Aus dem Zentrum für Neurochirurgie der Universität zu Köln Klinik und Poliklinik für
Funktionelle Neurochirurgie und Stereotaxie
Direktorin: Universitätsprofessorin Dr. med. V. Visser-Vandewalle

MR-Radiomics-based Prediction of Long-term Treatment Response of Vestibular Schwannoma after Stereotactic Radiosurgery

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Medizinischen Fakultät
der Universität zu Köln

vorgelegt von
Julius B. Garben
aus München

promoviert am 11. Juni 2026

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln

2026

Dekan: Universitätsprofessor Dr. med. G. R. Fink

1. Gutachter: Professor Dr. med. M. Kocher
2. Gutachter: Universitätsprofessor Dr. med. R. Goldbrunner

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertationsschrift ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich keine Unterstützungsleistungen.

Weitere Personen waren an der Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin/eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertationsschrift stehen.

Die Dissertationsschrift wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die dieser Arbeit zugrunde liegenden Bilddaten und Bestrahlungspläne wurden ohne meine Mitarbeit in der Klinik für Stereotaxie und Funktionelle Neurochirurgie des Zentrums für Neurochirurgie Priv.-Doz. Dr. Daniel Rueß zur Verfügung gestellt. Die Volumetrie Daten wurden teilweise durch mich aus den zur Verfügung gestellten Bilddaten erhoben, und teilweise zur Verfügung gestellt.

Die in dieser Arbeit angegebene Methoden zur Datenverarbeitung und Datenanalyse einschließlich Bildkonversion, Merkmalsextraktion, Erstellen eines Vorhersagemodells durch maschinelles Lernen und Evaluierung des Modells sind nach entsprechender Anleitung durch Herrn Prof. Dr. Martin Kocher von mir selbst ausgeführt worden.

Erklärung zur guten wissenschaftlichen Praxis:

Ich erkläre hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten (Amtliche Mitteilung der Universität zu Köln AM 132/2020) der Universität zu Köln gelesen habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen.

Köln, den 04.02.2026

Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the completion of this work.

First and foremost, I would like to thank my supervisor, Professor Martin Kocher, for his exceptional guidance, the many insightful discussions, constructive feedback, and time invested in reviewing and improving this work. His expertise and encouragement were invaluable at every stage of this work.

I am also grateful to PD Dr. Daniel Rueß for defining the clinical categories on which this work is based and for providing the clinical data used in this study as well as the great support and insight. His foundational work made this analysis possible.

I would like to thank the Department of Stereotactic and Functional Neurosurgery of the University Hospital of Cologne for providing the imaging, the workstation the machine-learning was performed on, and the stimulating academic environment necessary to conduct this research.

I am deeply grateful to the open-source community for creating and providing countless mature and established libraries like scikit-learn or pyradiomics among many others. This and many other studies would not be possible without.

I would like to thank Julia-Sarita Brand for her work on prediction modelling regarding brain metastasis, which provided valuable context and orientation for this study and Ricardo Loucao for his support in maintaining the research workstation.

Special thanks goes to Dr. Benedikt Thelen Schäfer for many discussions, critical feedback, and for his careful review of this work.

Finally, I would like to thank my family - especially my mother - and friends for their encouragement, and unwavering support throughout this journey.

