as radiosurgery and radiotherapy have been considered increasingly in the last decade obtaining good results with acceptable radiation-induced morbidities. However, it remains a controversial alternative to microsurgery, since one of the major contra-indications for GKRS on large VSs is the adverse effect of transient tumor enlargement (TTE). Therefore, the possibility of predicting TTE would be extremely beneficial as this would enable the selection of the most optimal treatment strategy on an individual basis.

It is hypothesized that the origin of this phenomenon can be found in the variations in individual tumor biology. We have explored the idea that the various tumor appearances on MRI reflect variations in tumor biology. Therefore, we have employed quantitative MRI texture features derived from conventional MR images in this research.

A comparative study on feature extraction methods has revealed that MRI tumor texture can provide information for predicting TTE. In this study, the information contained in MRI texture is best captured by GLCM features. Using these texture features extracted from MRI data, we are able to obtain classification sensitivity and specificity values of 0.77 and 0.89, respectively. These results clearly show that MRI tumor texture can provide information for predicting TTE. This can form a basis for individual VS treatment selection, further improving overall treatment results. Particularly for patients with large VSs, where the phenomenon of TTE is most relevant and for which our predictive model performs best, these findings can lead to an implementation in a clinical workflow such that for each patient the optimal treatment strategy can be determined.

The current chapter has shown that the prediction of TTE is feasible. However, TTE is not the only controversial effect of Gamma Knife radiosurgery on vestibular schwannomas. Another important critique that this modality receives, concerns the long-term treatment goal of VS tumors. In approximately 5 – 20% of the cases, tumor progression is not stopped and intervention is needed. This motivates that the predictability of long-term tumor control is investigated in the following chapter.

Chapter 6

Chapter 7

Long-term tumor control

7.1 Introduction

The previous chapter has discussed the difficulty in selecting the optimal treatment strategy for vestibular schwannomas (VSs), specifically for large tumors. It has introduced the first experiments in determining predictive factors for the transient tumor enlargement (TTE). This adverse effect is one of the reasons why microsurgical resection is considered for large VS tumors, even though several publications show that stereotactic radiosurgery (SRS) can be safely employed in the treatment of large VS tumors. Therefore, we have presented the possibility of predicting the TTE effect prior to treatment, by implementing MRI tumor texture features calculated by gray-level co-occurrence matrices (GLCMs). Using these features, in combination with support vector machine (SVM) learning, we are able to achieve a classification sensitivity and specificity of 0.77 and 0.89, respectively.

When taking a birds-eye view on the problem of TTE, this chapter is concentrating on a different direction. The prediction of the TTE effect is clinically highly beneficial in aiding physicians and patients in selecting the optimal care path, especially for patients in which the VS tumor already causes pressure on the neighboring critical brain structures. A transient swelling may cause life-threatening issues in these cases and should be avoided. However, the prediction of not developing TTE does not automatically imply that SRS will stop tumor progression. This long-term treatment goal is separate from the short-term TTE effect and necessitates a life-long follow-up of each patient. Furthermore, if tumor progression is not stopped, salvage microsurgical treatment may be needed. Such an intervention negates the reasons for electing SRS over microsurgery in the first place. Moreover, microsurgical excision of tumor tissue following SRS is considered more difficult than if the tumor was not irradiated [38]. Because SRS has lower overall post-treatment morbidity and overall increased preservation rates of cranial nerve functions when compared to microsurgery [215], it can be concluded that SRS is an attractive treatment modality for VSs, but only if the tumor responds well to this treatment.

Currently, it is not possible to *a-priori* predict the long-term SRS treatment response of a VS on an individual basis. To enable such prediction, tumor-specific information should be assessed. This ranges from macroscopic scale structure of the tumor to genetic profiling obtained by performing a biopsy [62]. However, in

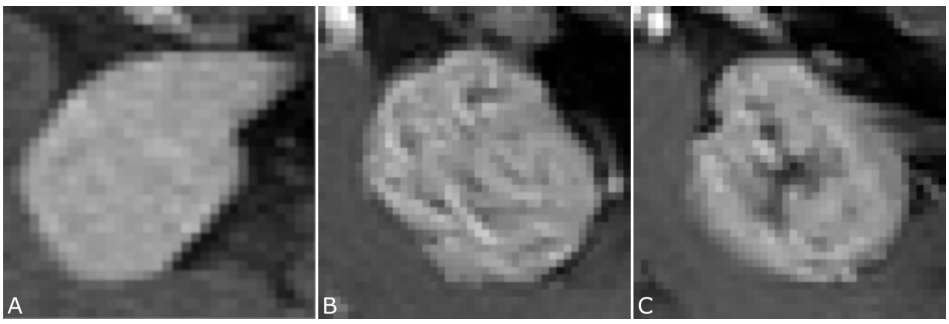


Figure 7.1 — *Examples of different vestibular schwannoma tumor textures appearing in contrast-enhanced MRI scans. The depicted tumors have comparable volumes. Part A: Near-homogeneously enhanced lesion. Part B: Small irregularities in texture. Part C: Heterogeneously enhanced lesion with an apparent hypo-intense area.*

the case of a VS, biopsy is not necessary for diagnosis. In fact, it poses a significant risk of complications due to surrounding critical neurovascular structures, making it an undesired procedure. Therefore, predictive tumor-specific parameters for SRS response in individual VS patients have to be obtained from readily available clinical data. For VS tumors, such data are available in the form of magnetic resonance imaging (MRI) scans. It is well-known that the MRI findings in VSs are highly variable in the gray-level inhomogeneity of the tissue itself. VS tumors can appear as micro- or macro-cystic [63], hemorrhagic [64], and with variable contrast-enhancement patterns. Some appearance examples can be found in Figure 7.1. These MRI appearances reflect variations in histology, such as cell proliferation and micro-vessel density [65], [66]. As such, these MRI images may provide sufficient information to enable the individual prediction of the SRS treatment response.

In this chapter we therefore investigate quantitative, tumor-specific parameters obtained from conventional MRI scans. These so-called *radiomic* features may provide information on differences in tumor biology, enabling the creation of a patient-specific tumor model that can be employed for predicting the long-term SRS treatment response. The aim of this chapter is to explore the prediction of long-term tumor control, employing radiomic features obtained from MRI scans. As a refinement of this general problem statement, we list the following aspects.

- *Patient-inclusion criteria*: In order to train a prediction model, MRI data of a sufficient number of patients need to be available. There are several trade-offs that influence the amount of patients that can be included. A strict definition of treatment failure and long-term tumor control are needed in order to prevent misclassifications.
- *Machine learning*: Several texture features will be analyzed and subsequently machine learning is applied as a technique for classifying the MRI observations.

- *Tumor-control prediction*: To facilitate prediction, a model is trained that enables the *a-priori* prediction of long-term tumor control on an individual basis. This model is then evaluated and, if possible, implemented in a clinical decision-support system, aiding physicians and patients in selecting the optimal care path.

This chapter is outlined as follows. First, Section 7.2 elaborates on the available patient data and the treatment protocol. Next, Section 7.3 describes the experimental setup for evaluating the predictability of long-term tumor control. Then, the results of the experiments are presented in Section 7.4, after which they are discussed in Section 7.5. Finally, Section 7.6 concludes this chapter.

7.2 Patient cohort and treatment protocol

This section introduces the data employed in this chapter and presents the applied treatment protocol. Furthermore, it discusses the employed classification labels based on objective treatment failure and long-term tumor control definitions.

7.2.1 Patient cohort

All patients with unilateral VS treated with SRS in our center between 2002 and 2014 were identified. This cohort consisted of all VS patients remaining after excluding patients with neurofibromatosis Type 2 (NF2), those previously treated for their VS, or with less than two years of post-SRS follow-up. Furthermore, we excluded patients with small VS tumors, because these tumors show little to no variation in texture. The associated volumes of such small tumors are ill-defined. Analysis of the entropy of voxel intensity variations in our data showed that tumor texture becomes discernible in tumors around one cm^3, see Figure 7.2. Since an exact volumetric cut-off is arbitrary, we opted for an approach based on the controversy to treat larger tumors with SRS. Thus, we investigated all tumors with a minimum volume of 1.42 cm^3, because this corresponds to the volumes reported in literature for Koos Grade IV tumors [137].

7.2.2 Treatment and follow-up

Stereotactic radiosurgery was performed using the Leksell Gamma Knife® model 4C or Perfexion (since November 2008; both Elekta AB, Stockholm, Sweden). A dose of 13 Gy was prescribed to the isodose line covering 90 – 99% of the tumor volume. For treatment planning, T1-weighted with (T1CE) and without (T1) Gadolinium administration, and T2-weighted (T2) MRI scans were obtained for each patient. Treatment was carried out in a single fraction with frame-based fixation.

After treatment, each patient was subjected to a follow-up schedule with a standard interval of one year. A T1CE MRI with a slice thickness of one mm was obtained at each follow-up visit. In the case of suspected radiological progression or new or worsening symptoms, the standard interval was reduced. If the

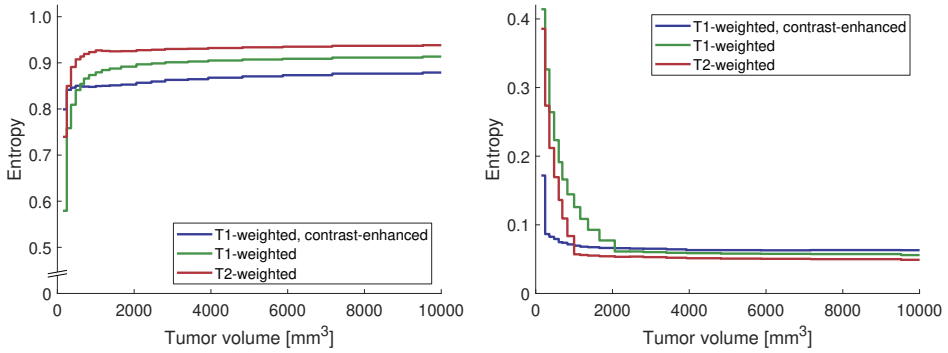


Figure 7.2 — *Entropy analysis of tumor voxels of all 735 patient data. Left: mean entropy of tumor voxel intensities, captured in 35 equal-sized volume bins. Right: standard deviation of the entropy values, divided over the same 35 bins.*

tumor displayed radiological regression or stability for several years, the standard interval was extended. All follow-up MRIs were employed for determining a volumetric treatment response. Tumor volume were determined by segmenting the tumor using the treatment planning software (GammaPlan Versions 10 and 11, Elekta AB, Stockholm Sweden).

7.2.3 Definitions of treatment outcome

A widely applied approach for obtaining models that can predict treatment outcomes, is supervised machine learning (sML) [216]. This technique employs training data with pre-determined classification labels to discover and identify specific patterns, possibly even indiscernible to the human eye. With these findings, such techniques can distinguish cases in the training data with correct labels. Thus, it is crucial that the input data has high-quality labels (i.e. a high certainty of correctness of the assigned classification label), such that the trained model is robust. Therefore, a well-defined classification labeling is needed in these experiments. These labels are defined as long-term tumor control, and true tumor progression. Consequently, strict definitions for both controlled and progressive tumors are required. These definitions were discussed in Section 2.4 and are summarized below. Patients that could not be labeled according to the strict criteria were excluded from further analysis.

True tumor progression was detected using linear measurements in a clinical setting. An increase in tumor size was accepted and considered as radiation-induced swelling during the first two years after treatment [131], [134], unless the enlargement was deemed too excessive for the considered patient by the radiosurgical team. These failures may have been the result from swelling and not due to true tumor progression. Since radiation-induced swelling is a radiobiological distinct response from true tumor progression, we excluded patients that had salvage treatment within the first two years following SRS, to avoid inaccuracy through misclassification. Volumetric progression after this period was considered

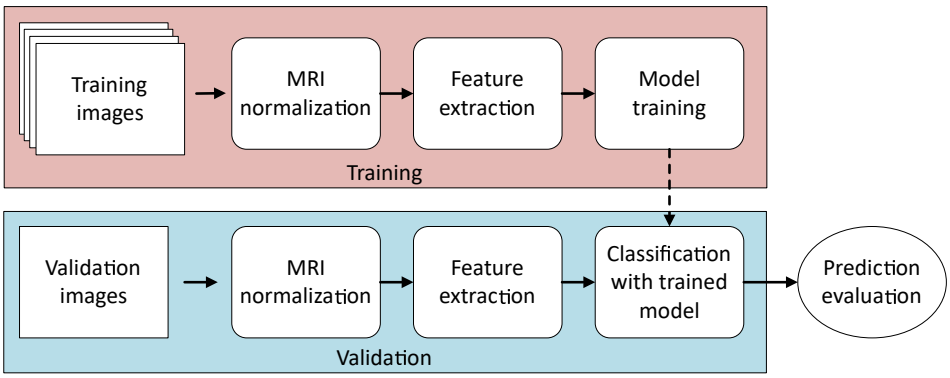


Figure 7.3 — *Processing block diagram of the experimental setup.*

true tumor progression. This was always confirmed by the radiosurgical team. In addition, we looked for discrete volume increases after the initial two years, which were undetected by linear measurements performed in the clinical setting. However, these increases were detectable with the volumetric analyses performed in this research. The potentially missed failures are defined as two consecutive significant increases in tumor volume among three proceeding follow-up MRIs, where a minimum of 10% increase in volume is deemed significant [97].

A definition for long-term tumor control highly depends on the time-period in which treatment failures may still occur. However, there is no certainty that a specific VS will not progress anymore after a predetermined time. Hence, long-term tumor control cannot be specified without concessions. Therefore, in this research, we have defined it as absence of progression beyond 129 months following treatment. This cutoff is based on the latest-occurring failure in our large database, which was identified at 129 months after treatment.

7.3 Experimental setup

For evaluating the predictability of long-term tumor control following SRS, several experiments are conducted. These experiments involve the evaluation of MRI tumor texture features and their relations to long-term tumor control. Figure 7.3 depicts the processing block diagram of the experimental setup. In this section, we will discuss each processing stage in this figure. First, in Section 7.3.1, we will elaborate on the input MRI images used in this chapter. Next, we will present the employed feature extractors in Section 7.3.2, followed by a discussion on the model training approach in Section 7.3.3.

7.3.1 MRI input data

For each patient, the T1-, T1CE-, and T2-MRI scans (Intera® and Ingenia®, both Philips Healthcare, Best, the Netherlands) were extracted from the treatment planning system, including tumor contours drawn by the neurosurgeon during treatment planning. The matrix sizes of these MRIs were 256×256 pixels per axial slice

for the T1 and T1CE, and 512×512 pixels for T2 scans.

As discussed in Section 6.2.2, MRI scans need to be normalized, in order to enable comparison between patients. The same generalized intensity-scale method is employed as in the previous chapter. This method performs piecewise-linear normalization utilizing tissue-specific landmarks. These landmarks included the brain stem and the stereotactic G-frame fiducial markers (Elekta AB, Stockholm, Sweden) in the T1 and T1CE scans. Additionally, for T2 scans, the cerebrospinal fluid is included as landmark.

7.3.2 Feature extraction methods

For evaluating the predictability of long-term tumor control, several texture features are extracted from the available MRI scans. These radiomic features are as follows.

- *Twenty first-order statistics (FOS)* features: statistical properties (e.g. average and variance) of all voxel values, ignoring spatial interaction between image voxels.
- *Four Minkowski functionals (MFs)* [162], [164]: morphological properties from groups of voxels whose intensities are above a specific threshold.
- *Thirteen gray-level size zone matrix (GLSZM)* features [88]: statistical properties on the size of homogeneously enhanced zones for each gray-level, depending on their quantization levels.
- *Four gray-level co-occurrence matrix (GLCM)* features [86]: spatial distribution properties of gray-levels in the image voxels, depending on inter-voxel distances, viewing angles, and their quantization levels.

The illustrations in Figure 7.4 demonstrates a graphical explanation of the MFs and Figure 7.5 shows the MF concept on an MRI scan. Furthermore, Figures 7.6 and 7.7 depict explanations of the GLSZM calculations and of the GLCM calculations, respectively, using a simple two-dimensional example image. These methods have been explained in more detail in Section 2.7. All features are calculated on the complete tumor, thus in three-dimensional space for all feature extraction methods.

7.3.3 Machine learning approach

The extracted features are applied to an sML stage to train predictive binary classification models, classifying either true tumor progression or long-term tumor control. Training is performed on a single feature vector, i.e. an array of numerical values, for each individual feature extraction method. This is to prevent overfitting, caused by too many included features. If the number of features is large

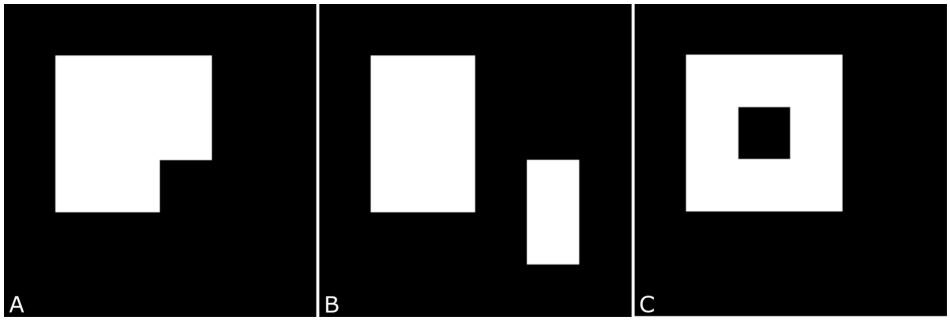


Figure 7.4 — *Visualization of the calculations of the Minkowski functionals in two-dimensional images. In these simple images of 6×6 pixels, we can calculate the number of white pixels (K), the length of the boundary of all white shapes (L) and the number of white shapes minus the number of black shapes within white shapes (M). These three values then represent the image contents. Part A: $K = 8$, $L = 12$, and $M = 1$. Part B: $K = 8$, $L = 16$, and $M = 2$. Part C: $K = 8$, $L = 16$, and $M = 0$.*

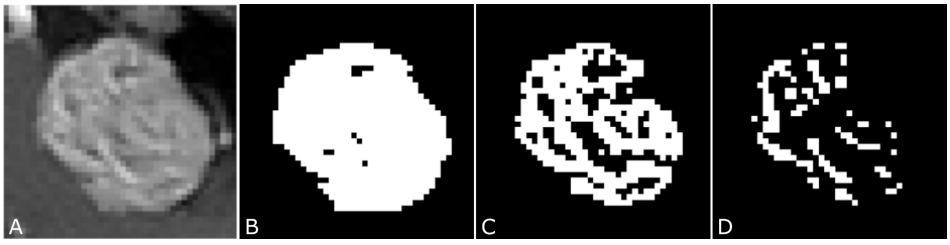


Figure 7.5 — *Visualization of the Minkowski functionals (MFs) for a VS tumor. For three-dimensional images like MRIs, four metrics can be calculated in the same way as depicted in Figure 7.4: (1) the number of voxels, (2) the number of open faces, (3) the number of open edges, and (4) the number of open vertices. This can be done for multiple threshold levels, resulting in different binary representations of the tumor. Here, Part A represents the original T1-weighted, contrast-enhanced MRI image. Parts B, C, and D are generated with thresholds T equal to 0.5, 0.7, and 0.8, respectively.*

compared to the number of tumors, each tumor can be uniquely identified by one individual feature vector. This results in high prediction scores, but the resulting model may have unreliable performance on new data. A single feature vector is individually created per MRI modality. Furthermore, for the combination of all three MRIs, the individual feature vectors are combined. Together with all the different parameter settings for each feature extraction and the different settings for the employed sML algorithm, a large number of models are trained and evaluated. All parameters, settings, and calculated features per feature extraction are listed in Table 7.1. For implementation of support vector machines, training is carried out by the classification learning application from MATLAB (MathWorks inc., Natick, Massachusetts, USA).