to the developers of Python, without whom this work would not exist

Contents

ABBREVIATIONS 10

1. SUMMARY (GERMAN) 11

2. INTRODUCTION 13

2.1. Vestibular Schwannoma 14

2.1.1. Histology 14

2.1.2. Epidemiology 15

2.1.3. Symptoms 15

2.1.4. Classification 16

2.1.5. Imaging 17

2.1.6. Treatment Approaches 17

2.1.7. Stereotactic Radiosurgery (SRS) 18

2.1.8. Radiobiology of SRS 19

2.1.9. Treatment Response and side effects after SRS for VS 20

2.1.10. Changes in the Therapeutic Approach 21

2.2. Introduction into Radiomics 21

2.2.1. Radiomics Features and Biology of VS 22

2.2.2. PyRadiomics and the Standardization of Radiomics Features 23

2.3. Introduction into Machine Learning 24

2.3.1. Machine Learning in Medicine 24

2.3.2. Choice of Model Type, Overfitting and Underfitting 25

2.3.3. Logistic Regression 26

2.3.4. Random Forest 29

2.4. Objectives and Research Questions 31

3. MATERIAL AND METHODS 32

3.1. Ethics Statement 32

3.2. Inclusion Criteria 32

3.3. Stereotactic Radiosurgery 32

3.4. Follow-up 33

3.5. Imaging Endpoints 33

3.6. Software 34

3.7. Image Preprocessing 35

3.8. Segmentation 36

3.9. MR-Radiomics Feature Extraction 36

3.10. Machine Learning Workflow 39

3.10.1. Feature Value Normalization 41

3.10.2. Feature Selection Based on Redundancy 41

3.10.3. Train/Test-Set Split for Initiation of Nested Cross-Validation 42

3.10.4. Feature Selection Based on Correlation with Clinical Data 43

3.10.5. Feature Selection Based on Regularization of Logistic Regression 44

3.10.6. Feature Selection with Recursive Feature Elimination and Cross-Validation (RFECV) 44

3.10.7. Hyperparameter Optimization 44

3.10.8. Model Training for Cross-Validation 45

3.10.9. Evaluation 45

4. RESULTS 47

4.1. Patient Characteristics 47

4.2. Progressive Disease (PD) 49

4.3. Pseudoprogression (PP) 52

4.4. Local Control (LC) 54

4.5. Multiclass Classification 56

5. DISCUSSION 58

5.1. Progressive Disease (PD) 58

5.2. Pseudoprogression (PP) 59

5.3. Local Control (LC) 60

5.4. Three-class Model (3-C) 61

5.5. Methodical Considerations 61

5.6. Comparison to Literature 62

5.7. Outlook 64

6. BIBLIOGRAPHY 66

7. APPENDIX 75

7.1. Supplementary Tables and Figures 75

7.2. List of Radiomics Features 76

7.2.1. Shape 76

7.2.2. Firstorder Statistics 76

7.2.3. GLCM 76

7.2.4. GLDM 76

7.2.5. GLRLM 76

7.2.6. GLSZM 77

7.2.7. NGTDM 77

7.3. Code blocks **77**

7.4. Figures **85**

7.5. Tables **87**

Abbreviations

2D	Two Dimensional
3-C	Three-Class Response
3D	Three Dimensional
AUC	Area Under Curve
CE	Contrast Enhancement
CR	Complete Response
CV	Cross-Validation
DICOM	Digital Imaging and Communications in Medicine
IQR	Inter Quartile Ratio
KW	Kruskal-Wallis-test
LC	Local Control
ML	Machine Learning
MRI	Magnet Resonance Imaging/Image
MS	Microsurgery
MWU	Mann-Whitney-U-test
NIFTI	Neuroimaging Informatics Technology Initiative
NPV	Negative Predictive Value
OvR	One-versus-Rest
PD	Progressive Disease
PP	Pseudoprogression
PPV	Positive Predictive Value
PR	Partial Response
PRC	Proposed Response Criteria
PTV	Planning Target Volume
RFECV	Recursive Feature Elimination and Cross-Validation
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RTSS	Radiotherapy Structure Sets
SD	Stable Disease
SRS	Stereotactic Radiosurgery
VOI	Volume of Interest
VS	Vestibular Schwannoma

1. Summary (German)

Vestibularis-Schwannome (VS) sind gutartige Tumoren des achten Hirnnervs,[1] bei deren Behandlung sich das Therapieziel sich in den letzten Jahrzehnten zunehmend von der vollständigen Tumorresektion hin zum Funktionserhalt verschoben hat.[2] Die stereotaktische Radiochirurgie (SRS) stellt heute insbesondere für kleine und mittelgroße Tumoren eine etablierte Behandlungsoption mit gutem Therapieansprechen und vergleichsweise günstigem periinterventionellen Risikoprofil dar.[3] Dennoch zeigen einzelne Tumoren nach SRS einen unerwünschten Verlauf in Form einer Progression, welche sich derzeit prospektiv nicht zuverlässig von einer Pseudoprogression, einer transienten Volumenzunahme ohne Zellvermehrung, abgrenzen lässt.[4] Eine prädiktive Einschätzung des Therapieansprechens vor Behandlungsbeginn existiert bislang nicht.[3,5,6]

Ziel dieser Arbeit war es, zu untersuchen, ob sich das langfristige volumetrische Therapieansprechen von VS nach SRS mithilfe von MR-Radiomics-basierten Machine-Learning-Modellen anhand prätherapeutischer Magnetresonanz-Tomogramme (MRT) vorhersagen lässt. Grundlage hierfür bildet die Radiomics-Hypothese, nach der sich histologische und biologische Tumoreigenschaften in der Bildtextur widerspiegeln, die quantitativ erfassbar sind.[7–9]

In dieser retrospektiven Single-Center-Studie wurden 122 Patientinnen und Patienten mit VS eingeschlossen, die zwischen 2012 und 2022 mittels SRS (CyberKnife®) an der Uniklinik Köln therapiert wurden. Aus prätherapeutischen MRT-Sequenzen wurden standardisierte Radiomics-Merkmale gemäß den Richtlinien der *Image Biomarker Standardization Initiative*[10] extrahiert. Zur Vorhersage der volumetrischen Verlaufsformen orientierten wir uns an den sechs Kategorien der vorgeschlagenen response-Kriterien von Rueß et. al.[4] und fassten diese, nach klinischer Implikation gruppiert, in drei Gruppen zusammen: Progressive Disease (PD), Pseudoprogression (PP) und Local Control (LC) bei Vorliegen von Stable Disease, Partial Remission oder Complete Remission. Es wurden logistische Regressionsmodelle[11,12] und Random-Forest-Klassifikatoren[13,14] verwendet. Die Modellierung erfolgte unter Verwendung von nested cross-validation[15] einschließlich Feature-Selektion und Hyperparameter-optimierung, um eine möglichst realistische Abschätzung der Generalisierbarkeit zu erreichen.

Die besten Modelle erzielten für die Vorhersage einer progressiven Erkrankung (PD) eine gute prädiktive Leistung mit einer mittleren ROC-AUC von bis zu 0.83. Die Vorhersage der Pseudoprogression (PP) erreichte eine moderate Trennschärfe (ROC-AUC bis 0.69), und die der LC eine ROC-AUC von 0.64. Besonders häufig selektierte radiomics features waren

statistische Merkmale der Graustufenverteilung, insbesondere auch aus wavelet-transformierten Sequenzen.

Zusammenfassend zeigen die Ergebnisse, dass MR-Radiomics-basierte Machine-Learning-Modelle grundsätzlich in der Lage sind, das volumetrische Therapieansprechen von VS nach SRS mit moderater bis guter Genauigkeit vorherzusagen, insbesondere im Hinblick auf eine Tumorprogression. Auch wenn die Vorhersageleistung derzeit noch nicht für eine unmittelbare klinische Anwendung ausreicht, unterstützen die Ergebnisse die Radiomics-Hypothese und legen nahe, dass mit größeren, multizentrischen Datensätzen und der Einbeziehung von zusätzlichen *non-radiomics Feature*s ein klinisch anwendbares Entscheidungshilfesystem entwickelt werden könnte.

2. Introduction

Vestibular schwannomas (VS) are benign, slow-growing tumors arising from the Schwann cells of the vestibular part of the eighth cranial nerve.[1] They represent one of the most common benign intracranial tumors and typically present with unilateral hearing loss, tinnitus, and vestibular symptoms.[16,17] Owing to advances in magnetic resonance imaging (MRI), vestibular schwannomas are increasingly diagnosed at earlier stages and smaller sizes,[18] leading to a shift in therapeutic strategies over recent decades.[2]

Management options for vestibular schwannoma include observation, microsurgical resection, stereotactic radiosurgery (SRS), and, in neurofibromatosis type 2 (NF2), systemic therapy.[19] Among these, SRS has become a widely accepted treatment modality for small to medium-sized tumors, offering high rates of long-term tumor control while minimizing treatment-related morbidity.[3] As a result, treatment paradigms have progressively shifted from radical tumor removal toward functional preservation, particularly of hearing and facial nerve function.[2]

Despite favorable overall outcomes, tumor response after SRS is heterogeneous. While most tumors remain stable or regress over time, a subset demonstrates volumetric progression requiring salvage treatment. In addition, transient tumor enlargement or pseudoprogression (PP) occurs in a considerable proportion of cases and cannot be reliably distinguished from true progressive disease (PD) using conventional imaging criteria alone. This diagnostic uncertainty may lead to unnecessary interventions or prolonged uncertainty for patients and clinicians alike.[4]

Currently, treatment response after SRS can only be assessed retrospectively by observing the temporal evolution of tumor volume. No established method exists to predict individual tumor behavior prior to treatment. A reliable pretherapeutic prediction of volumetric response could therefore substantially improve clinical decision-making, enabling a more personalized selection of treatment strategies and more informed patient counseling.[3,5,6]

Radiomics has emerged as a promising approach to extract high-dimensional quantitative information from medical images.[10,20] By converting imaging data into large sets of numerical features describing tumor shape, intensity distribution, and texture, radiomics aims to capture information with are plausibly related to underlying tumor biology, such as cellularity, vascularization, necrosis, and microstructural heterogeneity.[7–9] Combined with machine learning techniques, radiomics offers the potential to identify complex patterns associated with clinical outcomes and treatment response.[6,21,22]

Evidence for radiomics-based prediction of long-term response of VS to SRS is limited, and existing studies differ substantially in cohort size, response definitions, feature standardization, and methodological rigor. In particular, the distinction between progressive disease and pseudoprogression has often been insufficiently addressed, despite its high clinical relevance.[6,21,22]

The present study aims to examine whether MR-based radiomics features derived from pretherapeutic planning MRI can be used to predict the long-term volumetric response of vestibular schwannomas after stereotactic radiosurgery. By applying standardized radiomics feature extraction and rigorously validated machine learning workflows, this work seeks to evaluate the feasibility and limitations of radiomics-based outcome prediction in this clinical context and to contribute to the development of personalized treatment strategies for patients with vestibular schwannoma.

2.1. Vestibular Schwannoma

2.1.1. Histology

Vestibular schwannomas (VS) are benign nerve sheath tumors arising from the Schwann cells of the vestibular part of the eighth cranial nerve. The tumors are well demarcated, loosely attached to the nerve, and do not show invasive growth, which allows resection with preservation of function. Accordingly, the pathogenesis of symptoms is not due to direct invasion to the nerve, but to compression of the nerve, other cranial nerves, and surrounding brain tissue. The term acoustic neuroma is rarely used anymore, as VS is neither a neuroma nor does the tumor originate from the cochlear part of the nerve.[23,1]

As in other schwannomas, cell-rich Antoni A regions and cell-poor Antoni B regions can be identified microscopically in VS (see Figure 1). Antoni A regions are cell-rich tumor parts with closely arranged nuclei in so-called palisades. Antoni B regions are characterized by a few scattered cells with interspersed microcysts filled with basophilic mucin. Antoni B regions are interpreted as degenerated Antoni A regions. Common degenerative features are nuclear pleomorphism, xanthomatous changes, vascular hyalinization, cystic changes, and necrosis. The hyalinized vessels found in the Antoni B regions are interpreted as signs of hypoxia.[1,19,24]

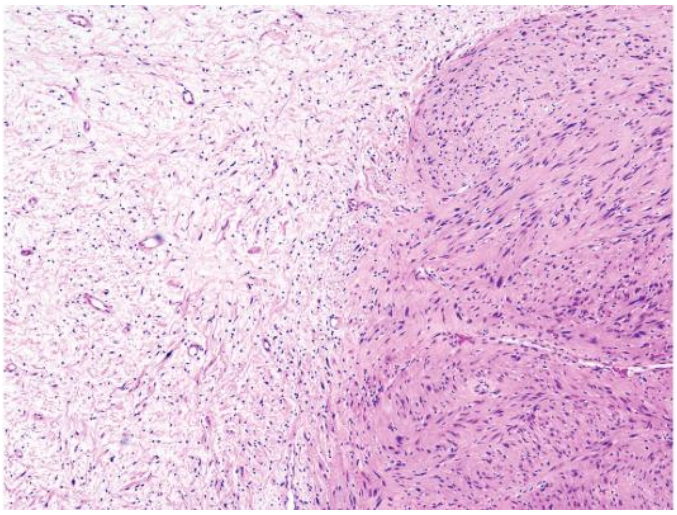

Figure 1: Microscopic appearance of a vestibular schwannoma. A cell-rich Antoni A region is present on the right side, while a comparatively loosened, cell-poor Antoni B region is visible on the left side (Hematoxylin-Eosin, x400)[24]

2.1.2. Epidemiology

Recent studies indicate that the incidence of VS is increasing globally, primarily due to improved diagnostic imaging techniques such as MRI. A systematic review reported incidence rates ranging from 3.0 to 5.2 per 100,000 person-years, with the highest rates observed in individuals aged 70 years and older.[16] In Denmark, a 40-year-long prospective study demonstrated an increase in annual incidence from approximately 3 per million in 1976 to 34 per million in 2015, alongside a decrease in tumor size at diagnosis from 26 mm to 7 mm and an increase in the median age at diagnosis from 49 to 60 years. These trends suggest that the rising incidence is rather attributable to the availability and quality of MRI than to an true increase in incidence.[18]

2.1.3. Symptoms

The most common symptom of VS is unilateral hypoacusis, which has often been present for several years at the time of diagnosis. In addition, tinnitus and vestibular symptoms such as dizziness and gait disorders are also common. Less frequent symptoms are facial hypesthesia, facial paresis, diplopia and symptoms due to caudal cranial nerve deficits.[17] The typical frequency and duration of the symptoms at diagnosis are displayed in Table 1.