With growing tumor volumes also the number of tumor voxels increases,

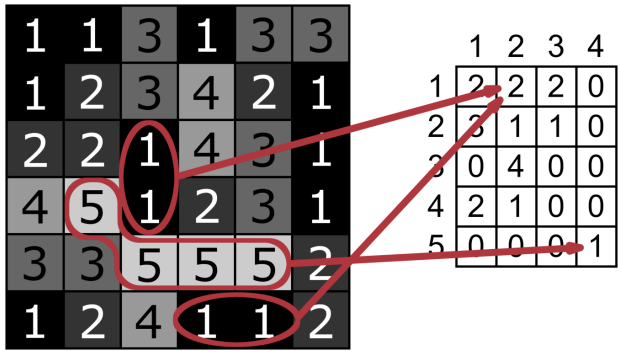


Figure 7.6 — *Graphical representation of the calculations of the gray-level size zone matrices (GLSZMs). For the GLSZM, the number of zones is counted, where a zone contains equally-valued connected pixels and has a specific size in pixels. The pixel values depend on the number of quantization levels, and for different levels distinct GLSZMs can be calculated. In this figure, some example zones are highlighted. These zones consist of connected pixels with the values "1" and "5" of size 2 and 4, respectively. As there are 2 zones with pixel value "1" of size 2, the corresponding position in the resulting matrix becomes "2". The same can be computed for the zone with the value "5". This results in a "1" on the corresponding position in the matrix. For each individual GLSZM, a single feature vector is calculated incorporating the above-described statistics, which is then employed for training.*

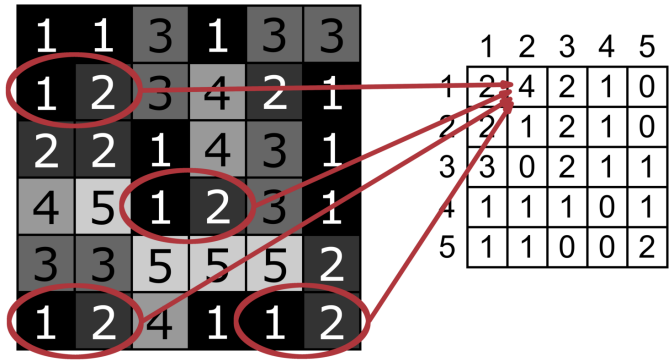


Figure 7.7 — *Graphical representation of the calculations of the gray-level co-occurrence matrices (GLCMs). The GLCM is calculated by counting inter-pixel relations. These relations depend on (1) the inter-pixel distance (in this example equal to unity), (2) the inter-pixel angle (in this illustration 0 degrees) and (3) the number of quantization levels, which is the number of different pixel values (here equal to 5). The resulting GLCM matrix is calculated by counting the number of each specific combination of pixel-pairs. In this example, the pair "1-2" is highlighted. The resulting value in the corresponding position of the final matrix is equal to "4", as there are 4 pairs "1-2". For each individual GLCM, a single feature vector is calculated incorporating the above-described statistics, which is then employed for training.*

Feature extraction	Parameters	Calculated features	# Trained models
FOS	-	Mean Standard deviation Skewness Kurtosis 16-bin histogram	24
MF	Various thresholds: 9 equally-spaced values between 0 and 1	Foreground volume Surface area Curvature Euler number	120
GLSZM	Levels: 2, 3, 4, 5, 6	Small-zone emphasis Large-zone emph. Gray-level non-uniformity Zone-size non-uniformity Zone percentage Low gl. zone emph. High gl. zone emph. Small-zone low gl. emph. Small-zone high gl. emph. Large-zone low gl. emph. Large-zone high gl. emph. Gray-level variance Zone size variance	96
GLCM	Distance: 1, 2, 3, 4, 5, 6 Angle: 13 unique 3D-directions Levels: 2, 3, 4, 5, 6	Contrast Energy Entropy Correlation	9360

Table 7.1 — *Description of all parameters and calculated features for each feature extractor. These include first-order statistics (FOS), Minkowski functionals (MFs), gray-level size zone matrices (GLSZMs), and gray-level co-occurrence matrices (GLCMs). The total number of trained models per feature extraction method is calculated by multiplying the number of options for each parameter, the number of MRI modalities (i.e. 4), and the number of SVM kernels (i.e. 6). For MFs, each feature is a calculated average per MRI modality of the individual functionals for all thresholds. These are employed individually as well as in combination. In this table, gray-level is abbreviated by 'gl'.*

thereby expanding the amount of texture information. Therefore, we have explored the impact of the tumor volume on the prediction results, by imposing various volume thresholds for specific selections of the data. The evaluated volume thresholds are 2, 3, 4, and 5 cm^3.

Validation of the resulting individual models is performed using 10-fold cross-validation. In this method, depicted in Figure 7.3, the model is subsequently trained with 90% of the MRIs and validated on the 10% left-out scans. This is repeated 10 times, each time leaving out a different set of MRIs. This method has resulted in values for accuracy (ACC; correct prediction rate), sensitivity (SENS; proportion of actual controlled tumors correctly identified), and specificity (SPEC; proportion of actual progressed tumors correctly identified). Furthermore, bootstrapping is performed to determine the area under the receiver operating characteristic (AUC; degree of distinction between the two prediction classes), including confidence bounds. An AUC value equal to unity translates to the ability to perfectly distinguish the two classes, while a value of 0.5 can be interpreted as random selection.

7.4 Results

In this section, first a description of the patient dataset is given. Next, the feature extraction parameters and results are presented, including the classifier performances with regards to the employed features and the tumor filtering.

7.4.1 Cohort

After exclusion of patients with NF2, prior treatment, and small tumors, 379 patients were extracted from the database of 735 patients. Of these, 30 patients (7.9%) did not have a post-SRS follow-up beyond two years, either due to salvage treatment for their VS (7 patients, 1.8%), or because they were lost to follow-up (23 patients, 6.1%). After excluding these, a total of 349 patients with a median follow-up of 74 months were identified for this study. Of these 349 patients, 30 (8.6%) needed salvage treatment due to recurrent tumor progression. Of the remaining patients, 13 (3.7%) displayed a volumetric tumor progression. By combining these two groups, a so-called progressed tumor cohort of 43 patients (12%) was defined. A total of 42 patients were identified as having obtained long-term tumor control, according to our strict definition of at least 129 months absence of true tumor progression. As such, a total of 85 patients were included in the model training.

After stratifying for volume, the total number of patients per volume threshold per group can be found in Table 7.2. The results of the optimal model for each individual feature extraction, including the additional volume thresholding, are described in the following subsection.

7.4.2 Evaluation of radiomic features

In this subsection, the results of each feature extraction algorithm are discussed. First, the results from the FOS-based features are presented. Next, the results

Volume threshold (>)	True tumor progression	Long-term tumor control
-	43	42
2 cm^3	38	38
3 cm^3	30	28
4 cm^3	22	18
5 cm^3	18	11

Table 7.2 — *Number of patients in both cohorts after additional volume thresholding.*

Vol. threshold (>)	ACC	SENS	SPEC	AUC
-	0.67	0.74	0.60	1.00
2 cm^3	0.63	0.74	0.53	0.86
3 cm^3	0.73	0.80	0.64	1.00
4 cm^3	0.66	0.68	0.65	0.94
5 cm^3	0.83	0.94	0.64	0.88

Table 7.3 — *Classification results obtained using first-order statistical features and additional volume thresholding (in cm^3). The results are given in accuracy (ACC), sensitivity (SENS), specificity (SPEC), and area under the receiver operating characteristic (AUC).*

obtained with MF-based features are given, followed by the results based on GLSZM-trained models. Finally, we present the results obtained using GLCM features.

A. First-order statistics

The classification scores of the optimal FOS-based models for each volume threshold can be found in Table 7.3. The optimal model trained with FOS features, without any additional volume thresholding, obtains ACC, SENS, SPEC, and AUC values of 0.67, 0.74, 0.60, and 1.00, respectively. The lower- and upper confidence bounds for the AUC values are 0.99 and 1.00, respectively. For increasing volume thresholds, the results increase up to 0.83, 0.94, 0.64, and 0.88 for tumors larger than 5 cm^3, with lower- and upper confidence bounds of the AUC value of 0.59 and 0.96, respectively. Even though the ACC, SENS, and SPEC increase, the AUC value decreases, with a lower confidence bound only slightly above 0.5. This is most likely caused by the reduced number of tumors in the dataset, resulting in a less robust model.

B. Minkowski functionals

The results of the optimal MF-based models for each volume threshold are depicted in Table 7.4. The performance scores of the optimal MF-based model obtain ACC, SENS, SPEC, and AUC values of 0.68, 0.64, 0.73, and 0.96, respectively. The

Vol. threshold (>)	ACC	SENS	SPEC	AUC
-	0.68	0.64	0.73	0.96
2 cm^3	0.71	0.74	0.69	0.91
3 cm^3	0.67	0.73	0.60	0.97
4 cm^3	0.72	0.64	0.82	0.76
5 cm^3	0.76	0.83	0.64	0.88

Table 7.4 — *Classification results obtained using Minkowski functionals and additional volume thresholding (in cm^3). The results are given in accuracy (ACC), sensitivity (SENS), specificity (SPEC), and area under the receiver operating characteristic (AUC).*

Vol. threshold (>)	ACC	SENS	SPEC	AUC
-	0.71	0.79	0.63	0.88
2 cm^3	0.73	0.71	0.74	0.78
3 cm^3	0.67	0.80	0.52	0.99
4 cm^3	0.77	0.86	0.65	0.84
5 cm^3	0.69	0.78	0.55	1.00

Table 7.5 — *Classification results obtained using gray-level size zone matrix features and additional volume thresholding (in cm^3). The results are given in accuracy (ACC), sensitivity (SENS), specificity (SPEC), and area under the receiver operating characteristic (AUC).*

lower- and upper confidence bounds for the AUC value are 0.90 and 0.98, respectively. For growing volume thresholds, the results are increasing up to 0.76, 0.83, 0.64, and 0.88 for tumors larger than 5 cm^3, with AUC lower- and upper confidence bounds of 0.66 and 0.96, respectively. Again, the AUC values decrease for increasing volumes. Furthermore, the specificity decreases significantly as well. However, the lower confidence bound of the AUC value now remains well above the 0.5 value.

C. Gray-level size zone matrix features

The resulting classification scores of the optimal GLSZM-based models for each volume threshold are presented in Table 7.5. For the GLSZM-trained models, the optimal model obtains ACC, SENS, SPEC, and AUC values of 0.71, 0.79, 0.63, and 0.88, respectively. The lower- and upper confidence bounds for the AUC value are 0.75 and 0.94, respectively. For the additional volume thresholds, these results remain comparable to the results of zero threshold. The best results are achieved for tumors larger than 4 cm^3, having an ACC, SENS, SPEC, and AUC of 0.77, 0.86, 0.65, and 0.84, respectively, with AUC lower- and upper confidence bounds of 0.67 and 0.95, respectively.

Vol. threshold (>)	ACC	SENS	SPEC	AUC
-	0.77	0.71	0.83	0.93
2 cm^3	0.75	0.76	0.73	0.92
3 cm^3	0.76	0.80	0.72	0.87
4 cm^3	0.82	0.86	0.76	0.90
5 cm^3	0.83	0.83	0.82	0.99

Table 7.6 — *Classification results obtained using gray-level co-occurrence matrix features and additional volume thresholding (in cm^3). The results are given in accuracy (ACC), sensitivity (SENS), specificity (SPEC), and area under the receiver operating characteristic (AUC).*

D. Gray-level co-occurrence matrix features

The classification scores of the optimal GLCM-based models for each volume threshold can be found in Table 7.6. For the GLCM-based features, the ACC, SENS, SPEC, and AUC values are 0.77, 0.71, 0.83, and 0.93, respectively. The lower- and upper confidence bounds of the AUC value are 0.83 and 0.98, respectively. If additional volume thresholding is applied, these results increase to 0.83, 0.83, 0.82, and 0.99, respectively, for tumors larger than 5 cm^3. With this threshold, the lower- and upper confidence bounds of the AUC value are 0.94 and 1.00, respectively. In contrast with the other feature extraction methods, the AUC values of GLCM features improve for growing tumor volumes, including the confidence bounds.

7.5 Discussion

This chapter has concentrated on finding indicators for prediction of the long-term treatment response of SRS-treated VS patients. Several experiments have been conducted to examine the predictability of long-term tumor control following stereotactic radiosurgery. The ability to *a-priori* predict such a treatment response can significantly impact the treatment selection process and may improve the overall treatment outcome. In this section, the different aspects of the implemented methodology and the obtained results are highlighted and discussed. First, a summary of the obtained results is given, after which some limitations of this research are addressed.

7.5.1 Obtained results

Recently, there is an increasing interest of radiomics in various oncology fields, including brain tumors, in relation to their specific pathology and treatment response [69], [81]–[84]. For VS tumors, only incidental literature is available and concerns the prediction of early treatment response, i.e. radiation-induced swelling [59], [107]. To the best of our knowledge, this chapter presents the first research to focus on the prediction of long-term tumor control following SRS treatment of VS tumors, exploiting MRI-based radiomics. We have evaluated tumor-specific parameters obtained from conventional MRI scans in a large database

with long follow-up, enabling a high-throughput mining of MRI data, which can be subsequently exploited for a machine learning approach.

Furthermore, we have employed objective definitions for true tumor progression and long-term tumor control. The highest obtained prediction accuracy, sensitivity, specificity, and AUC values were 0.83, 0.83, 0.82, and 0.99, respectively. In other words, the best-performing model is able to correctly predict the treatment outcome in our own data in 83% of all cases (accuracy), progressed tumor in 82% of the cases (specificity), and controlled tumor in 83% of the cases (sensitivity). The most predictive features were based on GLCMs and in tumors larger than 5 cm^3. These GLCMs measure the distribution of co-occurring voxels at a given gray-level, revealing certain properties about the spatial distribution of the gray-levels in the image. Furthermore, the obtained results show that the prediction improves for increasing tumor volumes. This is most likely caused by the amount of voxels: larger tumors have more tumor voxels, resulting in an increased amount of texture information. However, for growing volume thresholds, the number of available tumors decreases. This can cause overfitting in the machine learning stage, which results in models that are too fine-tuned to the data, giving high prediction results on the training data. As such, these models may be not sufficiently robust in predicting the treatment outcome on new unseen data.

7.5.2 Definitions of the treatment outcome

A problematic aspect in predicting treatment outcome is the classification of long-term tumor control and true tumor progression, which is crucial in a supervised machine learning approach. Generally, treatments are classified as failed if salvage treatment is needed. Although this is a valid clinical definition, various centers use different motivations prior to considering salvage treatment [97]. Therefore, there is no clear consensus on whether a treatment has failed. Furthermore, subtle progressions may have been missed in the clinical setting, using linear measurements. This is why we have implemented an objective measure for failure, i.e. true tumor progression, using the volumetric tumor response. Moreover, it is hypothesized that radiation-induced swelling is related to different radiobiological aspects of the tumor compared to true tumor progression. To avoid misclassification of this phenomenon, true tumor progression has been defined based on volumetric tumor assessments beyond two years after treatment.

Long-term tumor control is also difficult to define, because tumor progression can occur many years after SRS. Kondziolka *et al.* [138] reported that after four years following SRS, no further increase in volumes were identified. Contrarily, Hasegawa *et al.* [32] reported that only after 10 years of follow-up, no more treatment failures occurred. In our extensive database, we have determined the latest-occurring tumor progression at 129 months following treatment. Therefore, we have adopted this time instance as cut-off for determining tumor control.

The author realizes that these definitions may have had an impact on the obtained results, e.g. tumors now classified as controlled tumor may show tumor progression in the future. However, it can be argued that tumor progression is a

very unlikely phenomenon beyond 10 years after treatment. Nevertheless, this may have impacted the obtained results.

7.5.3 Data limitations

Another possible confounder in this study is its retrospective character. As a consequence of our long-term tumor control definition, the implemented algorithm used MRIs of at least 10 years old to find aspects of tumor texture that distinguish tumor control from progression. However, MRIs have improved in the course of time. It is therefore possible that the currently employed conventional MRIs exhibit more detailed radiomic features, leading to improved SRS outcome predictions.

Furthermore, the obtained results are based on the data from a single institution. As such, the obtained models may suffer from input bias errors. Moreover, overfitting of the trained models is a concern for larger tumors, because the number of tumors that meet our strict criteria of long-term tumor control and true tumor progression, decreases significantly with growing volume thresholds. To prove the robustness of the results, the obtained algorithms need to be validated on large datasets from multiple centers with sufficiently long follow-up.

7.6 Conclusions

Currently, it is not possible to *a-priori* predict the long-term SRS treatment outcome of a VS on an individual basis. If possible, this would be highly beneficial for the individual patient and their treating physician, since such a prediction can help in selecting the optimal patient-specific treatment strategy, thereby improving overall treatment results.