Cranial Nerve	Symptom	Incidence	Duration at Diagnosis
V	Trigeminal nerve disturbances	7-9%	1.3 y

VI	Double vision	1.8%	1.4 y
VII	Facial paresis	5.5%	1.9 y
VIII	Hypoacusis	95%	3.7 y
VIII	Vestibular disturbances	61%	2.1 y
IX-XII	Caudal Cranial nerve deficits	2.7%	2.7 y

Table 1: Incidence and duration of symptoms of VS at time of diagnosis [17]

2.1.4. Classification

The classification is usually made by accounting for size and location of the tumor and is based on the Koos grading scheme (see Figure 2). Grade I describes an exclusively intrameatal lesion in which there is usually only a mild hypoacusis of less than 50 dB. Grade II applies to tumors with protrusion into the cerebellopontine angle without contact to the brain stem. Grade III comprises tumors that fill the cerebellopontine cistern but do not compress the brainstem. Grade IV tumors are large tumors with associated brainstem compression.[25]

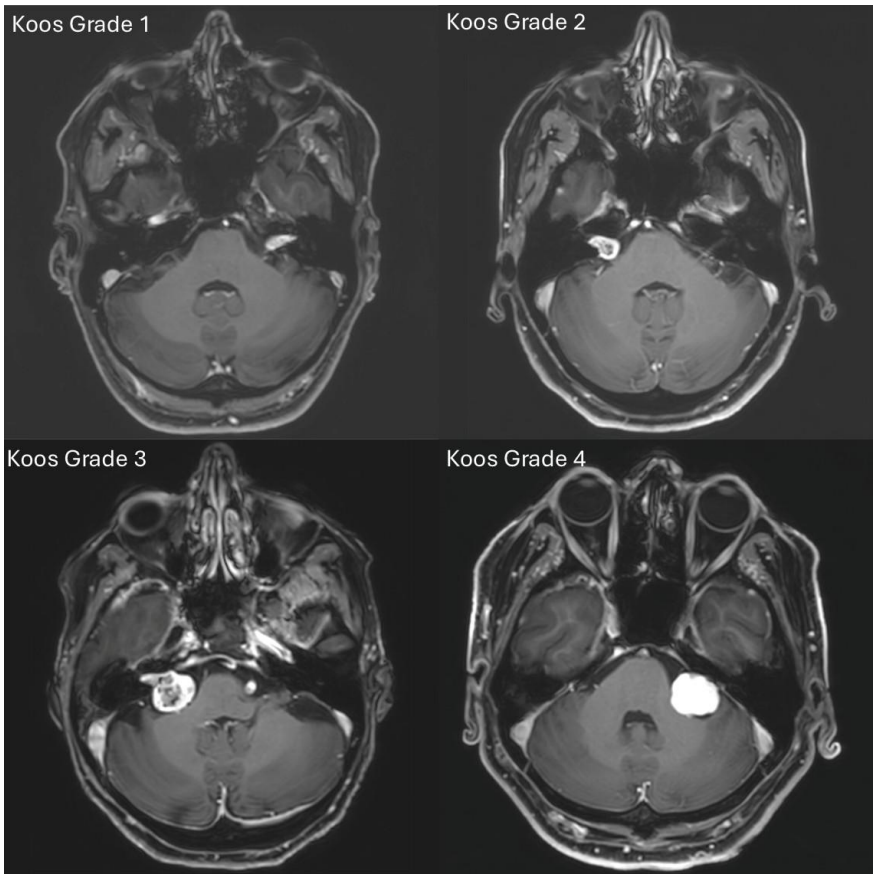

Figure 2: Contrast enhanced T1-weighted MRI depicting the classification of VS as described by Koos[25]

2.1.5. Imaging

As with most intracranial tumors, T1-weighted, contrast-enhanced magnetic resonance imaging (MRI) is the imaging modality of choice for VS. VS typically consist of a solid nodular mass with an intrameatal component that often leads to dilation of the internal acoustic porus.[26]

On MRI, VS appear isointense or hypointense on native T1-weighted images and hyperintense on T2-weighted images, and present with pronounced homogeneous or heterogeneous enhancement in the contrast-enhanced T1-weighted MR images.[27] The histological characteristics, as well as cell proliferation and microvessel density, also influence the magnetic resonance imaging texture of the tumor. The same applies to macrocystic changes or hemorrhages.[28–31]

2.1.6. Treatment Approaches

Besides stereotactic radiosurgery (SRS) which will be discussed in detail below, observation, microsurgery (MS) and systemic treatment are available. VS have average growth rates of approximately 1 mm increase in diameter per year. About half of all cases remain stable during the first five years after diagnosis,[32,33] so an immediate therapeutic intervention is not necessarily required in all cases. For small VS, a scan-and-wait strategy is a valid approach, especially for small (<2cm), intrameatal VS without tinnitus.[34] In opposite, a study including medium sized tumors with a relevant extrameatal tumor extension of up to 2cm suggested inferiority of the scan-and-wait strategy compared to upfront SRS.[35]

Microsurgical resection remains a cornerstone in the management of vestibular schwannomas, particularly for patients with large tumors, significant brainstem compression, or progressive symptoms. The choice of surgical approach – retrosigmoid , translabyrinthine, or middle cranial fossa – is tailored based on tumor size, location, and the patient's preoperative hearing status. The retrosigmoid approach offers the potential for hearing preservation and is suitable for tumors extending into the cerebellopontine angle. The translabyrinthine approach provides direct access to the internal auditory canal but results in complete hearing loss on the affected side. The middle cranial fossa approach is typically reserved for small tumors confined to the internal auditory canal in patients with good hearing.[36] Advancements in intraoperative neurophysiological monitoring, particularly of the facial nerve, have significantly improved postoperative outcomes.[37] In Koos grade 4 VS, subtotal resection followed by adjuvant radiosurgery may help balance tumor control with functional preservation.[38]

Bevacizumab is a monoclonal antibody binding to the vascular endothelial growth factor (VEGF). It hereby decreases angio-neogenesis and subsequently decreases tumor growth.[39]

Its use is primarily reserved for patients with neurofibromatosis type 2 (NF2)-associated VS in whom it has demonstrated efficacy in reducing tumor volume and improving hearing outcomes.[3,40]