It is hypothesized that the differences in treatment response originate from small variations in intrinsic tumor biology. Since taking a biopsy is undesired, we have explored the idea that variations in tumor appearance on MRI reflect variations in tumor biology. Therefore, we have employed quantitative MRI tumor texture features derived from MRI scans in this chapter.

The results obtained in this research show that prediction of long-term tumor control after SRS treatment of larger VS tumors is feasible with the use of radiomic features. We have obtained prediction accuracy, sensitivity, specificity, and AUC scores of 0.83, 0.83, 0.82, and 0.99, respectively, using GLCM-based features. These scores have been obtained in tumors larger than 5 cm^3.

The radiomics-based information can potentially be implemented in a clinical decision-support system. Individual MRI scans can serve as an input to software, which contains a well-trained tumor texture model. Such a system can then present the patient and treating physician with prediction scores, thereby facilitating the selection of a personalized optimal treatment strategy. This will enable overall improvement of the treatment of vestibular schwannomas. This aspect will be further discussed in the concluding chapter.

Chapter 7

Chapter 8 Conclusions

8.1 Conclusions of the individual chapters

This thesis has investigated the possibility of predicting the Gamma Knife treatment of vestibular schwannoma on an individual patient basis. Several contributions to this have been presented, such as the creation of a unique large database, introduction of various treatment outcome definitions, influence of the pre-treatment growth rate on the volumetric tumor response, a novel method for evaluation the heterogeneous Gamma Knife dose distribution, radiomic feature evaluation of VS tumors, and the development of short- and long-term treatment response prediction models. This concluding chapter will first summarize the most important findings of each chapter, after which it continues with addressing the posed research questions. Finally, a short outlook on the Gamma Knife treatment prediction of vestibular schwannomas is provided.

Chapter 2 has presented the state of the art in the evaluation of risk factors related to the Gamma Knife treatment response. The literature review highlights that the obtained results remain inconclusive. Nevertheless, the included risk factors show their potential and may provide information that can be included in more elaborate evaluation methods. Furthermore, this chapter has introduced the unique large database, which has been created for the research in this thesis. This database has enabled (1) the opportunity to assess the correlations of numerous parameters on various treatment outcomes, and (2) the careful analysis of the individual volumetric responses of the included tumors to debate and construct objective treatment outcome definitions. These definitions form the basis of determining predictive parameters. Finally, this chapter is concluded with a technical overview of the methods for machine learning and feature extractions employed in this thesis for creating a treatment outcome prediction model.

Chapter 3 has discussed the pre-treatment growth rate of a VS, which is considered a highly interesting patient-specific risk factor, closely related to the intrinsic tumor biology. Current state of the art reveals that this particular factor has not been researched extensively and methodological differences may explain the contradicting results found. The large number of patients, the long follow-up times, and the volumetric tumor assessments both prior to and after GKRS treatment in

the database have enabled the accurate investigation into the influence of the pre-treatment growth rate on the radiosurgical efficacy. The conducted experiments have shown that the so-called "bending-the-curve" effect is not present in the data. However, the influence of the pre-treatment growth rate of VS tumors on the long-term GKRS treatment effects with respect to the tumor volume is established in various Kaplan-Meier analyses. The resulting tumor control rates confirm the high efficacy of GKRS for slow growing VS. The 5- and 10-year rates are 98.8% and 91.4%, respectively. Conversely, fast growing tumors exhibit significantly lower tumor control rates, i.e. 84.6% and 66.4%, respectively, for the 33%-fastest growing tumors. For these cases, different treatment strategies may be considered. Furthermore, a Cox regression model is constructed that enables the prediction of the risk at treatment failure on an individual basis, thereby emphasizing the impact of the growth rate on the treatment outcome. This model can be implemented in the clinical workflow to facilitate physicians in selecting the optimal treatment strategy on an individual basis.

Chapter 4 has evaluated the various treatment-related parameters and their possible impacts on the treatment response. First, using the extensive database, all global treatment-related factors are investigated for their impact on the treatment response. The performed Cox regression analyses reveal that the dose to the tumor margin correlates to the long-term tumor control, with a resulting risk factor of 0.63. This implies that a lower dose will result in higher risks at treatment failure. However, after analyzing the variations in long-term tumor control among multiple sub-cohorts, this effect is not established. Therefore, it is concluded that the influence is limited within the boundaries of the treatment protocols.

Second, the heterogeneous character of the dose distribution is evaluated. The calculated homogeneity indices show no statistical differences within the data. However, this method ignores the actual spatial dose characteristics. Therefore, a novel method is introduced to enable the inclusion of these characteristics in the treatment response evaluation. The three-dimensional histogram of oriented gradients method yields features that are implemented in a machine learning environment. The resulting model obtains interesting results, with accuracy, true positive rate, true negative rate, and AUC of 77.5%, 80.0%, 75.0%, and 0.79, respectively. These results suggest that the spatial dose distribution has an impact on the treatment response. Nevertheless, these findings need to be more extensively analyzed on larger datasets to validate their impact on the treatment response.

Finally, the tumor segmentation step in the treatment planning is evaluated. Since a minimum radiation dose is prescribed to the tumor margin, inaccuracies in the tumor delineations may lead to underexposure of these tumor margins. As such, experiments are conducted to analyze differences between the treatment delineations and retrospectively created annotations, among two cohorts that have experienced significantly different treatment responses. It is illustrated in this chapter that the delineations of the fast-responding cohort show more variations compared to the retrospectively created annotations, by means of the Jaccard in-

dices. In contrast, the delineations of the failure cohort show less variation. This suggests that tumors that are more difficult to annotate, are expected to obtain a positive treatment outcome.

Chapter 5 has provided the first experiments on the radiomic features of VS tumors. These features are calculated on the treatment planning MRI scans of a limited dataset of 20 tumors showing a significantly fast response and 20 tumors for which the treatment did not stop the tumor progression. First, various tumor shape features are evaluated for their predictive value on the Gamma Knife treatment response. It is concluded that the shape appears to be a weak predictor. Both SVM- and DT-trained models indicate that classifying the treatment response of a VS, based on the calculated shape descriptors, do not provide a significant improvement over random classification. This shortcoming is explained by the observation that the inter-class differences, i.e. variations between the failure and the fast-responder cohort, are limited, whereas intra-class differences are considerable. Hence, the variations between both classes are deemed too restricted, thereby significantly reducing the prognostic value of these shape descriptors.

Second, experiments involving the radiomic tumor texture features show that popular second-order statistical metrics, like gray-level co-occurrence matrices and run-length matrices, are suitable for describing texture and predicting the Gamma Knife treatment response. Nevertheless, these metrics are slightly outperformed by simple first-order statistics, like mean, standard deviation and median, obtaining an accuracy, sensitivity, and specificity of 85.0%. Nevertheless, the best choice for texture description can be made only after performing more extensive analyses on larger datasets. In any case, the experiments from this chapter have provided useful texture measures for successful prediction of the Gamma Knife treatment outcome for VS and invoke further research on the patient-specific evaluation for VS treatment options.

Chapter 6 has presented the experiments on predicting the short-term adverse effect of radiosurgically induced transient tumor enlargement (TTE). This effect is one of the main causes of controversy in the Gamma Knife treatment for large VS tumors, since it can lead to severe life-threatening morbidities. Therefore, it would be extremely beneficial to predict this effect, since this would enable the selection of the optimal treatment strategy on an individual basis. It is hypothesized that the origin of this phenomenon can be found in the variations in individual tumor biology. To this end, we have explored the possibility that the various tumor appearances on MRI reflect such variations. Therefore, quantitative MRI texture features derived from conventional MR images are analyzed in this chapter. In conjunction with this, since tumor texture is size-dependent, this chapter has also explored the effect of volume thresholding. Furthermore, since MRI provides relative data, two distinct normalization methods are examined in this chapter.

The obtained results prove that first-order statistics and Minkowski functionals are not suited for predicting TTE. However, GLCM-based features obtain relevant

results, with sensitivity and specificity values of 0.82 and 0.69, respectively. After applying volume thresholding, these performance results improve. For a minimum volume inclusion criterion of 6 cm^3, the results increase to sensitivity and specificity values of 0.77 and 0.89, respectively. The AUC measurements of the resulting model obtain values of 0.95, with confidence intervals of approximately 0.80 up to 0.99. These results clearly show that MRI tumor texture can provide information for enabling the prediction of TTE. This characteristic can form a basis for individual VS treatment selection, further improving overall treatment results. This holds particularly for patients with large VS, where the phenomenon of TTE is most relevant and for which the obtained model performs best.

Chapter 7 has extended the experiments of the preceding chapter to the prediction of long-term tumor control. First-order statistics obtain accuracy, sensitivity, specificity and AUC values of 0.67, 0.74, 0.60, and 1.00, respectively. Increasing the volume threshold improves these values to 0.83, 0.94, 0.64, and 0.88 for tumors larger than 5 cm^3. The application of Minkowski functionals does not significantly improve these results. The optimal support vector machine model obtains accuracy, sensitivity, specificity, and AUC values of 0.76, 0.83, 0.64, and 0.88, respectively, for tumors larger than 5 cm^3. The same holds for the models trained on gray-level size zone matrix features, obtaining 0.77, 0.86, 0.65, and 0.84 for the validation results, for tumors larger than 4 cm^3. However, GLCM-based models obtain accuracy, sensitivity, specificity, and AUC values of 0.77, 0.71, 0.83, and 0.93, respectively, including all VS tumors of various sizes. These results increase up to 0.83, 0.83, 0.82, and 0.99, respectively, for tumors larger than 5 cm^3. This shows that radiomics-based information can be potentially used in a clinical decision-support system, enabling the overall improvement of the treatment of VS tumors.

8.2 Discussion on the research questions

This section will evaluate the proposed methods and solutions addressing the research questions formulated in Section 1.6.

RQ1: Data and treatment response measurements

RQ1a: Which patients and how many need to be included for determining predictive parameters of the GKRS treatment outcome?

The discussions on the state of the art and the treatment response definitions in Chapter 2 have highlighted several aspects of the vestibular schwannoma. First, VS tumors are considered rare, so it is difficult to obtain large numbers of patients. Furthermore, since the Gamma Knife obtains high levels of long-term tumor control (see Section 1.3), treatment failure rates are low. Therefore, in order to evaluate parameters that may enable treatment outcome prediction, the required number of included patients increases even further. Since the ETZ in Tilburg is the only institution in the Netherlands that treats VS patients with a Gamma Knife, their experts treat a significant amount of patients each year. Therefore, their unique large database facilitates the careful evaluation of this specific brain tumor. Sec-

ond, a VS can be unilateral and sporadic, but it can be caused also by a genetic condition called neurofibromatosis Type 2 (NF2). The resulting NF2-based VS tumors have different biological properties compared to their sporadic counterparts. As such, this type of tumors should be considered different and is therefore excluded from our research work. Third, the response of a VS tumor to GKRS is slow. Therefore, a long follow-up is required to enable the assessment of long-term treatment outcome. Fourth, for transient tumor enlargement, it is required that the tumor volume is known within the first 12 months after treatment. Since the transient swelling can be subsided after 12 months, the tumor volume at around 6 months provides crucial information for determining TTE. As such, a 6-month follow-up scan is required for this analysis. These four crucial observations have led to the inclusion of 735 patients in this work, with an additional 22 patients for the TTE analysis.

Since the amount of data points required for applying deep learning methods in this field is not known, we have opted for employing conventional machine learning approaches. These methods allow for obtaining meaningful results with clearly lower patient counts than initially required for deep learning. This thesis has proven with the obtained results in Chapters 5, 6, and 7, that prediction of the Gamma Knife treatment for vestibular schwannomas is feasible with the obtained dataset with the amount of patients varying between 40 and 99. These results pave the way for future explorations of more advanced machine learning strategies.

RQ1b: What are good clinical metrics for determining the different treatment outcomes?

The literature review in Chapter 2 has highlighted that there is no clear consensus in the employed treatment outcome definitions. First, long-term tumor control (or treatment success) is not considered in clinical evaluations. There is no clear cutoff as to how long patients need to be free of tumor progression before treatment is considered successful. In current research, survival analyses are conducted, thereby circumventing the necessity of such a definition. However, to enable treatment outcome prediction, clear and objective definitions are required. Following careful considerations with several medical specialists, we have derived a definition for long-term tumor control (Section 2.4.2). Based on the available data, we have found that no treatment failures occur after 129 months following treatment. Therefore, we have opted for employing this time-period as follow-up end-point. Each patient that has reached this end-point is considered to have been treated successfully.

Second, the definitions of treatment failure are highly variable in current state of the art. Section 2.4.1 has demonstrated that most centers consider intervention as treatment failure. However, several issues with this have been discussed. Furthermore, different measurement strategies result in varying failure rates. Therefore, we have introduced a clear and objective treatment failure definition (Section 2.4.1). For this, we have employed volumetric measurements to improve the measurement accuracy and introduced a mathematical model to determine a so-called *volumetric failure*. This model is defined as two consecutive significant

increases in tumor volume, among three consecutive follow-up MRI sessions. A volume change is considered significant if the tumor has grown at least 10% in volume. Furthermore, only MRI sessions are employed that were obtained at least two years after GKRS.

Finally, definitions of transient tumor enlargement (TTE) are mainly based on tumor-size measurements, and there is no clear consensus on the amount of increase required before progression is considered. Furthermore, since a swelling can subside within the first 12 months, it may be missed if patients are only scanned annually. Therefore, we have proposed a clear and objective definition of TTE, where we have incorporated a criterion for the change in tumor size and a criterion on the required follow-up scans (Section 2.4.3). A tumor has presented with TTE if there was a significant increase of the tumor volume within the first 12 months after treatment, followed by a volumetric reduction to at least the tumor volume at treatment. For this, a follow-up scan is required at about 6 months after treatment.

RQ2: Influence of the pre-treatment growth rate on the Gamma Knife treatment response

RQ2a: Is the pre-treatment growth rate influencing the rate of volume reduction following treatment?

The experiments in Chapter 3 have shown that the pre-treatment growth rate, calculated by the volume doubling time (VDT), did not correlate significantly with the short-term volumetric changes. For the 2-year MRI session, the obtained p-value was 0.07, which may suggest a trend in the data. However, since we have included 225 patients in the analysis for this specific correlation, it can be reasonably well concluded that the impact is not present in our data. For the other MRI sessions, i.e. after 6 months, 1 year, and 3 years, the resulting p-values clearly indicate that there is no relation present in our data. These results remain negative when considering a different approach for calculating the post-treatment volume change, i.e. tumor halving times, instead of the relative volume changes.

RQ2b: How does the pre-treatment growth rate relate to the long-term tumor control?

The results on the volume doubling time (VDT) data with respect to the long-term tumor control in Chapter 3 have presented that there is a clear influence of the pre-treatment growth rate on the long-term tumor control. Various survival analysis techniques have obtained p-values well beyond the generally accepted cutoff value of 0.05. First, a Kaplan-Meier analysis comparing slow growing and fast growing tumors showed that slow growing tumors obtained 5- and 10-year tumor control rates of 97.3% and 86.0%, respectively, whereas fast growing tumors obtained 85.5% and 67.6%, respectively ($p < 0.01$). This is confirmed in a Kaplan-Meier analysis comparing slow growing, average growing, and fast growing tumors. The obtained 5- and 10-year tumor control rates were 98.8% and 91.4%, 90.6% and 70.7%, and 84.6% and 66.4%, respectively ($p < 0.01$). Finally, a Cox regression analysis has been conducted. The resulting risk model is able to

compute the risk of treatment failure for a specific volume doubling time. In this model, the risk of loss of tumor control for a tumor with a given VDT will decrease with a multiplication factor of 0.97 for a tumor for which the VDT is one month larger, i.e. a slower growing tumor. This means the following for our database. The average patient, i.e. having a VDT of 15 months, had a risk at loss of tumor control within the first 5 years following treatment of 8.4%, whereas a patient with a VDT of 48 months had a risk of 3.1% at loss of tumor control within that same time period. Conversely, a tumor with a VDT of 6 months, i.e. a fast growing tumor, had a risk at loss of tumor control in 11.0% of the cases within the first 5 years.

RQ2c: In what way does the adopted clinical methodology influence the obtained prediction model results?

Since the treatment failure definition employed in the experiments from Chapter 3 is based on tumor volumes, inter- and intra-observer variations may have influenced the calculated VDT and the number of failures in each sub-cohort in the conducted analyses. In Section 2.5, we have determined that for tumors smaller than 250 mm^3, the variability in contouring increases beyond the 10% cutoff value chosen as reasonable for this work, in agreement with clinicians. Therefore, for these smaller tumors, the relative volume errors are significant. By introducing a threshold on the minimum required tumor volume (i.e. 250 mm^3), we have certainly reduced the impact of the relative volume errors. This statistically improved the results from the Cox regression (changes from $p < 0.05$ to $p = 0.01$). Furthermore, by excluding the volumetric failure, thus only considering intervention as failure, we still determined a significant difference between the slow- and fast-growing tumor cohorts ($p = 0.02$).

RQ3: Influence of the treatment planning on the treatment outcome

RQ3a: Does the marginal dose influence the long-term tumor control?

The conducted experiments in Chapter 4, Section 4.2, have presented no clear statistical correlation between globally calculated treatment parameters and the long-term tumor control. First, in univariate Cox regression analyses, the p-value for the dose covering 99% of the tumor volume (DOSE99) was slightly higher than the required 0.05 for obtaining statistical significance. Nevertheless, in a multivariate Cox regression, both tumor volume at treatment and the DOSE99 setting are determined to show significant covariation with respect to the long-term tumor control. This suggests the influence of both factors on the long-term tumor control. However, Kaplan-Meier analyses revealed that, after splitting the cohort based on the median of DOSE99, the resulting tumor control rates did not differ significantly. Dividing the complete cohort in three sub-groups did not improve these results. These analyses suggest that, even though the Cox regression determined that DOSE99 showed a significant covariation, the differences between tumor control rates within the protocolized doses are small and statistically not significant. The small variation possibilities within the protocolized treatments are guided by the minimization of the prescribed radiation dose over the last decades, where the

treatment protocols have been optimized to reduce the radiation toxicity, while maintaining good tumor control rates.