2.1.7. Stereotactic Radiosurgery (SRS)

While microsurgical resection involves the surgical removal of tumor tissue, SRS aims to arrest tumor growth by irradiation. The method of focusing a single dose of high energy radiation on an intracranial target with a high degree of spatial precision was first described by Leksell in 1951,[41] who developed the first version of the Gamma-Knife^{R}. The gamma radiation originates from the decay of cobalt-60 emitting 1.17 MeV and 1.33 MeV photons.[42] Newer versions of the Gamma-Knife^{R} using the same principle and radiation sources are still in practice today. An alternative approach is to use a high energy (e.g. 6MV[i]) focused, collimated X-ray beam from a linear accelerator (LINAC) with multiple beams directed to the same small target. These specially equipped LINACs became available in the 1980s.[43]

While both methods are frame based, i.e. requiring the often painful invasive fixation of the skull in a stereotactic frame, since the development of the Cyberknife^{R} by Adler starting in 1989 and its commercial introduction in 2001,[44] the patient's head can be held in place by an individually fitted thermoplastic aquaplast mask. Stereotactic guidance is provided via intraprocedural X-ray imaging, and the therapeutic radiation dose is administered by a LINAC mounted on a robotic arm. Overall, the Cyberknife^{R}, the Gamma-Knife^{R} and the modified LINACs differ only slightly in terms of their clinical effectiveness.[45,46]

In Figure 3, the course of a VS treated with SRS is showcased. All images are contrast enhanced T1 MRI. a) is the image acquired prior to SRS, b) displays the segmented VS defining the target volume for the irradiation, c) displays the isodoses of the treatment plan, d) is the follow-up MRI after 6 months and e) another follow-up after 20 months.

[i] It is megavolt (MV) not mega electron volt (MeV) since the LINAC emits a spectrum of photons of up to 6 MV. MeV is conventionally used to describe discrete energies as in radioactive decay.

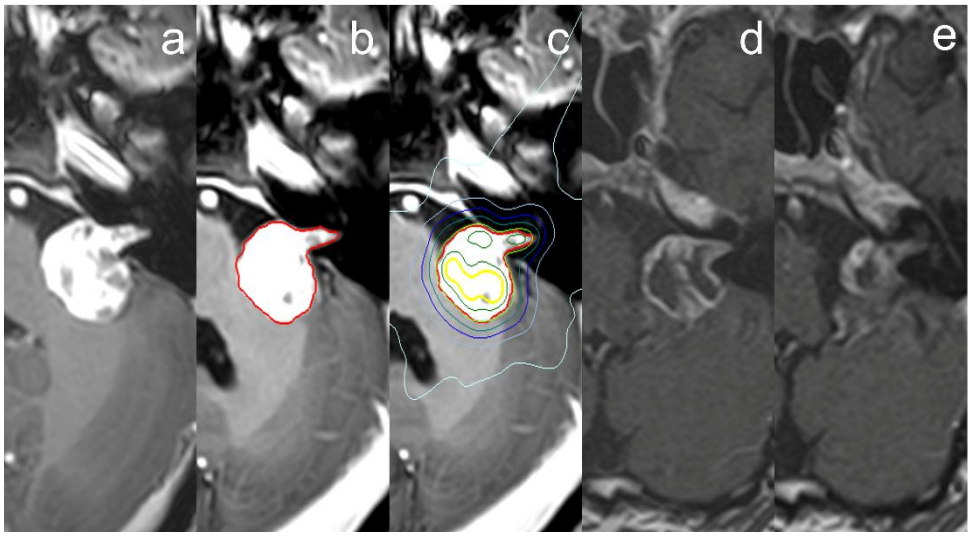

Figure 3: Contrast enhanced T1 MRI of a VS treated with SRS. a) MR image for treatment planning, b) manual segmentation, c) radiation planning with isodoses displayed, d) 6 month follow up MRI, e) 20 month follow up MRI

It is noticeable that the tumor presented in Figure 3 decreased in size and contrast affinity over time. The latter may also be explained by the lower amount of contrast agent used for following up imaging as opposed to imaging for treatment planning. Nevertheless, also non-contrast-enhancing regions are present in d) and e) indicating possible necrotic regions. Overall, this indicates a good treatment response with sufficient tumor control. The volume reduction over the course of that very case is displayed in Figure 4.

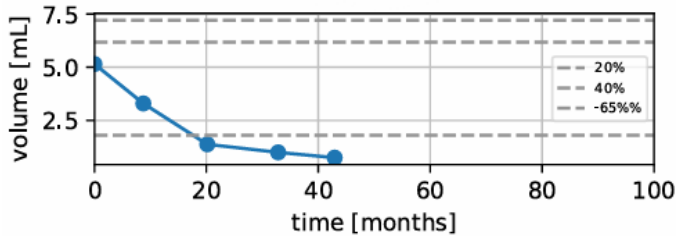

Figure 4: Development of the VS volume after SRS in the case of Fig. 3

2.1.8. Radiobiology of SRS

Radiobiology describes the biological effect of ionizing radiation on both target and off-target tissue. In contrast to radiation with particles like protons with high linear energy transfer (LET) that causes direct DNA damage in the cells of the irradiated tissue, in SRS, like in all radiation methods using high energy photons, the radiation effect is mainly caused by secondary electrons released in the tissue, and their DNA damaging properties. By interaction of these secondary electrons with oxygen or water, free radicals and other reactive oxygen species (ROS) are created. These highly reactive molecules with unpaired valence electrons chemically react with DNA, causing single or double-strand breaks. This triggers a DNA

damage response, which either leads to cell cycle arrest and DNA repair, apoptosis, or senescence, i.e. entering an irreversible G1 or G2 state. In addition, double-strand breaks in particular lead to cell death at the next mitotic attempt of the cell, although an accumulation of single-strand breaks is also potentially lethal. Interestingly, since the creation of ROS is partially dependent on the presence of oxygen, hypoxic tissues – commonly found especially in malignant tumors – are less susceptible to radiation.[46,47]

Withers introduced four main radiobiological mechanisms, also called the "four Rs" of radiobiology, namely repair of sublethal damage, reoxygenation of the tumor, redistribution of tumor cells throughout the cell cycle and regeneration, in order to explain the effect of different irradiation regiments.[48] Since then, Steel added radiosensitivity of the irradiated tissue as another factor,[49] Boustani supplemented the reactivation of tumor response[50] and Taghizadeh-Hesary the reinforcement by the tumor microenvironment.[51] An important idea is, that through different radiotherapeutic regimens with different dosage and fractionation, these different radiobiological processes can be controlled to some extent.[46]

For a single dose treatment as in SRS, mostly repair and radiosensitivity are relevant factors. Another important mechanism is likely the apoptosis of vascular endothelial tissue in microvessels as its sensitivity to doses >10 Gy/fraction has been demonstrated in experimental tumors resulting in secondary hypoxia in the target tissue.[52,53] Additionally, Boustani proposed that DNA damage as well as DNA release to the cytosol activates an innate immune response that could exhibit an sustained anti-tumor effect.[50,54] In contrast, fractionated radiotherapy might rather have only a transient immunosuppressive effect.[55]

2.1.9. Treatment Response and side effects after SRS for VS

The main aspect of the response of VS to SRS is the course of the tumor volume. In addition to remission (decrease in volume) and progression (increase in volume), RS of VS often leads to so-called pseudoprogression (PP),[4,56] which has also been described for other tumors and is characterized by a transient and spontaneous increase in tumor volume or contrast media affinity. Currently, in VS, differentiation from true tumor progression is only possible from the observation of the time course of the tumor.[4,56]

Referencing to the existing RANO criteria for the volumetric response of meningioma after SRS,[57] Rueß et al.[4] proposed an analog classification for VS. The classification comprises pseudoprogression (>20% volume increase, then decrease within 24-36 months) which can be subdivided into early PP (volume peak <12 months after SRS) and late PP (volume peak >12 months after SRS), complete response (complete cessation of the tumor), partial response

(volumetric decrease of >= 65%), stable disease (<65% decrease, <20% increase) and progressive disease (>40% increase or change in Koos grade from I or II to III or IV).

Another aspect is the functional or clinical course, particularly regarding hearing and symptoms of dizziness. In a recent meta-analysis, long-term hearing preservation is observed in 67% of the cases after a median follow-up period of 6.7 years. Young age, tinnitus rather than hearing loss as the first symptom of VS and good pre-therapeutic hearing ability indicated a positive prognosis. New onset or worsening facial nerve palsy (House-Brackmann > II) was found in 1.3% of cases. Trigeminal neuropathy leading to facial hypesthesia occurred in 3.2% of cases. New onset or transiently increased tinnitus or dizziness were also observed,[58] but symptoms of dizziness also improved in a proportion of patients.

2.1.10. Changes in the Therapeutic Approach

The primary treatment goal for VS has shifted in recent decades from complete surgical tumor removal to functional preservation of the vestibulo-cochlear nerve and the adjacent cranial nerves.[2] Especially for small VS (Koos grade I–II) and medium VS (Koos grade III–IV, <3 cm), SRS is regarded as an alternative to microsurgery in terms of preserving facial and cochlear nerve function, as shown in a meta-analysis with a follow-up period of up to 2 years.[59] Also the rates of therapy-induced morbidity and mortality are lower in SRS,[59] and the hospital stay is shorter.[3]

The long-term preservation of serviceable hearing on the other hand, proves to be better in microsurgery (74.5%) than in SRS (18.1%).[60] The use of SRS as a treatment strategy for large VS is not recommended by current EANO guidelines.[3] However, there are studies showing safe and effective treatment of large VS without symptomatic brainstem compression by SRS.[5,61,62]

2.2. Introduction into Radiomics

Analogous to concepts such as genomics and proteomics, the term ‘radiomics’ has become established for the machine-learning supported extraction and evaluation of medical image features. The focus is particularly on oncological MRI. The field of radiomics and other automated analysis methods for medical imaging has made great progress in recent years.[63]

Methodologically, the image section to be examined – namely the volume of interest (VOI) or region of interest (ROI) – is first defined and segmented manually or automatically in the

respective images. In contrast to computer tomography (CT) which provides widely standardized image intensities by means of the Hounsfield scale, for MRI, intensity normalization is necessary, as MRI intensity values stem from an arbitrary linear scale. Methods such as Z-value intensity normalization can be used. The next step is feature extraction. Here, image properties are automatically computed and output as numerical values. In order to emphasize certain image properties, filters such as a wavelet filter can be applied beforehand[63]. From the potentially large number of features, a subset is chosen for further analysis, e.g. by ranking them according to their correlation with the endpoints under investigation; redundant and non-robust features are excluded. Eventually, an ML-model can be fitted and tested using the selected features. A graphic overview of the workflow is displayed in Figure 5.[64]

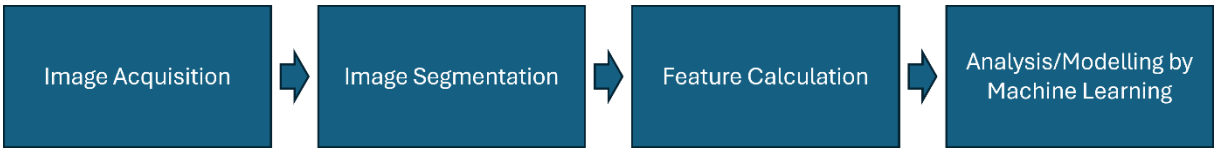

Figure 5: Radiomics workflow

2.2.1. Radiomics Features and Biology of VS

On visual inspection of MRI of different VS, it is noticeable that they come in various sizes and shapes and also exhibit different textures (see Figure 6).