RQ3b: Is there an influence of the specific heterogeneous dose distribution on the long-term treatment outcome?

The analyses in Chapter 4, Section 4.3, have presented that the specific heterogeneous dose distributions may have an influence on the treatment response. In a limited dataset of 20 tumors in which the treatment resulted in a failure and 20 tumors that obtained significant volume reduction within the first year following treatment (fast responders), the differences in calculated homogeneity indices (HIs) were found statistically not significant. Indeed, in a machine learning environment, the resulting HI values have obtained classification accuracy values of about 50%. This did not yield improvement over random classification. However, these indices do not incorporate actual spatial dose information. To this end, we have introduced a novel metric for assessing the heterogeneous dose distribution by incorporating the spatial information. The three-dimensional histograms of oriented gradients have been calculated for each of the 40 included tumors. The resulting feature vectors have been supplied to a well-known machine learning algorithm to evaluate their impact on the treatment response. The highest accuracy value obtained was 77.5%, thereby suggesting that there are measurable differences that could potentially influence the treatment efficacy, even though treatment plannings are considered uniform.

RQ3c: How does the inter-observer tumor segmentation variability work out on the Gamma Knife treatment response?

It is observed in Chapter 4, Section 4.4, that the delineations of the fast-responding group show more variations compared to the retrospectively created annotations. Tumor volume differences are statistically not significant, although it may be considered that a trend is present in the data since the resulting p-value was 0.08. By introducing a novel metric that normalizes the differences in slice area with respect to the height location within the tumor, we have shown that both the summed absolute pixel differences, as well as the summed relative area differences differ significantly between the two included cohorts ($p = 0.02$). Finally, calculated Jaccard indices have highlighted that the differences within the fast-responding cohort are significantly larger than within the failure cohort. With these results, it can be deduced that there are significant inter-observer variation differences between the included cohorts. Furthermore, it is shown that these variations are larger within the fast-responding cohort, compared to the cohort containing patients for which the treatment failed. This suggests that tumors that are more difficult to segment, are expected to obtain a positive treatment outcome.

RQ4: Selection and application of informative MRI-based quantitative features for predicting the GKRS treatment response

RQ4a: Can quantitative tumor shape descriptors enable the prediction of the GKRS treatment response?

In Chapter 5, the examination of the impact of shape descriptors on the treatment response has revealed that shape is a weak predictor of the treatment outcome, in the limited dataset of 20 fast-responding tumors and 20 failures. In the conducted assessments, 19 two-dimensional and 6 three-dimensional shape descriptors have been considered. First, we have evaluated whether each calculated shape feature differed significantly between both cohorts. Visual inspection of scatter plots, histograms, and boxplots have revealed that inter-class variations are small, while intra-class variations are large. Only the roughness index provided a clear difference between both classes. Statistical analyses further showed that perimeter, area, and four contour-sequence-moments features differed significantly among the two cohorts. However, none of the three-dimensional features reached statistical significance. Implementing the resulting features in a machine learning environment obtained accuracy values of around 65.0%. The best model achieved accuracy, sensitivity, specificity, and area under the receiver operating characteristic of 67.5%, 65.0%, 70.0%, and 0.70, respectively. Although this is an improvement over random classification, the obtained results highlight that shape can only be considered as weak predictor for the treatment response.

RQ4b: Which texture features are informative for the various treatment responses?

The evaluation of various texture features in Chapter 5 have exhibited that popular second-order statistical metrics like GLCM and RLM, are suitable for describing texture and predicting the Gamma Knife treatment response. These metrics obtained accuracy values of 77.5% and 82.5%, respectively, using a broadly accepted machine learning algorithm. The corresponding sensitivity and specificity values were 90.0% and 75.0%, respectively, for the GLCM-based model, and 75.0% and 80.0%, respectively, for the RLM-based model. Nevertheless, these metrics are slightly outperformed by simple first-order statistics (FOS), like mean, standard deviation, and median. The best-performing FOS-based model obtained accuracy, sensitivity, and specificity values of 85.0%.

RQ4c: What is the influence of the imbalance in data and the variations in tumor volumes on the prediction results?

The conducted experiments in Chapter 6 have displayed that the imbalance in the data can have an impact on the obtained prediction results. In the experiments on FOS features and on Minkowski features, the inclusion of all data resulted in an increased sensitivity, but a decreased specificity. This is most likely caused by the fact that the imbalance in the data can result in trained models that have a large preference for the majority class. As such, there are very few true negatives, resulting in a low specificity. However, for GLCM-trained models, both sensitivity and

specificity values are increased. This suggests that the additional data improved the trained model, thereby creating a more detailed decision boundary.

Chapters 6 and 7 have investigated the influence of the various tumor volumes on the obtained prediction models. Both chapters have concluded that this influence is present for the GLCM-based models. For FOS-based, Minkowski-based, and GLSZM-based models, increasing the volume threshold resulted in comparable accuracy, sensitivity, and specificity values, whereas the AUC values decreased. This is most likely caused by the reduced number of tumors included in the analyses. However, for GLCM-based models, all considered validation values increased with the volume threshold. This suggests that the models obtained with GLCM features are robust.

RQ4d: Is it possible to develop models that can predict transient tumor enlargement and the long-term treatment success, based on MRI texture features?

The experiments in Chapters 6 and 7 have achieved various results, where most of the included features showed to have predictive value of the treatment outcomes. However, GLCM-based models have produced the best results. For TTE prediction, the best model realized sensitivity and specificity values of 0.82 and 0.69, respectively. For a minimum volume inclusion criterion of 6 cm^3, the highest sensitivity and specificity values of 0.77 and 0.89 have been obtained, respectively. The resulting models achieved AUC values in the range 0.90 – 0.95, with confidence intervals of approximately 0.80 up to 0.99.

For long-term tumor control, the same conclusions can be drawn. Features from FOS, Minkowski, and GLSZM have realized interesting results, thereby highlighting their predictive value. However, GLCM-based models again outperform the other feature extractors, obtaining accuracy, sensitivity, specificity, and AUC values of 0.77, 0.71, 0.83, and 0.93, respectively. If additional volume thresholding is applied, these results increase to 0.83, 0.83, 0.82, and 0.99, respectively, for tumors larger than 5 cm^3. Both results exhibit the possibility of creating a model that enables the prediction of TTE and of long-term tumor control on an individual patient-basis.

8.3 Future outlook on Gamma Knife treatment prediction

Based on the continuous improvement in machine learning, the application of advanced imaging techniques, such as perfusion and diffusion MRI and/or multi-spectral sensing, may provide additional and valuable data for improving the prediction models. Combining multi-modal imaging sequences increases the amount of information obtained, since each imaging sequence highlights different biological aspects of the tumor tissue. Therefore, the combination could prove to be crucial in creating models that provide highly accurate treatment outcome predictions.

In the case of vestibular schwannomas, several innovation developments are already happening. Various research groups are investigating the possibility to

predict the treatment response, and these are increasingly based on individual patient-specific information. However, since VS is a rare pathology, it is necessary to undertake large multi-center investigations. This will lead to an increased number of included patients and it will make the data more heterogeneous. Each medical center has its own strategy and methods for treating a patient. Each treatment planning has many parameters and is therefore highly heterogeneous and possibly dependent on which center is creating and executing the treatments. Therefore, collaboration is of paramount importance and will lead to an improvement of the prediction models, as proposed in this thesis. Enabling accurate prediction of the treatment results will in the end further improve overall treatment outcomes for vestibular schwannomas.

Furthermore, in this work, we have shown that the pre-treatment growth rate can predict the risk at loss of tumor control. This is an easy model to implement in a clinical workflow. However, there are studies suggesting that up-front active treatment leads to improved cranial nerve functions, when compared to a wait-and-see approach. Others advocate against active treatment, since it may not be necessary to treat a tumor that is not growing. It would be therefore highly interesting to investigate the possibility of predicting the natural growth rate of a tumor. If possible, a wait-and-scan policy may become obsolete, and the predicted tumor growth rate can be used for determining the risk at loss of tumor control, thereby enabling a clinical decision-support system.

The ever-increasing abilities of machine learning methods enable the further enhancement of the already developed models. In the past years, deep learning techniques have clearly proven to outperform the conventional machine learning models in many fields of research, including medicine. These developments are so strong, that their results surpass the expert assessments already in well-defined cases. This suggests that these techniques will most likely become important tools not only for medical image analysis, but also for medical experts and general physicians. However, the explainability of these advanced AI techniques is currently hampering their implementations in the clinical workflow. It is therefore expected that this field will further grow in the coming years, so that the acceptance for using AI techniques is further broadened. The technology will develop so rapidly and broadly that the solutions will be very advanced and sophisticated, thereby making AI an medical expert field of its own. This implies that collaborations should expand beyond the border of medical experts and physicians and should involve technical AI experts as well.

In the last decade, the technology and the involved applications of AI have been increasing rapidly. This trend is also visible in medicine, where the data is abundantly available on multiple diseases, conditions, and treatments. Machine learning and the availability of so-called *big data* enable physicians and researchers to develop models that lead to significant improvements of the various care paths available. These models can detect for instance diseases earlier on, improve disease classifications, and enable the risk assessments of the actions taken in the

clinical care path by predicting the results. Furthermore, the inclusion of metadata, multi-modal data processing, and multi-dimensional decision-making are actual developments taking place in clinical research. Currently, several clinical applications have been and are proposed, including those from large technology companies like Google, Apple and IBM. In the next decade, the number of AI applications in medicine will increase even further, due to the awareness of possibilities and increased exposure of the developed applications and their results.

In summary, while this thesis has presented the first promising results in predicting the various Gamma Knife treatment responses of vestibular schwannomas, the obtained models need to be validated on multiple external datasets. Furthermore, the availability of additional data from these external datasets can improve the prediction models, since more complex machine learning methods can be explored. Moreover, the results in this thesis are based on individual features, calculating first- and second-order statistics on the gray-level MRI intensities. Combining the various features on multi-modal MRI sequences may further improve the results. Also the implementation of advanced imaging protocols may provide valuable information on the intrinsic tumor biology, thereby possibly improving the predictability of the various treatment outcomes. Additionally, since we have concluded that GLCM-based patterns are informative for Gamma Knife treatment responses, frequency-based features may improve upon the existing features. Nevertheless, the obtained results highlight the possibilities of predicting the various treatment outcomes, thereby enabling the improvement of the overall treatment results. Finally, the ability to predict the treatment outcome on an individual basis will significantly aid patients and their treating physicians in selecting the optimal treatment strategy and follow-up care path.

Bibliography

[1] Q. T. Ostrom, H. Gittleman, P. Liao, T. Vecchione-Koval, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan. "CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014". In: *Neuro-Oncology* 19.suppl_5 (2017), pp. v1–v88. ISSN: 1522-8517.

[2] M. Kleijwegt, V. Ho, O. Visser, W. Godefroy, and A. van der Mey. "Real Incidence of Vestibular Schwannoma? Estimations From a National Registry". In: *Otology & Neurotology* 37.9 (Oct. 2016), pp. 1411–1417. ISSN: 1531-7129.

[3] Blausen.com staff. "Medical gallery of Blausen Medical 2014". In: *WikiJournal of Medicine* 1.2 (2014), p. 10.

[4] M. Hentschel, M. Rovers, L. Markodimitraki, S. Steens, and H. Kunst. "An international comparison of diagnostic and management strategies for vestibular schwannoma". In: *European Archives of Oto-Rhino-Laryngology* 276.1 (2019), pp. 71–78. ISSN: 1434-4726.

[5] E. P. Lin and B. T. Crane. "The Management and Imaging of Vestibular Schwannomas." In: *AJNR. American journal of neuroradiology* 38.11 (Nov. 2017), pp. 2034–2043. ISSN: 1936-959X.

[6] M. L. Carlson, E. B. Habermann, A. E. Wagie, C. L. Driscoll, J. J. Van Gompel, J. T. Jacob, and M. J. Link. "The Changing Landscape of Vestibular Schwannoma Management in the United States-A Shift Toward Conservatism". In: *Otolaryngology-Head and Neck Surgery* 153.3 (2015), pp. 440–446.

[7] J. Zou and T. Hirvonen. ""Wait and scan" management of patients with vestibular schwannoma and the relevance of non-contrast MRI in the follow-up". In: *Journal of Otology* 12.4 (Dec. 2017), pp. 174–184. ISSN: 1672-2930.

[8] J. J. Van Gompel, J. Patel, C. Danner, A. N. Zhang, A. A. Youssef, H. R. Van Loveren, and S. Agazzi. "Acoustic neuroma observation associated with an increase in symptomatic tinnitus: Results of the 2007-2008 Acoustic Neuroma Association survey". In: *Journal of Neurosurgery* 119.4 (Oct. 2013), pp. 864–868. ISSN: 0022-3085.

[9] J. Régis, R. Carron, M. C. Park, O. Soumare, C. Delsanti, J. M. Thomassin, and P.-H. Roche. "Wait-and-see strategy compared with proactive Gamma Knife surgery in patients with intracanalicular vestibular schwannomas." In: *Journal of neurosurgery* 113.Special_Supplement (2010), pp. 105–111. ISSN: 1933-0693.

[10] J. G. Wolbers, A. H. Dallenga, A. Mendez Romero, and A. van Linge. "What intervention is best practice for vestibular schwannomas? A systematic review of controlled studies". In: *BMJ Open* 3.2 (2013), pp. 1–10. ISSN: 2044-6055.

[11] D. E. Anderson, J. Leonetti, J. J. Wind, D. Cribari, and K. Fahey. "Resection of large vestibular schwannomas: Facial nerve preservation in the context of surgical approach and patient-assessed outcome". In: *Journal of Neurosurgery* 102.4 (Apr. 2005), pp. 643–649. ISSN: 0022-3085.

[12] S. Jung, S. S. Kang, T. S. Kim, H. J. Kim, S. K. Jeong, Seok-Chul, J. K. Lee, J. H. Kim, S. H. Kim, and J. H. Lee. "Current surgical results of retrosigmoid approach in extralarge vestibular schwannomas". In: *Surgical Neurology* 53.4 (Apr. 2000), pp. 370–378. ISSN: 0090-3019.

[13] M. L. Carlson, Ø. V. Tveiten, C. L. Driscoll, et al. "Long-term quality of life in patients with vestibular schwannoma: an international multicenter cross-sectional study comparing microsurgery, stereotactic radiosurgery, observation, and nontumor controls." In: *Journal of neurosurgery* 122.4 (Apr. 2015), pp. 833–842. ISSN: 1933-0693.

[14] H. Abou-Al-Shaar, M. A. Azab, M. Karsy, J. Guan, G. Alzhrani, Y. M. Gozal, R. L. Jensen, and W. T. Couldwell. "Assessment of costs in open surgery and stereotactic radiosurgery for vestibular schwannomas". In: *Journal of Neurosurgery* 131.2 (Oct. 2018), pp. 561–568. ISSN: 0022-3085.

[15] Z. Schnurman, J. G. Golfinos, D. Epstein, D. R. Friedmann, J. T. Roland, and D. Kondziolka. "Comparing costs of microsurgical resection and stereotactic radiosurgery for vestibular schwannoma". In: *Journal of Neurosurgery* 131.5 (Nov. 2018), pp. 1395–1404. ISSN: 0022-3085.

[16] D. Starnoni, R. T. Daniel, C. Tuleasca, M. George, M. Levivier, and M. Messerer. "Systematic review and meta-analysis of the technique of subtotal resection and stereotactic radiosurgery for large vestibular schwannomas: A "nerve-centered" approach". In: *Neurosurgical Focus* 44.3 (Mar. 2018), E4. ISSN: 1092-0684.

[17] W. Y. Chung, D. H. C. Pan, C. C. Lee, H. M. Wu, K. D. Liu, Y. S. Yen, W. Y. Guo, C. Y. Shiau, and Y. H. Shih. "Large vestibular schwannomas treated by Gamma Knife surgery: long-term outcomes." In: *Journal of neurosurgery* 113.Special_Supplement (Dec. 2010), pp. 112–121. ISSN: 1933-0693.

[18] C. W. Huang, H. T. Tu, C. Y. Chuang, C. S. Chang, H. H. Chou, M. T. Lee, and C. F. Huang. "Gamma Knife radiosurgery for large vestibular schwannomas greater than 3 cm in diameter". In: *Journal of Neurosurgery* 128.5 (May 2018), pp. 1380–1387. ISSN: 1933-0693.

[19] C. Iorio-Morin, F. Alsubaie, and D. Mathieu. "Safety and efficacy of gamma knife radiosurgery for the management of Koos Grade 4 vestibular schwannomas". In: *Neurosurgery* 78.4 (Apr. 2016), pp. 521–530. ISSN: 1524-4040.

[20] M. Lefranc, L. M. Da Roz, A. Balossier, J. M. Thomassin, P. H. Roche, and J. Regis. "Place of Gamma Knife Stereotactic Radiosurgery in Grade 4 Vestibular Schwannoma Based on Case Series of 86 Patients with Long-Term Follow-Up". In: *World Neurosurgery* 114 (June 2018), e1192–e1198. ISSN: 1878-8769.

[21] B. D. Milligan, B. E. Pollock, R. L. Foote, and M. J. Link. "Long-term tumor control and cranial nerve outcomes following Gamma knife surgery for larger-volume vestibular schwannomas: Clinical article". In: *Journal of Neurosurgery* 116.3 (Mar. 2012), pp. 598–604. ISSN: 0022-3085.

[22] M. Bailo, N. Boari, A. Franzin, F. Gagliardi, A. Spina, A. del Vecchio, M. Gemma, A. Bolognesi, and P. Mortini. "Gamma Knife Radiosurgery as Primary Treatment for Large Vestibular Schwannomas: Clinical Results at Long-Term Follow-Up in a Series of 59 Patients". In: *World Neurosurgery* 95 (Nov. 2016), pp. 487–501. ISSN: 1878-8769.

[23] R. van de Langenberg, P. E. J. Hanssens, J. J. van Overbeeke, J. B. Verheul, P. J. Nelemans, B. J. de Bondt, and R. J. Stokroos. "Management of large vestibular schwannoma. Part I. Planned subtotal resection followed by Gamma Knife surgery: radiological and clinical aspects". In: *Journal of Neurosurgery* 115.5 (Nov. 2011), pp. 875–884. ISSN: 0022-3085.