The idea behind radiomics is that these properties of appearance in MRI are linked to differences in histology, e.g. the presence, distribution and composition of Antoni A and B areas, as well as vascularization, necrosis or (micro)cysts. These histologic properties are plausibly determined by genetics and linked to oxygenation, both influencing tumor growth, radiosensitivity and other factors relevant for the clinical course or the response to treatment.[7–9]

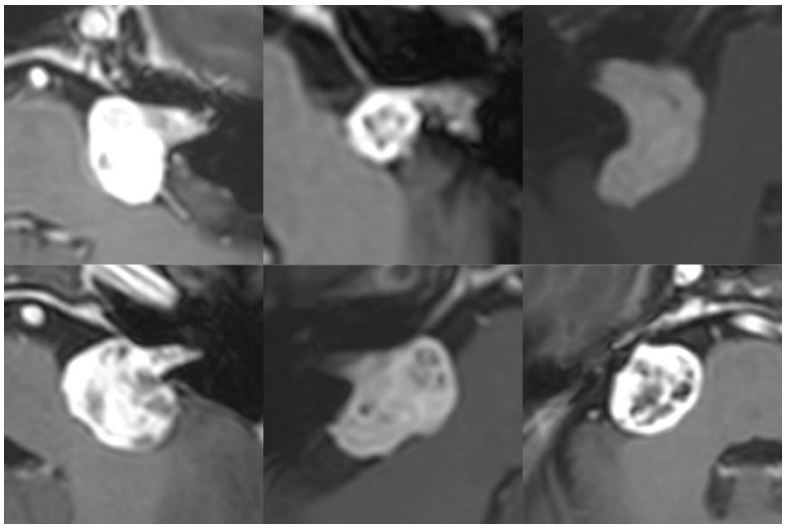

Figure 6: VS of different shapes, sizes and textures

The application of radiomics aims to exploit this connection in order to make diagnostic statements or predictions based on these image properties. To examine this connection in a systematic way, the size, shape and texture of the VS need to be described in an objective and reproducible manner. This is achieved by defining radiomics features derived from the tumor contour such as size and shape, from the numeric distribution of voxel signal intensities, and from the spatial distribution of intensity values representing textural features.[65–67]

2.2.2. PyRadiomics and the Standardization of Radiomics Features

Historically, different scientific groups were using their own, in-house developed feature sets, leading to difficulties in verification and reproduction. The Image Biomarker Standardization Initiative (IBSI) launched in 2016 assessed radiomics features of 25 research teams and aimed to create a consensus-based reference set of features to be used between radiomics software, and defined a standardized procedure for image processing and the calculation of features.[68]

PyRadiomics is an open-source IBSI-compliant platform, designed to make radiomics widely accessible, especially via its implementation in Python or software like 3D Slicer. It supports both 2D and 3D image data and offers a wide range of feature classes, including first-order statistics, shape descriptors, textural features as well as features calculated from wavelet-transformed MRI.[20]

2.3. Introduction into Machine Learning

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms capable of making predictions or decisions by learning from data without being explicitly programmed. In supervised learning, these algorithms learn to identify relationships between patterns in input datasets and predefined endpoints by exposing them to a set of input data together with the labeled endpoint. These models are able to improve their performance, i.e. learning, by adjusting a large set of internal variables when they are exposed to more data.[69–72]

A typical kind of task that is addressed using ML is classification. The aim is to assign new observations, which in clinical practice could be new cases, to the correct class, e.g. a correct diagnosis. While conventional diagnostics may use criteria that focus on a single variable or calculate scores based on several variables, ML tries to find an intricate mathematical model that identifies and uses key features from a large quantity of available features to achieve diagnostic or predictive conclusions. The idea is to utilize information that does not rely on single features but is hidden in linear or non-linear combinations of multiple, selected features.

In contrast to conventional statistical tests that analyze possible relations between independent variables and the endpoints of interest, ML models are mainly used to make predictions for single, new cases from the developed model.[73,74]

2.3.1. Machine Learning in Medicine

Machine learning has a long history in medicine. A famous early example is the computer based medical consultation system MYCIN described by E.H. Shortliffe in 1976.[75] The system draws conclusions based on decision rules:

$$P(h|e) = X$$

Meaning if *e* is known to be true, *h* is true with the probability of *X*. Based on a large set of predefined rules, clinical and contextual data, the system was supposed to provide information about the patients current clinical condition, the diagnosis and was supposed to provide recommendations about the best medical treatment.[75]

While early rule-based systems proved to be brittle, hard to maintain, and costly, the application of AI in medicine has come a long way since then, coming close to and occasionally surpassing human capabilities.[76] In 2016 Wang et al. trained a neural network to detect metastatic breast cancer in lymph node biopsies. The algorithm reached an accuracy of 0.925

coming close to the accuracy of 0.966 of a trained pathologist. Interestingly, there was no strong correlation between the errors of the human and the trained model. Combining both approaches resulted in an accuracy of 0.995. The authors concluded that the strength of AI in diagnostics might be an integrated approach combining human and non-human applications.[77]

Especially in radiology, where large volumes of data are generated in the form of medical images, AI has become increasingly important. It is used for image and lesion classification, detection of organs, regions, landmarks as well as objects or lesions, for segmentation of physiological or non-physiological conditions, image registration, content-based image retrieval (i.e., identification of similar cases from large databases) and image enhancement.[78]

2.3.2. Choice of Model Type, Overfitting and Underfitting

A problem that occurs when training AI models is that they perform very well on the training data, but fail to generalize, so they perform poorly when applied to new data. This occurs particularly in complex models with many degrees of freedom, which can perfectly fit the training data without learning a meaningful, generalizable relationship between the input and the output in previously unseen datasets. Underfitting, on the other hand, occurs when a too simple model is chosen to explain complex relationships in the data. Here the performance is bad on both the training and new data. The solution in case of underfitting is to choose a more complex model.[79]

Most models are prone to overfitting, and the main method to overcome overfitting is to choose a model that is simple enough and is still able to generalize to unseen datasets. In the case of the training scenario investigated here with a little more than a hundred patients and less than a hundred selected radiomics features, a neural network would very likely overfit. Therefore, two simpler models were applied. The first is *logistic regression*, a relatively simple linear model, that can be used in conjunction with regularization, a method that reduces overfitting while training the respective model. This is achieved by penalizing the coefficients, starting with those with the highest value. Thereby, the chance that the importance of single features is overestimated is reduced, and many of the smaller coefficients may be excluded from the model, leading to a lean mathematical model.[80] The second model applied here is based on decision trees and called *random forest*. It is a non-linear model that includes a measure of feature importance which again can be used to develop a sparse model with the ability to generalize to unseen data sets.[13,14]

2.3.3. Logistic Regression

This chapter summarizes key principles of logistic regression[80]. Logistic regression, first described by Berkson in 1944[11,12], is a statistical modeling technique used for binary and multivariate classification problems. It is built around the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

For a graphic representation see Figure 7.