[24] F. A. Zeiler, M. Bigder, A. Kaufmann, P. J. McDonald, D. Fewer, J. Butler, G. Schroeder, and M. West. "Gamma knife radiosurgery for large vestibular schwannomas: a Canadian experience". In: *The Canadian journal of neurological sciences. (Le journal canadien des sciences neurologiques)* 40.3 (May 2013), pp. 342–347. ISSN: 0317-1671.

[25] J. J. Olson, S. N. Kalkanis, and T. C. Ryken. "Congress of Neurological Surgeons Systematic Review and Evidence-Based Guidelines on the Treatment of Adults with Vestibular Schwannomas: Executive Summary". In: *Clinical Neurosurgery*. Vol. 82. 2. Narnia, Feb. 2018, pp. 129–134.

[26] M. L. Carlson, A. E. Glasgow, B. R. Grossardt, E. B. Habermann, and M. J. Link. "Does where you live influence how your vestibular schwannoma is managed? Examining geographical differences in vestibular schwannoma treatment across the United States". In: *Journal of Neuro-Oncology* 129.2 (Sept. 2016), pp. 269–279. ISSN: 1573-7373.

[27] B. E. Pollock, M. J. Link, S. L. Stafford, I. F. Parney, Y. I. Garces, and R. L. Foote. "The Risk of Radiation-Induced Tumors or Malignant Transformation After Single-Fraction Intracranial Radiosurgery: Results Based on a 25-Year Experience". In: *International Journal of Radiation Oncology Biology Physics* 97.5 (Apr. 2017), pp. 919–923. ISSN: 1879-355X.

[28] L. Ma, A. Nichol, S. Hossain, et al. "Variable dose interplay effects across radiosurgical apparatus in treating multiple brain metastases". In: *International Journal of Computer Assisted Radiology and Surgery* 9.6 (2014), pp. 1079–1086. ISSN: 1861-6429.

[29] S. Braunstein and L. Ma. "Stereotactic radiosurgery for vestibular schwannomas". In: *Cancer Management and Research* 10 (2018), pp. 3733–3740. ISSN: 1179-1322.

[30] T. Hasegawa, Y. Kida, T. Kato, H. Iizuka, and T. Yamamoto. "Factors associated with hearing preservation after Gamma Knife surgery for vestibular schwannomas in patients who retain serviceable hearing: Clinical article". In: *Journal of Neurosurgery* 115.6 (Dec. 2011), pp. 1078–1086. ISSN: 0022-3085.

[31] N. Boari, M. Bailo, F. Gagliardi, A. Franzin, M. Gemma, A. del Vecchio, A. Bolognesi, P. Picozzi, and P. Mortini. "Gamma Knife radiosurgery for vestibular schwannoma: clinical results at long-term follow-up in a series of 379 patients". In: *Journal of neurosurgery* 121.Suppl_2 (2014), pp. 123–142. ISSN: 1933-0693.

[32] T. Hasegawa, Y. Kida, T. Kato, H. Iizuka, S. Kuramitsu, and T. Yamamoto. "Long-term safety and efficacy of stereotactic radiosurgery for vestibular schwannomas: Evaluation of 440 patients more than 10 years after treatment with Gamma Knife surgery". In: *Journal of Neurosurgery* 118.3 (Mar. 2013), pp. 557–565. ISSN: 0022-3085.

[33] O. W. Meijer, E. J. Weijmans, D. L. Knol, B. J. Slotman, F Barkhof, W. P. Vandertop, and J. A. Castelijns. "Tumor-volume changes after radiosurgery for vestibular schwannoma: Implications for follow-up MR imaging protocol". In: *American Journal of Neuroradiology*. Vol. 29. 5. 2008, pp. 906–910.

[34] R. A. Friedman, D. E. Brackmann, W. E. Hitselberger, M. S. Schwartz, Z. Iqbal, and K. I. Berliner. "Surgical salvage after failed irradiation for vestibular schwannoma". In: *Laryngoscope* 115.10 (Oct. 2005), pp. 1827–1832. ISSN: 0023-852X.

[35] S. T. Husseini, E. Piccirillo, A. Taibah, T. Almutair, G. Sequino, and M. Sanna. "Salvage surgery of vestibular schwannoma after failed radiotherapy: The Gruppo Otologico experience and review of the literature". In: *American Journal of Otolaryngology - Head and Neck Medicine and Surgery* 34.2 (Mar. 2013), pp. 107–114. ISSN: 0196-0709.

[36] Y. Iwai, K. Ishibashi, Y. Nakanishi, Y. Onishi, S. Nishijima, and K. Yamanaka. "Functional Outcomes of Salvage Surgery for Vestibular Schwannomas after Failed Gamma Knife Radiosurgery". In: *World Neurosurgery* 90 (June 2016), pp. 385–390. ISSN: 1878-8769.

[37] S. C. Wise, M. L. Carlson, Ø. V. Tveiten, C. L. Driscoll, E. Myrseth, M. Lund-Johansen, and M. J. Link. "Surgical salvage of recurrent vestibular schwannoma following prior stereotactic radiosurgery". In: *Laryngoscope* 126.11 (Nov. 2016), pp. 2580–2586. ISSN: 1531-4995.

[38] H. J. Lee, M. J. Kim, S. H. Koh, W. S. Chang, and I. S. Moon. "Comparing Outcomes Following Salvage Microsurgery in Vestibular Schwannoma Patients Failing Gamma-knife Radiosurgery or Microsurgery". In: *Otology and Neurotology* 38.9 (Oct. 2017), pp. 1339–1344. ISSN: 1537-4505.

[39] T. Hasegawa, Y. Kida, T. Kobayashi, M. Yoshimoto, Y. Mori, and J. Yoshida. "Long-term outcomes in patients with vestibular schwannomas treated using gamma knife surgery: 10-year follow up." In: *Journal of neurosurgery* 102.1 (Jan. 2005), pp. 10–16. ISSN: 1933-0693.

[40] B. J. Arthurs, W. T. Lamoreaux, A. R. MacKay, J. J. Demakas, N. A. Giddings, R. K. Fairbanks, B. S. Cooke, A. L. Elaimy, B. Peressini, and C. M. Lee. "Gamma knife radiosurgery for vestibular schwannomas: Tumor control and functional preservation in 70 patients". In: *American Journal of Clinical Oncology: Cancer Clinical Trials* 34.3 (June 2011), pp. 265–269. ISSN: 0277-3732.

[41] F. C. A. Timmer, J. J. S. Mulder, P. E. J. Hanssens, J. J. Van Overbeeke, R. T. Donders, C. W. R. J. Cremers, and K. Graamans. "Gamma knife radiosurgery for vestibular schwannomas: Identification of predictors for continued tumor growth and the influence of documented tumor growth preceding radiation treatment". In: *Laryngoscope* 121.9 (2011), pp. 1834–1838. ISSN: 0023-852X.

[42] J. K. Varughese, T. Wentzel-Larsen, P. H. Pedersen, R. Mahesparan, and M. Lund-Johansen. "Gamma knife treatment of growing vestibular schwannoma in Norway: A prospective study". In: *International Journal of Radiation Oncology Biology Physics* 84.2 (Oct. 2012), e161–e166. ISSN: 0360-3016.

[43] B. J. Williams, Z. Xu, D. J. Salvetti, I. T. McNeill, J. Larner, and J. P. Sheehan. "Gamma Knife surgery for large vestibular schwannomas: A single-center retrospective case-matched comparison assessing the effect of lesion size". In: *Journal of Neurosurgery* 119.2 (Aug. 2013), pp. 463–471. ISSN: 0022-3085.

[44] S. Larjani, E. Monsalves, H. Pebdani, B. Krischek, F. Gentili, M. Cusimano, N. Laperriere, C. Hayhurst, and G. Zadeh. "Identifying predictors of early growth response and adverse radiation effects of vestibular schwannomas to radiosurgery". In: *PLoS ONE* 9.10 (2014), e110823. ISSN: 1932-6203.

[45] T. Wangerid, J. Bartek, M. Svensson, and P. Förander. "Long-term quality of life and tumour control following gamma knife radiosurgery for vestibular schwannoma". In: *Acta Neurochirurgica* 156.2 (2014), pp. 389–396. ISSN: 0001-6268.

[46] C. D. Frisch, J. T. Jacob, M. L. Carlson, R. L. Foote, C. L. Driscoll, B. A. Neff, B. E. Pollock, and M. J. Link. "Stereotactic Radiosurgery for Cystic Vestibular Schwannomas". In: *Neurosurgery* 80.1 (Aug. 2016), pp. 112–118. ISSN: 1524-4040.

[47] S. Klijn, J. B. Verheul, G. N. Beute, S. Leenstra, J. J. S. Mulder, H. P. M. Kunst, and P. E. J. Hanssens. "Gamma Knife radiosurgery for vestibular schwannomas: evaluation of tumor control and its predictors in a large patient cohort in The Netherlands." In: *Journal of Neurosurgery* 124.6 (2016), pp. 1619–26. ISSN: 1933-0693.

[48] J. M. Lee, D. H. Kwon, C. J. Kim, and J. H. Kim. "Treatment Outcome of Gamma Knife Radiosurgery of Vestibular Schwannomas with Cystic Component". In: *The Nerve* 2.1 (Apr. 2016), pp. 1–4. ISSN: 2465-891X.

[49] A. P. Marston, J. T. Jacob, M. L. Carlson, B. E. Pollock, C. L. W. Driscoll, and M. J. Link. "Pretreatment growth rate as a predictor of tumor control following Gamma Knife radiosurgery for sporadic vestibular schwannoma". In: *Journal of Neurosurgery* 127.2 (2017), pp. 380–387.

[50] G. Bowden, J. Cavaleri, E. M. III, A. Niranjan, J. Flickinger, and D. D. Lunsford. "Cystic Vestibular Schwannomas Respond Best to Radiosurgery". In: *Neurosurgery* 81.3 (Sept. 2017), pp. 490–497. ISSN: 1524-4040.

[51] A. Camargo, T. Schneider, L. Liu, J. Pakpoor, L. Kleinberg, and D. M. Yousem. "Pretreatment ADC values predict response to radiosurgery in vestibular schwannomas". In: *American Journal of Neuroradiology* 38.6 (2017), pp. 1200–1205. ISSN: 1936-959X.

[52] J. H. Kim, H. H. Jung, J. H. Chang, J. W. Chang, Y. G. Park, and W. S. Chang. "Predictive Factors of Unfavorable Events After Gamma Knife Radiosurgery for Vestibular Schwannoma". In: *World Neurosurgery* 107 (2017), pp. 175–184. ISSN: 1878-8769.

[53] C.-C. Wu, W.-Y. Guo, W.-Y. Chung, H.-M. Wu, C.-J. Lin, C.-C. Lee, K.-D. Liu, and H.-c. Yang. "Magnetic resonance imaging characteristics and the prediction of outcome of vestibular schwannomas following Gamma Knife radiosurgery". In: *Journal of Neurosurgery* 127.6 (2017), pp. 1384–1391.

[54] H. Borghei-Razavi, M. Poturalski, C. Karakasis, J. Bullen, P. Recinos, J. Lee, and V. Kshettry. "Pretreatment ADC Values to Predict Response of Vestibular Schwannoma to Gamma Knife Radiosurgery". In: *29th Annual Meeting North American Skull Base Society*. Vol. 80. S 01. Georg Thieme Verlag KG, Feb. 2019, A219.

[55] J. Chang, J. D. Breshears, A. M. Molinaro, P. K. Sneed, M. W. McDermott, P. V. Theodosopoulos, and A. D. Tward. "Impact of pretreatment growth on Tumor control for vestibular schwannomas following gamma knife". In: *Laryngoscope* 129.3 (Mar. 2019), pp. 743–747. ISSN: 1531-4995.

[56] J. M. Frischer, E. Gruber, V. Schöffmann, et al. "Long-term outcome after Gamma Knife radiosurgery for acoustic neuroma of all Koos grades: A single-center study". In: *Journal of Neurosurgery* 130.2 (Feb. 2019), pp. 388–397. ISSN: 1933-0693.

[57] S. H. Lim, C. K. Park, B. J. Park, and Y. J. Lim. "Long-Term Outcomes of Gamma Knife Radiosurgery for Cystic Vestibular Schwannomas". In: *World Neurosurgery* 132 (Dec. 2019), e34–e39. ISSN: 1878-8769.

[58] D. R. Smith, H. J. Saadatmand, C. C. Wu, P. J. Black, Y. R. Wuu, J. Lesser, M. Horan, S. R. Isaacson, T. J. Wang, and M. B. Sisti. "Treatment Outcomes and Dose Rate Effects Following Gamma Knife Stereotactic Radiosurgery for Vestibular Schwannomas". In: *Clinical Neurosurgery* 85.6 (2019), E1084–E1094. ISSN: 1524-4040.

[59] H. Speckter, J. Santana, J. Bido, G. Hernandez, D. Rivera, L. Suazo, S. Valenzuela, J. Oviedo, C. F. Gonzalez, and P. Stoeter. "Texture Analysis of Standard Magnetic Resonance Images to Predict Response to Gamma Knife Radiosurgery in Vestibular Schwannomas". In: *World Neurosurgery* 132 (Sept. 2019), e228–e234. ISSN: 1878-8769.

[60] I. M. Germano, J. Sheehan, J. Parish, T. Atkins, A. Asher, C. G. Hadjipanayis, S. H. Burri, S. Green, and J. J. Olson. "Congress of Neurological Surgeons Systematic Review and Evidence-Based Guidelines on the Role of Radiosurgery and Radiation Therapy in the Management of Patients with Vestibular Schwannomas". In: *Clinical Neurosurgery* 82.2 (Feb. 2018), E49–E51. ISSN: 0148-396X.

[61] R. Van De Langenberg, P. E. J. Hanssens, J. B. Verheul, J. J. Van Overbeeke, P. J. Nelemans, A. J. C. Dohmen, B. J. De Bondt, and R. J. Stokroos. "Management of large vestibular schwannoma. Part II. Primary Gamma Knife surgery: Radiological and clinical aspects - Clinical article". In: *Journal of Neurosurgery* 115.5 (Nov. 2011), pp. 885–893. ISSN: 0022-3085.

[62] T. E. Yankeelov, N. Atuegwu, D. Hormuth, J. A. Weis, S. L. Barnes, M. I. Miga, E. C. Rericha, and V. Quaranta. "Clinically relevant modeling of tumor growth and treatment response". In: *Science Translational Medicine* 5.187 (May 2013), 187ps9. ISSN: 1946-6234.

[63] S. Charabi, M. Mantoni, M. Tos, and J. Thomsen. "Cystic vestibular schwannomas: Neuroimaging and growth rate". In: *The Journal of Laryngology & Otology* 108.5 (May 1994), pp. 375–379. ISSN: 1748-5460.

[64] E. Spickler, R. Lufkin, L. Teresi, L. Chiu, U. Batzdorf, R. Rand, and D. Becker. "MR of hemorrhagic acoustic neuromas." In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 15.5 (1991), pp. 333–337. ISSN: 0895-6111.

[65] A. Gomez-Brouchet, M. B. Delisle, C. Cognard, A. Bonafe, J. P. Charlet, O. Deguine, and B. Fraysse. "Vestibular schwannomas: correlations between magnetic resonance imaging and histopathologic appearance." In: *Otology & neurotology* 22.1 (Jan. 2001), pp. 79–86. ISSN: 1531-7129.

[66] M. de Vries, P. C. W. Hogendoorn, I. Briaire-de Bruyn, M. J. A. Malessy, and A. G. L. van der Mey. "Intratumoral hemorrhage, vessel density, and the inflammatory reaction contribute to volume increase of sporadic vestibular schwannomas". In: *Virchows Archiv* 460.6 (June 2012), pp. 629–636. ISSN: 0945-6317.

[67] M. E. Linskey, L. Dade Lunsford, and J. C. Flickinger. "Radiosurgery for acoustic neurinomas: Early experience". In: *Neurosurgery* 26.5 (May 1990), pp. 736–744. ISSN: 0148-396X.

[68] T. F. Witham, H. Okada, W. Fellows, R. L. Hamilton, J. C. Flickinger, W. H. Chambers, I. F. Pollack, S. C. Watkins, and D. Kondziolka. "The characterization of tumor apoptosis after experimental radiosurgery". In: *Stereotactic and Functional Neurosurgery* 83.1 (2005), pp. 17–24. ISSN: 1011-6125.

[69] R. J. Gillies, P. E. Kinahan, and H. Hricak. "Radiomics: Images are more than pictures, they are data". In: *Radiology* 278.2 (Feb. 2016), pp. 563–577. ISSN: 1527-1315.

[70] R. Forghani, P. Savadjiev, A. Chatterjee, N. Muthukrishnan, C. Reinhold, and B. Forghani. "Radiomics and Artificial Intelligence for Biomarker and Prediction Model Development in Oncology". In: *Computational and Structural Biotechnology Journal* 17 (Jan. 2019), pp. 995–1008. ISSN: 2001-0370.

[71] M. Avanzo, J. Stancanello, and I. El Naqa. "Beyond imaging: The promise of radiomics". In: *Physica Medica* 38 (June 2017), pp. 122–139. ISSN: 1724-191X.

[72] K. Doi. "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential". In: *Computerized Medical Imaging and Graphics* 31.4-5 (June 2007), pp. 198–211. ISSN: 0895-6111.

[73] M. L. Giger. "Machine Learning in Medical Imaging". In: *Journal of the American College of Radiology* 15.3 (Mar. 2018), pp. 512–520. ISSN: 1558-349X.

[74] H. Soltanian-Zadeh, F. Rafiee-Rad, and D. Siamak Pourabdollah-Nejad. "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms". In: *Pattern Recognition* 37.10 (Oct. 2004), pp. 1973–1986. ISSN: 0031-3203.

[75] A. V. Alvarenga, W. C. Pereira, A. F. C. Infantosi, and C. M. De Azevedo. "Classification of breast tumours on ultrasound images using morphometric parameters". In: *2005 IEEE International Workshop on Intelligent Signal Processing - Proceedings*. 2005, pp. 206–210. ISBN: 078039030X.

[76] A. Boujelben, A. C. Chaabani, H. Tmar, and M. Abid. "Feature extraction from contours shape for tumor analyzing in mammographic images". In: *2009 Digital Image Computing: Techniques and Applications*. 2009, pp. 395–399. ISBN: 9780769538662.

[77] E. I. Zacharaki, S. Wang, S. Chawla, D. S. Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos. "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme". In: *Magnetic Resonance in Medicine* 62.6 (Dec. 2009), pp. 1609–1618. ISSN: 0740-3194.

[78] A. Wibmer, H. Hricak, T. Gondo, et al. "Haralick Texture Analysis of prostate MRI: Utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason Scores". In: *European Radiology* 25 (2015), pp. 2840–2850.

[79] A. Chaddad, C. Desrosiers, and M. Toews. "GBM heterogeneity characterization by radiomic analysis of phenotype anatomical planes". In: *Medical Imaging 2016: Image Processing*. Ed. by M. A. Styner and E. D. Angelini. Vol. 9784. SPIE, 2016, pp. 568–574.

[80] W. Zhou, L. Zhang, K. Wang, S. Chen, G. Wang, Z. Liu, and C. Liang. "Malignancy characterization of hepatocellular carcinomas based on texture analysis of contrast-enhanced MR images". In: *Journal of Magnetic Resonance Imaging* 45.5 (May 2017), pp. 1476–1484. ISSN: 1522-2586.

[81] D. Yang, G. Rao, J. Martinez, A. Veeraraghavan, and A. Rao. "Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma". In: *Medical Physics* 42.11 (Nov. 2015), pp. 6725–6735. ISSN: 0094-2405.

[82] P Tiwari, P Prasanna, L Wolansky, et al. "Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric mri: A feasibility study". In: *American Journal of Neuroradiology* 37.12 (Dec. 2016), pp. 2231–2236. ISSN: 1936-959X.

[83] Z. Zhang, J. Yang, A. Ho, et al. "A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images". In: *European Radiology* 28.6 (June 2018), pp. 2255–2263. ISSN: 1432-1084.

[84] L. Peng, V. Parekh, P. Huang, et al. "Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics". In: *International Journal of Radiation Oncology Biology Physics* 102.4 (Nov. 2018), pp. 1236–1243. ISSN: 1879-355X.

[85] C. Wang, W. Sun, J. Kirkpatrick, Z. Chang, and F.-F. Yin. "Assessment of concurrent stereotactic radiosurgery and bevacizumab treatment of recurrent malignant gliomas using multimodality MRI imaging and radiomics analysis." In: *Journal of radiosurgery and SBRT* 5.3 (2018), pp. 171–181. ISSN: 2156-4647.

[86] R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural Features for Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3.6 (Nov. 1973), pp. 610–621. ISSN: 0018-9472.

[87] M. M. Galloway. "Texture analysis using gray level run lengths". In: *Computer Graphics and Image Processing* 4.2 (1975), pp. 172–179. ISSN: 0146-664X.

[88] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari. "Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification". In: *International Conference on Pattern Recognition and Information Processing, PRIP 2009*. 2009, pp. 140–145.

[89] M. Zhou, J. Scott, B. Chaudhury, et al. "Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches". In: *American Journal of Neuroradiology* 39.2 (Feb. 2018), pp. 208–216. ISSN: 1936-959X.

[90] T. Scheeve, M. R. Struyvenberg, W. L. Curvers, A. J. de Groof, E. J. Schoon, J. J. G. H. M. Bergman, F. van der Sommen, and P. H. N. de With. "A novel clinical gland feature for detection of early Barrett's neoplasia using volumetric laser endomicroscopy". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. SPIE, 2019, pp. 498–503.

[91] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger. "Deep learning in medical imaging and radiation therapy". In: *Medical Physics* 46.1 (Jan. 2019), e1–e36. ISSN: 0094-2405.

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90. ISSN: 1557-7317.

[93] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson. "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions". In: *Journal of Digital Imaging* 30.4 (Aug. 2017), pp. 449–459. ISSN: 1618-727X.

[94] A. J. de Groof, M. R. Struyvenberg, J. van der Putten, et al. "Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking". In: *Gastroenterology* 158.4 (Mar. 2020), 915–929.e4. ISSN: 1528-0012.

[95] J. Z. Cheng, D. Ni, Y. H. Chou, J. Qin, C. M. Tiu, Y. C. Chang, C. S. Huang, D. Shen, and C. M. Chen. "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans". In: *Scientific Reports* 6 (Apr. 2016). ISSN: 2045-2322.

[96] P. Courtiol, C. Maussion, M. Moarii, et al. "Deep learning-based classification of mesothelioma improves prediction of patient outcome". In: *Nature Medicine* 25.10 (Oct. 2019), pp. 1519–1525. ISSN: 1546-170X.

[97] P. P. J. H. Langenhuizen, S. Zinger, P. E. J. Hanssens, H. P. M. Kunst, J. J. S. Mulder, S. Leenstra, P. H. N. de With, and J. B. Verheul. "Influence of pretreatment growth rate on Gamma Knife treatment response for vestibular schwannoma: a volumetric analysis". In: *Journal of Neurosurgery* 131.5 (Nov. 2019), pp. 1405–1412. ISSN: 0022-3085.

[98] P. Langenhuizen, S. Zinger, P. Hanssens, H. Kunst, J. Mulder, S. Leenstra, P. H. N. de With, and H. B. Verheul. "Correlation between pre-treatment growth rate and tumor control of vestibular schwannomas after gamma knife radiosurgery in the dutch database". In: *ISRS 2017 Abstracts of presentations from the 13th International Stereotactic Radiosurgery Society Congress (ISRS), 28 May - 1 June 2017, Montreux, Switzerland*. Ed. by S. Ryu. Journal of radiosurgery and SBRT. United States: Old City Publishing, 2017, p. 4. ISBN: 978-1-933153-34-6.

[99] P. P. J. H. Langenhuizen and H. B. Verheul. "Invloed van groeisnelheid op de effectiviteit van Gamma Knife-behandeling van vestibulair schwannomen". In: *Tijdschift voor Neurologie en Neurochirurgie*. Vol. 119. 2. Ariez BV, 2018, p. 67.

[100] H. B. Verheul, P. P. J. H. Langenhuizen, B. van der Pol, S. Leenstra, G. Beute, S. te Lie, and P. Hanssens. "The role of marginal dose on tumor control in vestibular schwannoma: a large single institution matched cohort study". 2016.

[101] P. P. J. H. Langenhuizen, H. van Gorp, S. Zinger, H. B. Verheul, S. Leenstra, and P. H. N. de With. "Dose distribution as outcome predictor for Gamma Knife radiosurgery on vestibular schwannoma". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Ed. by K. Mori and H. K. Hahn. Vol. 10950. SPIE, 2019, pp. 1090 –1098.

[102] P. P. J. H. Langenhuizen, Y. Zeng, S. Zinger, H. B. Verheul, S. Leenstra, and P. H. N. de With. "Treatment delineation impact on Gamma Knife radiosurgical response of vestibular schwannoma". In: *Proceedings of the 2017 Symposium on Information Theory and Signal Processing in the Benelux, May 11-12, 2017, Delft, The Netherlands*. Ed. by R. Heusdens and J. Weber. Netherlands: Delft University of Technology, 2017, pp. 133–140.

[103] T. Scheeve, P. P. J. H. Langenhuizen, S. Zinger, and P. H. N. de With. "Outcome prediction of Gamma Knife radiosurgery on vestibular schwannoma using contour-based shape descriptors". 2017.

[104] P. P. J. H. Langenhuizen, M. J. W. Legters, S. Zinger, H. B. Verheul, P. H. N. de With, and S. Leenstra. "MRI textures as outcome predictor for Gamma Knife radiosurgery on vestibular schwannoma". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Ed. by K. Mori and N. Petrick. Vol. 10575. SPIE, Feb. 2018, pp. 112 –120. ISBN: 9781510616394.

[105] P. Langenhuizen, S. H. P. Sebregts, S. Zinger, S. Leenstra, P. Hanssens, P. de With, and H. B. Verheul. "Predictability of transient tumor enlargement following gamma knife radiosurgery on vestibular schwannoma". In: *The 14th International Stereotactic Radiosurgery Society Congress; Rio de Janeiro, Brazil*. 2019, Abstract nr. a28–5.

[106] P. P. J. H. Langenhuizen, S. Zinger, S. Leenstra, P. E. J. Hanssens, P. H. N. de With, and H. B. Verheul. "Predictability of Gamma Knife radiosurgical response of vestibular schwannoma". 2019.

[107] P. P. J. H. Langenhuizen, S. H. P. Sebregts, S. Zinger, S. Leenstra, J. B. Verheul, and P. H. N. de With. "Prediction of transient tumor enlargement using MRI tumor texture after radiosurgery on vestibular schwannoma". In: *Medical Physics* 47.4 (Apr. 2020), pp. 1692–1701. ISSN: 0094-2405.

[108] P. P. J. H. Langenhuizen, S. Zinger, S. Leenstra, H. P. M. Kunst, J. J. S. Mulder, P. E. J. Hanssens, P. H. N. de With, and J. B. Verheul. "Radiomics-based prediction of long-term treatment response of vestibular schwannomas following stereotactic radiosurgery". In: *Otology & Neurotology* 41.10 (Dec. 2020), e1321–e1327. ISSN: 1531-7129.

[109] L Leksell. "A note on the treatment of acoustic tumours." In: *Acta chirurgica Scandinavica* 137.8 (1971), pp. 763–765. ISSN: 0001-5482.

[110] W. T. Millar, J. W. Hopewell, I. Paddick, C. Lindquist, H. Nordström, P. Lidberg, and J. Gårding. "The role of the concept of biologically effective dose (BED) in treatment planning in radiosurgery". In: *Physica Medica* 31.6 (2015), pp. 627–633. ISSN: 1724-191X.

[111] J. K. Varughese, C. Breivik, T. Wentzel-Larsen, and M. Lund-Johansen. "Growth of untreated vestibular schwannoma: a prospective study". In: *Journal of Neurosurgery* 116.4 (2012), pp. 706–712.

[112] C. Tuleasca, M. George, R. Maire, L. Schiappacasse, M. Marguet, R. T. Daniel, and M. Levivier. "Letter: Cystic vestibular schwannomas respond best to radiosurgery". In: *Clinical Neurosurgery* 81.6 (2017), E80–E82. ISSN: 0148-396X.

[113] M. K. Bassim, K. I. Berliner, L. M. Fisher, D. E. Brackmann, and R. A. Friedman. "Radiation therapy for the treatment of vestibular schwannoma: A critical evaluation of the state of the literature". In: *Otology and Neurotology* 31.4 (June 2010), pp. 567–573. ISSN: 1531-7129.

[114] H. Speckter, J. Bido, G. Hernandez, D. Rivera, L. Suazo, S. Valenzuela, I. Miches, J. Oviedo, C. Gonzalez, and P. Stoeter. "Pretreatment texture analysis of routine MR images and shape analysis of the diffusion tensor for prediction of volumetric response after radiosurgery for meningioma". In: *Journal of Neurosurgery* 129.Suppl_1 (2018), pp. 31–37. ISSN: 1933-0693.

[115] V. Kaul and M. K. Cosetti. "Management of Vestibular Schwannoma (Including NF2): Facial Nerve Considerations". In: *Otolaryngologic Clinics of North America* 51.6 (Dec. 2018), pp. 1193–1212. ISSN: 1557-8259.

[116] P. Therasse, S. G. Arbuck, E. A. Eisenhauer, et al. "New guidelines to evaluate the response to treatment in solid tumors". In: *Journal of the National Cancer Institute* 92.3 (2000), pp. 205–216. ISSN: 0027-8874.

[117] J. Kanzaki, M. Tos, M. Sanna, and D. A. Moffat. "New and modified reporting systems from the consensus meeting on systems for reporting results in vestibular schwannoma". In: *Otology and Neurotology* 24.4 (July 2003), pp. 642–649.

[118] E. A. Eisenhauer, P. Therasse, J. Bogaerts, et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)". In: *European Journal of Cancer* 45.2 (Jan. 2009), pp. 228–247. ISSN: 0959-8049.

[119] D. Kondziolka and A. Wolf. "Commentary: Ten-Year Follow-up on Tumor Growth and Hearing in Patients Observed With an Intracanalicular Vestibular Schwannoma". In: *Neurosurgery* 80.1 (2016), pp. 57–59. ISSN: 1524-4040.

[120] R. Van De Langenberg, B. J. De Bondt, P. J. Nelemans, B. G. Baumert, and R. J. Stokroos. "Follow-up assessment of vestibular schwannomas: Volume quantification versus two-dimensional measurements". In: *Neuroradiology* 51.8 (2009), pp. 517–524. ISSN: 0028-3940.

[121] P. David Mozley, C. Bendtsen, B. Zhao, L. H. Schwartz, M. Thorn, Y. Rong, L. Zhang, A. Perrone, R. Korn, and A. J. Buckler. "Measurement of tumor volumes improves RECIST-based response assessments in advanced lung cancer". In: *Translational Oncology* 5.1 (2012), pp. 19–25. ISSN: 1936-5233.

[122] J. K. Varughese, T. Wentzel-Larsen, F. Vassbotn, G. Moen, and M. Lund-Johansen. "Analysis of vestibular schwannoma size in multiple dimensions: A comparative cohort study of different measurement techniques". In: *Clinical Otolaryngology* 35.2 (2010), pp. 97–103. ISSN: 1749-4478.

[123] C. P. Yu, J. Y. C. Cheung, S. Leung, and R. Ho. "Sequential volume mapping for confirmation of negative growth in vestibular schwannomas treated by gamma knife radiosurgery". In: *Journal of Neurosurgery* 93.suppl_3 (2000), pp. 82 –89.

[124] F. F. Mohammed, M. L. Schwartz, A. Lightstone, D. J. Beachey, and M. N. Tsao. "Pseudoprogression of vestibular schwannomas after fractionated stereotactic radiation therapy". In: *Journal of Radiation Oncology* 2.1 (2013), pp. 15–20. ISSN: 1948-7894.

[125] L. Hathout, C. Lambert, J.-F. Carrier, J.-P. Bahary, Y. Hervieux, R. A. Moumdjian, M.-A. Fortin, and D. Roberge. "Transient Tumor Volume Increase in Vestibular Schwannomas After Radiotherapy". In: *Cureus* 4.11 (Nov. 2012), e70. ISSN: 2168-8184.

[126] H. Aoyama, S. Onodera, N. Takeichi, R. Onimaru, S. Terasaka, Y. Sawamura, and H. Shirato. "Symptomatic outcomes in relation to tumor expansion after fractionated stereotactic radiation therapy for vestibular schwannomas: Single-institutional long-term experience". In: *International Journal of Radiation Oncology Biology Physics* 85.2 (Feb. 2013), pp. 329–334. ISSN: 0360-3016.

[127] H. Nakamura, H. Jokura, K. Takahashi, N. Boku, A. Akabane, and T. Yoshimoto. "Serial follow-up MR imaging after gamma knife radiosurgery for vestibular schwannoma". In: *American Journal of Neuroradiology* 21.8 (2000), pp. 1540–1546. ISSN: 0195-6108.

[128] B. Wowra, A. Muacevic, A. Jess-Hempen, J. M. Hempel, S. Müller-Schunk, and J. C. Tonn. "Outpatient gamma knife surgery for vestibular schwannoma: definition of the therapeutic

profile based on a 10-year experience." In: *Journal of neurosurgery* 102.Special_Supplement (Jan. 2005), pp. 114–118. ISSN: 1933-0693.

[129] A. T. C. J. van Eck and G. A. Horstmann. "Increased preservation of functional hearing after gamma knife surgery for vestibular schwannoma." In: *Journal of neurosurgery* 102.Special_-Supplement (2013), pp. 204–206. ISSN: 1933-0693.

[130] T. Okunaga, T. Matsuo, N. Hayashi, Y. Hayashi, H. K. Shabani, M. Kaminogo, M. Ochi, and I. Nagata. "Linear accelerator radiosurgery for vestibular schwannoma: Measuring tumor volume changes on serial three-dimensional spoiled gradient-echo magnetic resonance images". In: *Journal of Neurosurgery* 103.1 (July 2005), pp. 53–58. ISSN: 0022-3085.

[131] R. Van De Langenberg, A. J. Dohmen, B. J. De Bondt, P. J. Nelemans, B. G. Baumert, and R. J. Stokroos. "Volume changes after stereotactic LINAC radiotherapy in vestibular schwannoma: Control rate and growth patterns". In: *International Journal of Radiation Oncology Biology Physics* 84.2 (2012), pp. 343–349. ISSN: 0360-3016.

[132] O. Nagano, Y. Higuchi, T. Serizawa, J. Ono, S. Matsuda, I. Yamakami, and N. Saeki. "Transient expansion of vestibular schwannoma following stereotactic radiosurgery: Clinical article". In: *Journal of Neurosurgery* 109.5 (Nov. 2008), pp. 811–816. ISSN: 0022-3085.

[133] C. C. Lee, H. M. Wu, W. Y. Chung, C. J. Chen, D. H. C. Pan, and S. P. Hsu. "Microsurgery for vestibular schwannoma after Gamma Knife surgery: challenges and treatment strategies". In: *Journal of neurosurgery* 121.Suppl_2 (Dec. 2014), pp. 150–159. ISSN: 1933-0693.

[134] C. Hayhurst and G. Zadeh. "Tumor pseudoprogression following radiosurgery for vestibular schwannoma". In: *Neuro-Oncology* 14.1 (2012), pp. 87–92. ISSN: 1522-8517.

[135] B. E. Pollock, C. L. Driscoll, R. L. Foote, M. J. Link, D. A. Gorman, C. D. Bauch, J. N. Mandrekar, K. N. Krecke, and C. H. Johnson. "Patient Outcomes After Vestibular Schwannoma Management: a Prospective Comparison of Microsurgical Resection and Stereotactic Radiosurgery". In: *Neurosurgery* 59.1 (July 2006), pp. 77–85. ISSN: 0148-396X.

[136] K.-M. Kim, C.-K. Park, H.-T. Chung, S. H. Paek, H.-W. Jung, and D. G. Kim. "Long-term Outcomes of Gamma Knife Stereotactic Radiosurgery of Vestibular Schwannomas". In: *Journal of Korean Neurosurgical Society* 42.4 (Oct. 2007), p. 286. ISSN: 1225-8245.

[137] T. Mindermann and I. Schlegel. "How to distinguish tumor growth from transient expansion of vestibular schwannomas following Gamma Knife radiosurgery". In: *Acta Neurochirurgica* 156.6 (2014), pp. 1121–1123. ISSN: 0942-0940.

[138] D. Kondziolka, L. D. Lunsford, M. R. Mclaughlin, and J. C. Flickinger. "Long-term outcomes after radiosurgery for acoustic neuromas". In: *New England Journal of Medicine* 339.20 (Nov. 1998), pp. 1426–1433. ISSN: 0028-4793.

[139] F. van der Sommen, S. Zinger, E. J. Schoon, and P. H. N. de With. "Sweet-spot training for early esophageal cancer detection". In: *Medical Imaging 2016: Computer-Aided Diagnosis*. Ed. by G. D. Tourassi and S. G. A. III. Vol. 9785. International Society for Optics and Photonics. SPIE, 2016, pp. 330 –336.

[140] S. K. Warfield, K. H. Zou, and W. M. Wells. "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation". In: *IEEE Transactions on Medical Imaging* 23.7 (July 2004), pp. 903–921. ISSN: 0278-0062.

[141] J. Martin Bland and D. G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The Lancet* 327.8476 (Feb. 1986), pp. 307–310. ISSN: 0140-6736.

[142] K. Suzuki. "Overview of deep learning in medical imaging". In: *Radiological Physics and Technology* 10.3 (Sept. 2017), pp. 257–273. ISSN: 1865-0341.

[143] A. F. Swager, F. van der Sommen, S. R. Klomp, S. Zinger, S. L. Meijer, E. J. Schoon, J. J. Bergman, P. H. de With, and W. L. Curvers. "Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy". In: *Gastrointestinal Endoscopy* 86.5 (2017), pp. 839–846. ISSN: 1097-6779.

[144] F. Ghazvinian Zanjani, S. Zinger, and P. H. N. de With. "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces". In: *Medical Imaging 2018: Digital Pathology*. Ed. by M. N. Gurcan and J. E. Tomaszewski. Vol. 10581. SPIE, Mar. 2018, p. 17. ISBN: 9781510616516.

[145] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (Mar. 1986), pp. 81–106. ISSN: 0885-6125.

[146] B. E. Boser, I. M. Guyon, and V. N. Vapnik. "Training algorithm for optimal margin classifiers". In: *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*. New York, New York, USA: Publ by ACM, 1992, pp. 144–152. ISBN: 089791497X.

[147] Q. Li. "Reliable Evaluation of Performance Level for Computer-Aided Diagnostic Scheme". In: *Academic Radiology* 14.8 (Aug. 2007), pp. 985–991. ISSN: 1076-6332.

[148] M. M. Santoni, D. I. Sensuse, A. M. Arymurthy, and M. I. Fanany. "Cattle Race Classification Using Gray Level Co-occurrence Matrix Convolutional Neural Networks". In: *Procedia Computer Science*. Vol. 59. 2015, pp. 493–502.

[149] Q. Chen and E. Agu. "Exploring Statistical GLCM Texture Features for Classifying Food Images". In: *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*. 2015, p. 453. ISBN: 9781467395489.

[150] J. Liu, Y. Mao, Z. Li, D. Zhang, Z. Zhang, S. Hao, and B. Li. "Use of texture analysis based on contrast-enhanced MRI to predict treatment response to chemoradiotherapy in nasopharyngeal carcinoma". In: *Journal of Magnetic Resonance Imaging* 44.2 (Aug. 2016), pp. 445–455. ISSN: 1522-2586.

[151] R. Ortiz-Ramón, A. Larroza, S. Ruiz-España, E. Arana, and D. Moratal. "Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study". In: *European Radiology* 28.11 (Nov. 2018), pp. 4514–4523. ISSN: 1432-1084.

[152] K. Ekert, C. Hinterleitner, K. Baumgartner, J. Fritz, and M. Horger. "Extended texture analysis of non-enhanced whole-body mri image data for response assessment in multiple myeloma patients undergoing systemic therapy". In: *Cancers* 12.3 (Mar. 2020), p. 761. ISSN: 2072-6694.

[153] P. Brynolfsson, D. Nilsson, T. Torheim, T. Asklund, C. T. Karlsson, J. Trygg, T. Nyholm, and A. Garpebring. "Haralick texture features from apparent diffusion coefficient (ADC) MRI images depend on imaging and pre-processing parameters". In: *Scientific Reports* 7.1 (Dec. 2017). ISSN: 2045-2322.

[154] A. Kunimatsu, N. Kunimatsu, K. Kamiya, T. Watadani, H. Mori, and O. Abe. "Comparison between glioblastoma and primary central nervous system lymphoma using MR image-based texture analysis". In: *Magnetic Resonance in Medical Sciences* 17.1 (2018), pp. 50–57. ISSN: 1880-2206.

[155] H. Minkowski. "Volumen und Oberfläche". In: *Mathematische Annalen* 57 (1903), pp. 447–495.

[156] H. Mantz, K. Jacobs, and K. Mecke. "Utilizing Minkowski functionals for image analysis: A marching square algorithm". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.12 (2008). ISSN: 1742-5468.

[157] M. Kerscher, K. Mecke, J. Schmalzing, C. Beisbart, T. Buchert, and H. Wagner. "Morphological fluctuations of large-scale structure: The PSCz survey". In: *Astronomy and Astrophysics* 373.1 (July 2001), pp. 1–11. ISSN: 0004-6361.

[158] C. H. Arns, M. A. Knackstedt, and K. R. Mecke. "Reconstructing complex materials via effective grain shapes". In: *Physical Review Letters* 91.21 (Nov. 2003), p. 215506. ISSN: 1079-7114.

[159] K. R. Mecke and D. Stoyan. "Morphological characterization of point patterns". In: *Biometrical Journal* 47.4 (Aug. 2005), pp. 473–488. ISSN: 0323-3847.

[160] H. F. Boehm, T. Vogel, A. Panteleon, D. Burklein, H. Bitterling, and M. Reiser. "Differentiation between post-menopausal women with and without hip fractures: Enhanced evaluation of clinical DXA by topological analysis of the mineral distribution in the scan images". In: *Osteoporosis International* 18.6 (June 2007), pp. 779–787. ISSN: 0937-941X.

[161] M. B. Huber, M. B. Nagarajan, G. Leinsinger, R. Eibel, L. A. Ray, and A. Wismüller. "Performance of topological texture features to classify fibrotic interstitial lung disease patterns". In: *Medical Physics* 38.4 (2011), pp. 2035–2044. ISSN: 0094-2405.

[162] X. Li, P. R. S. Mendonça, and R. Bhotika. "Texture analysis using Minkowski functionals". In: *Medical Imaging 2012: Image Processing*. Ed. by D. R. Haynor and S. Ourselin. Vol. 8314. International Society for Optics and Photonics, Feb. 2012, 83144Y.

[163] M. B. Nagarajan, M. B. Huber, T. Schlossbauer, G. Leinsinger, A. Krol, and A. Wismüller. "Classification of small lesions in dynamic breast MRI: Eliminating the need for precise lesion segmentation through spatio-temporal analysis of contrast enhancement". In: *Machine Vision and Applications*. Vol. 24. 7. NIH Public Access, 2013, pp. 1371–1381.

[164] H. Hadwiger. *Vorlesungen Über Inhalt, Oberfläche und Isoperimetrie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1957. ISBN: 978-3-642-94703-2.

[165] N. N. Niu, A. Niemierko, M. Larvie, H. Curtin, J. S. Loeffler, M. J. McKenna, and H. A. Shih. "Pretreatment growth rate predicts radiation response in vestibular schwannomas". In: *International Journal of Radiation Oncology Biology Physics* 89.1 (May 2014), pp. 113–119. ISSN: 1879-355X.

[166] J. Régis, C. Delsanti, and P.-h. Roche. "Editorial: Vestibular schwannoma radiosurgery: progression or pseudoprogression?" In: *Journal of Neurosurgery* 127.2 (2017), pp. 374–379. ISSN: 0022-3085.

[167] R. Del Valle, M. Pérez, J. Ortiz, et al. "Stereotactic noninvasive volume measurement compared with geometric measurement for indications and evaluation of gamma knife treatment". In: *Journal of Neurosurgery* 102.Special_Supplement (Nov. 2005), pp. 140–142. ISSN: 0022-3085.

[168] B. Sahin, N. Acer, O. F. Sonmez, M. Emirzeoglu, H. Basaloglu, A. Uzun, and S. Bilgic. "Comparison of four methods for the estimation of intracranial volume: A gold standard study". In: *Clinical Anatomy* 20.7 (Oct. 2007), pp. 766–773. ISSN: 0897-3806.

[169] D. A. Moffat, J. Golledge, D. M. Baguley, and D. G. Hardy. "Clinical correlates of acoustic neuroma morphology". In: *The Journal of Laryngology & Otology* 107.4 (1993), pp. 290–294. ISSN: 1748-5460.

[170] T. Sakamoto, S. Fukuda, and Y. Inuyama. "Hearing loss and growth rate of acoustic neuromas in follow-up observation policy". In: *Auris Nasus Larynx* 28.Supplement_1 (2001), S23–S27. ISSN: 0385-8146.

[171] Y. Agrawal, J. H. Clark, C. J. Limb, J. K. Niparko, and H. W. Francis. "Predictors of vestibular schwannoma growth and clinical implications". In: *Otology and Neurotology* 31.5 (July 2010), pp. 807–812. ISSN: 1531-7129.

[172] P. C. Roehm and B. J. Gantz. "Management of acoustic neuromas in patients 65 years or older". In: *Otology and Neurotology* 28.5 (Aug. 2007), pp. 708–714. ISSN: 1531-7129.

[173] U. Patnaik, S. C. Prasad, H. Tutar, A. L. Giannuzzi, A. Russo, and M. Sanna. "The long-term outcomes of wait-and-scan and the role of radiotherapy in the management of vestibular schwannomas". In: *Otology and Neurotology* 36.4 (Apr. 2015), pp. 638–646. ISSN: 1537-4505.

[174] C. E. Reddy, H. G. Lewis-Jones, M. Javadpour, I. Ryland, and T. H. Lesser. "Conservative management of vestibular schwannomas of 15 to 31 mm intracranial diameter". In: *Journal of Laryngology and Otology* 128.9 (Sept. 2014), pp. 752–758. ISSN: 1748-5460.

[175] W. El Bakkouri, R. E. Kania, J. P. Guichard, G. Lot, P. Herman, and P. T. B. Huy. "Conservative management of 386 cases of unilateral vestibular schwannoma: Tumor growth and consequences for treatment - Clinical article". In: *Journal of Neurosurgery* 110.4 (Apr. 2009), pp. 662–669. ISSN: 0022-3085.

[176] P. Caye-Thomasen, S. Hansen, T. Dethloff, S. E. Stangerup, and J. Thomsen. "Sublocalization and volumetric growth pattern of intracanalicular vestibular schwannomas". In: *Laryngoscope* 116.7 (July 2006), pp. 1131–1135. ISSN: 0023-852X.

[177] A. Mohyuddin, E. A. Vokurka, D. G. Evans, R. T. Ramsden, and A. Jackson. "Is clinical growth index a reliable predictor of tumour growth in vestibular schwannomas?" In: *Clinical Otolaryngology and Allied Sciences* 28.2 (Apr. 2003), pp. 85–90. ISSN: 0307-7772.

[178] M. Diensthuber, T. Lenarz, and T. Stöver. "Determination of the clinical growth index in unilateral vestibular schwannoma". In: *Skull Base* 16.1 (Feb. 2006), pp. 31–38. ISSN: 1531-5010.

[179] A. Herwadker, E. A. Vokurka, D. G. R. Evans, R. T. Ramsden, and A. Jackson. "Size and growth rate of sporadic vestibular schwannoma: Predictive value of information available at presentation". In: *Otology and Neurotology* 26.1 (Jan. 2005), pp. 86–92. ISSN: 1531-7129.

[180] E. A. Sarapata and L. G. de Pillis. "A Comparison and Catalog of Intrinsic Tumor Growth Models". In: *Bulletin of Mathematical Biology* 76.8 (Sept. 2014), pp. 2010–2024. ISSN: 1522-9602.

[181] A. Talkington and R. Durrett. "Estimating Tumor Growth Rates In Vivo". In: *Bulletin of Mathematical Biology* 77.10 (Oct. 2015), pp. 1934–1954. ISSN: 1522-9602.

[182] V. X. Fu, J. B. Verheul, G. N. Beute, S. Leenstra, H. P. Kunst, J. J. Mulder, and P. E. Hanssens. "Retreatment of vestibular schwannoma with Gamma Knife radiosurgery: Clinical outcome, tumor control, and review of literature". In: *Journal of Neurosurgery* 129.1 (July 2018), pp. 137–145. ISSN: 1933-0693.

[183] A. H. Yeung, M. E. Sughrue, A. J. Kane, T. Tihan, S. W. Cheung, and A. T. Parsa. "Radiobiology of vestibular schwannomas: mechanisms of radioresistance and potential targets for therapeutic sensitization". In: *Neurosurgical Focus* 27.6 (2009), E2. ISSN: 1092-0684.

[184] T. M. Pawlik and K. Keyomarsi. "Role of cell cycle in mediating sensitivity to radiotherapy". In: *International Journal of Radiation Oncology Biology Physics* 59.4 (July 2004), pp. 928–942. ISSN: 0360-3016.

[185] L. Zhu, K. Wu, S. Ma, and S. Zhang. "HDAC inhibitors: a new radiosensitizer for non-small-cell lung cancer". In: *Tumori* 101.3 (2015), pp. 257–262. ISSN: 0300-8916.

[186] K. Hastak, S. Bhutra, R. Parry, and J. M. Ford. "Poly (ADP-ribose) polymerase inhibitor, an effective radiosensitizer in lung and pancreatic cancers". In: *Oncotarget* 8.16 (Apr. 2017), pp. 26344–26355. ISSN: 1949-2553.

[187] H. Wang, X. Mu, H. He, and X.-D. Zhang. "Cancer Radiosensitizers". In: *Trends in Pharmacological Sciences* 39.1 (2017), pp. 24–48. ISSN: 0165-6147.

[188] W. Y. Yue, J. J. Clark, M. Telisak, and M. R. Hansen. "Inhibition of c-Jun N-terminal kinase activity enhances vestibular schwannoma cell sensitivity to γ-irradiation". In: *Neurosurgery* 73.3 (2013), pp. 506–516. ISSN: 1537-8276.

[189] R. Chopra, D. Kondziolka, A. Niranjan, L. D. Lunsford, and J. C. Flickinger. "Long-Term Follow-up of Acoustic Schwannoma Radiosurgery With Marginal Tumor Doses of 12 to 13 Gy". In: *International Journal of Radiation Oncology Biology Physics* 68.3 (2007), pp. 845–851. ISSN: 0360-3016.

[190] J Flickinger, D. Kondziolka, M. C. Niranjan, A. Maitz, G. Voynov, and L. D. Lunsford. "Acoustic neuroma radiosurgery with marginal tumor doses of 12 to 13 Gy". In: *International Journal of Radiation Oncology Biology Physics* 60.1 (2004), pp. 225–230. ISSN: 0360-3016.

[191] E. Myrseth, P. Møller, P.-H. Pedersen, and M. Lund-Johansen. "Vestibular Schwannoma: Surgery or Gamma Knife Radiosurgery? A Prospective, Nonrandomized Study". In: *Neurosurgery* 64.4 (2009), pp. 654–661.

[192] K. Nakaya, A. Niranjan, D. Kondziolka, H. Kano, A. A. Khan, B. Nettel, C. Koebbe, S. Pirris, J. C. Flickinger, and L. D. Lunsford. "Gamma Knife Radiosurgery for Benign Tumors With Symptoms From Brainstem Compression". In: *International Journal of Radiation Oncology Biology Physics* 77.4 (2010), pp. 988–995. ISSN: 0360-3016.

[193] C. Yu, G. Jozsef, M. L. Apuzzo, Z. Petrovich, J. C. Flickinger, F. Colombo, and B. E. Pollock. "Dosimetric Comparison of Cyberknife with Other Radiosurgical Modalities for an Ellipsoidal Target". In: *Neurosurgery* 53.5 (Nov. 2003), pp. 1155–1163. ISSN: 0148-396X.

[194] T. Miller, T. Lau, R. Vasan, C. Danner, A. Samy Youssef, H. Van Loveren, and S. Agazzi. "Reporting success rates in the treatment of vestibular schwannomas: Are we accounting for the natural history?" In: *Journal of Clinical Neuroscience* 21.6 (2014), pp. 914–918. ISSN: 1532-2653.

[195] T. Kataria, K. Sharma, V. Subramani, K. Karrthick, and S. Bisht. "Homogeneity Index: An objective tool for assessment of conformal radiation treatments". In: *Journal of Medical Physics* 37.4 (2012), p. 207. ISSN: 0971-6203.

[196] N. Dalal and W. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05* 1.3 (2004), pp. 886–893. ISSN: 1063-6919.

[197] A. Fenster and B. Chiu. "Evaluation of Segmentation algorithms for Medical Imaging." In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. February. IEEE, 2005, pp. 7186–7189. ISBN: 0-7803-8741-4.

[198] D Nicoll and M Pignone. "Basic Principles of Diagnostic Test Use and Interpretation". In: *Pocket guide to diagnostic tests*. third edit. New York: McGraw-Hill, 2001. Chap. 1, pp. 1–21. ISBN: 9780443065705.

[199] P Jaccard. "The distribution of the flora in the alphine zone". In: *The New Phytologist* XI.2 (1912), pp. 37–50. ISSN: 1469-8137.

[200] L. R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302.

[201] A. A. Taha and A. Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." In: *BMC medical imaging* 15 (2015), p. 29. ISSN: 1471-2342.

[202] R. Shi, K. N. Ngan, and S. Li. "The objective evaluation of image object segmentation quality". In: *Advanced Concepts for Intelligent Vision Systems. ACIVS 2013. Lecture Notes in Computer Science*. Vol. 8192. 2013. ISBN: 9783319028941.

[203] R. S. Montero and E. Bribiesca. "State of the art of compactness and circularity measures". In: *International Mathematical Forum* 4.27 (2009), pp. 1305 –1335.

[204] R. M. Haralick. "A Measure for Circularity of Digital Figures". In: *IEEE Transactions on Systems, Man and Cybernetics* 4.4 (1974), pp. 394–396. ISSN: 2168-2909.

[205] P. E. Danielson. "A new shape factor". In: *Computer Graphics and Image Processing* 7.2 (Apr. 1978), pp. 292–299. ISSN: 0146-664X.

[206] E. Bribiesca. "Measuring 2-D shape compactness using the contact perimeter". In: *Computers and Mathematics with Applications* 33.11 (June 1997), pp. 1–9. ISSN: 0898-1221.

[207] A. Amanatiadis, V. G. Kaburlasos, A. Gasteratos, and S. E. Papadakis. "Evaluation of shape descriptors for shape-based image retrieval". In: *IET Image Processing* 5.5 (Aug. 2011), pp. 493–499. ISSN: 1751-9659.

[208] L. Gupta and M. D. Srinath. "Contour sequence moments for the classification of closed planar shapes". In: *Pattern Recognition* 20.3 (Jan. 1987), pp. 267–272. ISSN: 0031-3203.

[209] C. Parmar, E. R. Velazquez, R. Leijenaar, et al. "Robust radiomics feature quantification using semiautomatic volumetric segmentation". In: *PLoS ONE* 9.7 (July 2014). Ed. by G. E. Woloschak, e102107. ISSN: 1932-6203.

[210] X. Wei. *Gray Level Run Length Matrix Toolbox v1.0, Software, Beijing Aeronautical Technology Research Center*. 2007.

[211] N. Czarnek, K. Clark, K. B. Peters, and M. A. Mazurowski. "Algorithmic three-dimensional analysis of tumor shape in MRI improves prognosis of survival in glioblastoma: a multi-institutional study". In: *Journal of Neuro-Oncology* 132.1 (Mar. 2017), pp. 55–62. ISSN: 1573-7373.

[212] H. Shirato, T. Sakamoto, N. Takeichi, H. Aoyama, K. Suzuki, K. Kagei, T. Nishioka, S. Fukuda, Y. Sawamura, and K. Miyasaka. "Fractionated stereotactic radiotherapy for vestibular schwannoma (VS): Comparison between cystic-type and solid-type VS". In: *International Journal of Radiation Oncology Biology Physics* 48.5 (Dec. 2000), pp. 1395–1401. ISSN: 0360-3016.

[213] A. Madabhushi and J. K. Udupa. "New methods of MR image intensity standardization via generalized scale". In: *Medical Physics* 33.9 (Aug. 2006), pp. 3426–3434. ISSN: 0094-2405.

[214] F. van der Sommen, S. Zinger, E. J. Schoon, and P. H. de With. "Supportive automatic annotation of early esophageal cancer using local gabor and color features". In: *Neurocomputing* 144 (Nov. 2014), pp. 92–106. ISSN: 1872-8286.

[215] D. Kondziolka, S. H. Mousavi, H. Kano, J. C. Flickinger, and L. D. Lunsford. "The newly diagnosed vestibular schwannoma: radiosurgery, resection, or observation?" In: *Neurosurgical Focus* 33.3 (2012), E8. ISSN: 1092-0684.

[216] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. "Machine learning applications in cancer prognosis and prediction". In: *Computational and Structural Biotechnology Journal* 13 (2015), pp. 8–17. ISSN: 2001-0370.

Publication List

The following conference and journal papers have been published based on the research presented in this thesis.

Journal articles

[1] P. P. J. H. Langenhuizen, S. Zinger, P. E. J. Hanssens, H. P. M. Kunst, J. J. S. Mulder, S. Leenstra, P. H. N. de With, and J. B. Verheul. "Influence of pretreatment growth rate on Gamma Knife treatment response for vestibular schwannoma: a volumetric analysis". In: *Journal of Neurosurgery* 131.5 (Nov. 2019), pp. 1405–1412. ISSN: 0022-3085.

[2] P. P. J. H. Langenhuizen, S. H. P. Sebregts, S. Zinger, S. Leenstra, J. B. Verheul, and P. H. N. de With. "Prediction of transient tumor enlargement using MRI tumor texture after radiosurgery on vestibular schwannoma". In: *Medical Physics* 47.4 (Apr. 2020), pp. 1692–1701. ISSN: 0094-2405.

[3] P. P. J. H. Langenhuizen, S. Zinger, S. Leenstra, H. P. M. Kunst, J. J. S. Mulder, P. E. J. Hanssens, P. H. N. de With, and J. B. Verheul. "Radiomics-based prediction of long-term treatment response of vestibular schwannomas following stereotactic radiosurgery". In: *Otology & Neurotology* 41.10 (Dec. 2020), e1321–e1327. ISSN: 1531-7129.

International conference contributions

[1] P. P. J. H. Langenhuizen, Y. Zeng, S. Zinger, H. B. Verheul, S. Leenstra, and P. H. N. de With. "Treatment delineation impact on Gamma Knife radiosurgical response of vestibular schwannoma". In: *Proceedings of the 2017 Symposium on Information Theory and Signal Processing in the Benelux, May 11-12, 2017, Delft, The Netherlands*. Ed. by R. Heusdens and J. Weber. Netherlands: Delft University of Technology, 2017, pp. 133–140. ISBN: 978-94-6186-811-4.

[2] P. Langenhuizen, S. Zinger, P. Hanssens, H. Kunst, J. Mulder, S. Leenstra, P. H. N. de With, and H. B. Verheul. "Correlation between pre-treatment growth rate and tumor control of vestibular schwannomas after gamma knife radiosurgery in the dutch database". In: *ISRS 2017 Abstracts of presentations from the 13th International Stereotactic Radiosurgery Society Congress (ISRS)*. Ed. by S. Ryu. Journal of radiosurgery and SBRT. 28 May–1 June 2017, Montreux, Switzerland: Old City Publishing, 2017, p. 4. ISBN: 978-1-933153-34-6.

[3] P. P. J. H. Langenhuizen, M. J. W. Legters, S. Zinger, H. B. Verheul, P. H. N. de With, and S. Leenstra. "MRI textures as outcome predictor for Gamma Knife radiosurgery on vestibular schwannoma". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Ed. by K. Mori and N. Petrick. Vol. 10575. SPIE, 2018, pp. 112 –120. ISBN: 9781510616394.

[4] P. P. J. H. Langenhuizen, H. van Gorp, S. Zinger, H. B. Verheul, S. Leenstra, and P. H. N. de With. "Dose distribution as outcome predictor for Gamma Knife radiosurgery on vestibular schwannoma". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Ed. by K. Mori and H. K. Hahn. Vol. 10950. SPIE, 2019, pp. 1090 –1098. ISBN: 9781510625471.

[5] P. Langenhuizen, S. H. P. Sebregts, S. Zinger, S. Leenstra, P. Hanssens, P. de With, and H. B. Verheul. *Predictability of transient tumor enlargement following gamma knife radiosurgery on vestibular schwannoma*. URL: https://2019.isrscongress.org/en/abstract-book/. Abstract nr. a28–5.

[6] P. Langenhuizen, V. Fu, S. Zinger, S. Leenstra, P. Hanssens, P. de With, and H. B. Verheul. *Gamma Knife radiosurgery following partial resection of large vestibular schwannomas: evaluation of long-term tumor control*. Presented at 14th International Stereotactic Radiosurgery Society Congress, 9–13 June, 2019, Rio de Janeiro, Brazil. URL: https://2019.isrscongress.org/en/abstract-book/. Abstract nr. a28–6.

[7] P. Langenhuizen, S. Zinger, S. Leenstra, P. Hanssens, P. de With, and H. B. Verheul. *Short-term volumetric tumor response as predictor for long-term tumor control after Gamma Knife radiosurgery of vestibular schwannoma*. Presented at 14th International Stereotactic Radiosurgery Society Congress, 9–13 June, 2019, Rio de Janeiro, Brazil. URL: https://2019.isrscongress.org/en/abstract-book/. Abstract nr. c491–3.

[8] P. P. J. H. Langenhuizen, S. Zinger, S. Leenstra, P. E. J. Hanssens, P. H. N. de With, and H. B. Verheul. "Predictability of Gamma Knife radiosurgical response of vestibular schwannoma". Presented at the 8th Quadrennial International Conference on Vestibular Schwannoma and Other CPA Tumors: Advancing Care through Ideas and Innovation, June 18-21, 2019, Rochester, MN.

National conference and other contributions

[1] H. B. Verheul, P. P. J. H. Langenhuizen, B. van der Pol, S. Leenstra, G. Beute, S. te Lie, and P. Hanssens. "The role of marginal dose on tumor control in vestibular schwannoma: a large single institution matched cohort study". Presented at the Leksell Gamma Knife Society Meeting, May 15–19, 2016, Amsterdam, the Netherlands.

[2] T. Scheeve, P. P. J. H. Langenhuizen, S. Zinger, and P. H. N. de With. "Outcome prediction of Gamma Knife radiosurgery on vestibular schwannoma using contour-based shape descriptors". Presented at the 11th Biomedica Summit, May 9–10, 2017 , Eindhoven, The Netherlands.

[3] P. Langenhuizen and H. B. Verheul. "Invloed van groeisnelheid op de effectiviteit van Gamma Knife-behandeling van vestibulair schwannomen". In: *Tijdschrift voor Neurologie en Neurochirurgie*. Vol. 119. 2. Ariez BV, 2018, p. 67.

Acronyms

3D-HOG Three-Dimensional Histograms of Oriented Gradients

ADC Apparent Diffusion Coefficient

AUC Area Under the Curve

BED Biologically Effective Dose

CAD Computer-Aided Diagnosis

CT Computed Tomography

DVH Dose-Volume Histograms

DWI Diffusion-Weighted Imaging

FPR False Positive Rate

GKRS Gamma Knife RadioSurgery

GLCM Gray-Level Co-occurrence Matrix

GTV Gross Target Volume

HI Homogeneity Indices

HOG Histogram of Oriented Gradients

LINAC Linear Accelerator

LOOCV Leave-One-Out Cross Validation

MF Minkowski Functional

MLIN Multi-Landmark Intensity Normalization

mRECIST Modified Response Evaluation Criteria In Solid Tumors

MRI Magnetic Resonance Imaging

NF2 NeuroFibromatosis type 2

PCA Principal Component Analysis

PIV Planned Isodose Volume

ROC Receiver Operating Characteristic

RECIST Response Evaluation Criteria In Solid Tumors

ROI Region Of Interest

SAP Summed Absolute Pixel

SRA Summed Absolute Area

SVM Support Vector Machine

T1 T1-weighted

T1CE T1-weighted, Contrast-Enhanced

T2 T2-weighted

TNR True Negative Rate

TPR True Positive Rate

TTE Transient Tumor Enlargement

VDT Volume Doubling Time

VHT Volume Halving Time

VS Vestibular Schwannoma

Acknowledgements

After such an intense and wonderful period with its ups and downs, it is nice to look back at all the things that have happened. And a lot has happened. I remember quite vividly that, after studying for a little over 10 years, I was unsure what the future would hold for me and what I wanted to do. That was until I was asked to do a PhD. I became interested in the medical field through my graduation project, and decided that a PhD in that specific domain really appealed to me. I wanted to contribute to this field, and the vestibular schwannoma project sounded really interesting. In my first year, I mainly focused on the creation of the unique database that we now have. This 'monnikenwerk', as my second co-promotor called it, was not very inspiring, but it has helped me to really getting to know the data and the database. It formed the basis of all the work that is presented in this thesis (and some more) and of which I am really proud. We have published several articles and have presented data and results at various conferences across the world, with the most exotic location being Rio de Janeiro, Brasil, where I gave three different oral presentations. In the final summary of that conference, the ISRS 2019, one of these presentations was highlighted by the keynote speaker. This of course made me very proud of my work. At the last conference I attended, the eight quadrennial conference on vestibular schwannoma and other CPA tumors at the Mayo Clinics in Rochester MN, USA, the presentation I gave was again addressed by others during a rather heated discussion between a speaker and a neurosurgeon in the audience. This really boosted my confidence.

So, looking back, I have grown tremendously during the last five years, both professionally as well as personally. I have been to great conferences and met interesting people. I have learned both medical and technical aspects of this project and how to create a bridge between these two very different worlds. I have witnessed and participated in very fascinating discussions and I have shared in the creation of a follow-up project, which recently started. I am really proud of all the achievements made during this period, including the new collaborations with international medical centers. All of this would not have been possible without the help of others.

First of all, I would like to thank my promotor, prof. dr. ir. Peter de With. Although I successfully executed an internship on salient object detection and a master graduation project on optical coherence tomography, in cooperation with AMC hospital, in the SPS-VCA group of Peter, I still had doubts about my future in this field of work. However, Peter saw my potential and asked me to consider a PhD position in his group. After some consideration, I realized that this was a great opportunity and said yes. Looking back at those moments, I am now

wondering why I was so hesitant about this. Peter, thank you for the wonderful experience you have introduced me to and for seeing my potential. Your trust in me and your always positive energy have been a significant contribution to the success of this thesis. Our meetings always energized me and motivated me to investigate all facets of this project.

Next, I would like to thank my first co-promotor dr. Sveta Zinger. We met during my first internship at the VCA group, and I have enjoyed working with you since then. Your guidance during the last five years has helped me through some tough moments and inspired me to continue discussing the various aspects of this work with the medical experts and to always come to a consensus with them. I have enjoyed our frequent (and sometimes infrequent) meetings, in which I could discuss the medical and technical aspects of my work (and express the inherent frustrations that occurred alongside). Thank you for your patience, guidance and positive energy.

I would also like to express my gratitude to my second co-promotor dr. Jeroen Verheul. With my singular technical background, it was quite a challenge to adapt to the medical setting in which this project took place. My first few days at your side have been very inspiring. I learned a lot during the outpatient visits, where I had the opportunity to meet several patients suffering from a vestibular schwannoma. It really boosted the affinity I had with this project. Jeroen, your enthusiasm concerning this project and everything related to the disease has really motivated me and it provided a basis for our strong relationship. This has led to numerous wonderful results, including several presentations at various international conferences, and we are now being recognized for the work we did by the neurosurgical community. This in turn has led to collaborations with other hospitals in a future project. I will never forget the caipirinhas we had during the ISRS 2019 in Rio de Janeiro. Those really made an impact. Thank you for the wonderful time and inspiring discussions we have had, and hopefully we can continue our scientific successes in the coming years during the TZO project. You have been an important driving force during my time as a PhD candidate and I am truly grateful for everything you have done.

Not only Jeroen has provided me with a good time at the Gamma Knife center in Tilburg. The other neurosurgeons (Sieger Leenstra, Bram van der Pol, Guus Beute, Liselotte Lamers, and Suan Te Lie), both radiotherapists (Patrick Hanssens and Diana Grootenboers), both nurses (Marion en Reneé), both clinical physicists (Wim de Jong and Jannie Schasfoort), the secretary (Anja), and the various radiographers have helped me feel welcome and part of the team. I could always count on their support and I have enjoyed their professionalism in their work as well as our personal conversations. All neurosurgeons that work at the Gamma Knife center have provided me with motivation and support (both in professional as well as in personal situations), and were always available for discussing the various ideas, experiments and results. I would like to thank Sieger Leenstra for his input during our bimonthly meetings and his expertise in tumor biology. Special thanks go to Anja for providing numerous MRI scans and follow-up data during

the start of the project. Without her, it would have taken me so much more time to compile the database. Furthermore, Patrick Hanssens has provided me with additional clinical input for my publications, for which I am grateful. I will never forget the epic bike tours he organizes each year for the people working at the Gamma Knife center, including the impressive lunch, dinner, and Belgium beers. Finally, I would also like to thank Jannie and Suan Te for the coffee during the always "tough" Monday mornings and the sushi dinners we enjoyed. These will always be remembered.

I would also like to thank my colleagues at the VCA group. Although I worked there part-time, I have enjoyed the talks and the atmosphere at the office, as well as their company during conferences. Their insights and remarks have helped me grow scientifically as well as personally. I would also like to express my gratitude to all students that have contributed to this work: Yan, Thom, Mark, Hans, and Sander. All your hard work has led to various publications that are incorporated in this thesis, for which I am grateful. Furthermore, special thanks go to Anja, who has helped me throughout the project with all administrative tasks and the organization of our bimonthly meetings with the clinical experts.

During my time as a PhD, I learned that it is good to relax and wind down from time to time. Taking up too much stress is never good, and you should find a way to let this all out. For these moments, I would like to thank my dear friends, who have stood by me during the last five years. Tom and Esther, Merijn and Alette, and Tim and Sabine, thank you for all the moral support you have given me. I am grateful that you all believed in me.

Not only friends are important in life. Family perhaps even more so. That is why I want to let Jeroen and Karin, Etienne and Rozan, and Moniek know that I truly appreciate what you have done for me. I could always go to you for moral and emotional support during the times it got tough. I cannot express how grateful I am for having you close by and I will never forget what you have done for me.

Finally, and by far the most important person in my life, I would like to thank my wife Inge for always being there for me and for her unconditional love. You have been an inspiration throughout my PhD, and I could always count on you to be there for me. You have helped me through the tough moments and you have motivated me to investigate the ideas that popped in my head. You have listened to me ranting about science stuff and have helped me with any medical-related questions. I will never forget my first days at the ETZ. I came home to you with a million questions, which you all answered in-depth, including and not limited to images, books, papers and websites. I definitely could not have done this without you, and I am proud to have you by my side!

Curriculum vitae

Patrick Langenhuizen obtained his BSc. degree in Electrical Engineering in 2015 from Eindhoven University of Technology. In that same year he also received his MSc. degree in Electrical Engineering. During his master thesis, Patrick investigated the possibilities of employing optical coherence tomography for scanning the digestive tract in a new manner. This project was in collaboration with AMC hospital in Amsterdam. Prior to his graduation project, Patrick already worked in the Video Coding and Architectures group at Eindhoven University of Technology, on an ROI-detection algorithm during an internship.

After his master studies, Patrick started a PhD research project, working on image analysis methods for predicting the Gamma Knife treatment response of vestibular schwannomas at the ETZ hospital, Tilburg, in collaboration with the Eindhoven University of Technology. His work has enabled the collaboration with other European hospitals, in which the results on individual treatment prediction will be exploited for further investigations into creating an clinical decision support system.

After he finished the project at the ETZ hospital, he started as a post-doctorate researcher at the Video Coding and Architectures group at the Eindhoven University of Technology. Patrick has always been interested in science and specifically technological developments. In his spare time, he continues his passion for images by experimenting with photography.