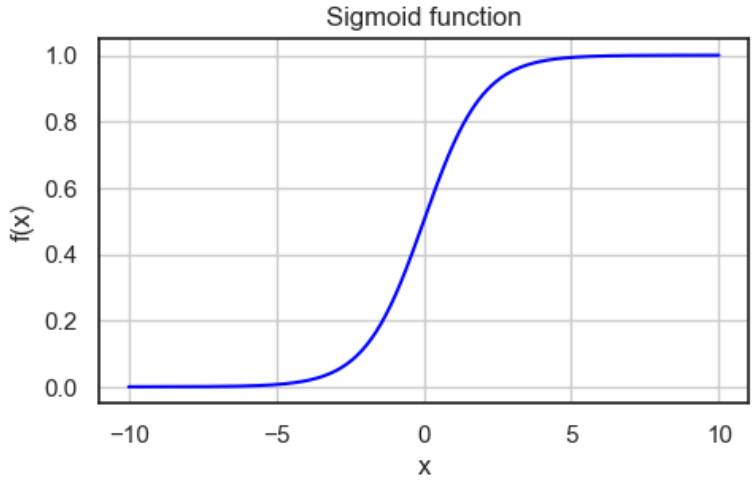

Figure 7: Plot of a sigmoid function

Binary logistic regression classifies observations into two categories, where the probability of occurrence of one of the classes is modelled. In the present study, e.g. the occurrence of pseudo-progression (PP) could be labeled category 1, opposed to category 0 for the non-occurrence of PP. The function used in logistic regression models the probability of a given observation y=p(x) to fall into category 1, given the value of the independent variable x. The formula can be written as

$$p(x) = \frac{1}{1 + e^{-(mx+t)}}$$

where *m* is the coefficient and *t* the intercept. *m* and *t* are to be optimized during model training to get the best prediction of y.

A logistic regression-based classifier not yet trained might look like in Figure 8, where m=1 and t=0.

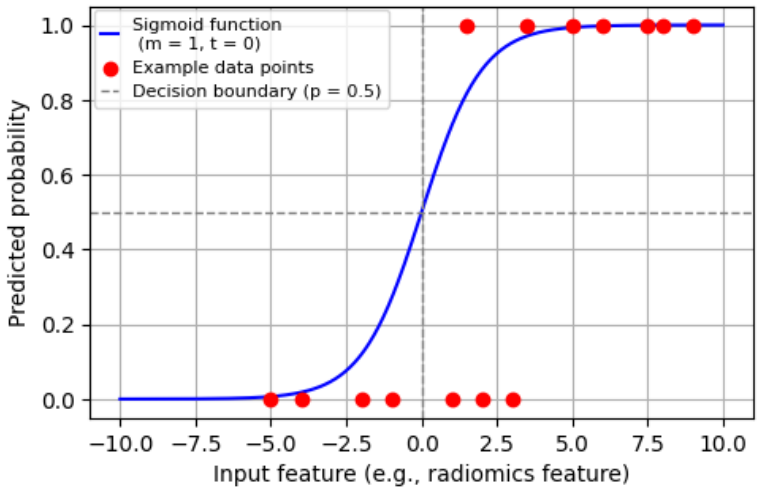

Figure 8: Sigmoid function with added example data

The red dots represent the example data points, where x is an input feature such as a radiomics feature and y is their actual category. If the score given by the sigmoid function *p(x)* displayed in blue is greater than 0.5, the model makes the prediction $\hat{y}$ = 1. *If p(x)* is lower than 0.5, $\hat{y}$ = 0. The value of *x* where *p(x)* = 0.5 is called decision boundary. Observations of larger *x* tend to be classified as 1, while smaller *x* are more frequently classified as 0. The predictions made in Figure 8 would be correct in 11 of 14 times giving it an accuracy of ~0.79. Since the parameters used in this example are arbitrary, its rather good performance is coincidence. In the graphic representation it's obvious that a decision boundary more to the right would give better results, so optimization is clearly possible.

A measure of how good a logistic regression model is performing is the logistic likelihood function *L* that is computed from all data points and yields higher values if the predictions match the observed values:

$$L = \sum_{n=1}^{N} (y_n \, ln(\, p_n) + (1 - y_n) \ln(1 - p_n))$$

where *n* is the individual observation and *N* their total number, y_n is the observed class (0 or 1) and p_n is the score calculated by the model. The last term equals 0 when y = 1, and the first term equals zero if y = 0, so both situations can be handled in one term and be added up.

If e.g. *y* is 1 and *p* is 0.9, this represents a rather good estimate, $y_n \ln(p_n) = -0.11$. If y is 1 and p is 0.1, a bad estimate, $y_n \ln(p_n) = -2.30$, and a higher number is subtracted from the logistic likelihood function. The negative of the likelihood function is called logistic loss function.

For an optimal model we want to compute values for *m* and *t* in order to maximize *L*. A computational method to achieve this is to start with a random or standard *m* and *t* and calculate the gradient of *L* in that point using the derivates with respect to *m* and *t* respectively. By calculating the gradient, we know for *m* and *t*, if we would need to increase or decrease their value to increase *L*, so they are altered in that direction by a small amount defined by the current value and the learning rate *α*. This is repeated until the algorithm converges (the values are close to an optimum), or it has reached a defined number of repetitions. This method of repeated tiny optimizations is called gradient descent.

Especially for multivariate logistic regression which uses multiple coefficients *m*, one for each variable, a regularization term is added to the likelihood function or loss function. This subtracts from the likelihood function for high values of the coefficient *m*. High values for *m* can be interpreted as overestimation of the importance of one feature constituting overfitting. Another aspect is that regularization decreases the coefficients of features not contributing to the prediction to either very small numbers or to 0. Features that are not important are sorted out in that way, representing a sort of internal feature selection process simultaneous to model training.

Scikit learn[81] is an open-source machine learning library for Python. Among many other modules it also provides a class for implementing logistic regression[82]. The model can be trained with the data using the gradient decent algorithm explained above and such identifies the optimal values for m and t, such creating a model that optimally fits the data. In graphical representation, the sigmoid curve is shifted, and its steepness is adapted (m = 0.82, t = -2.14, see Figure 9).

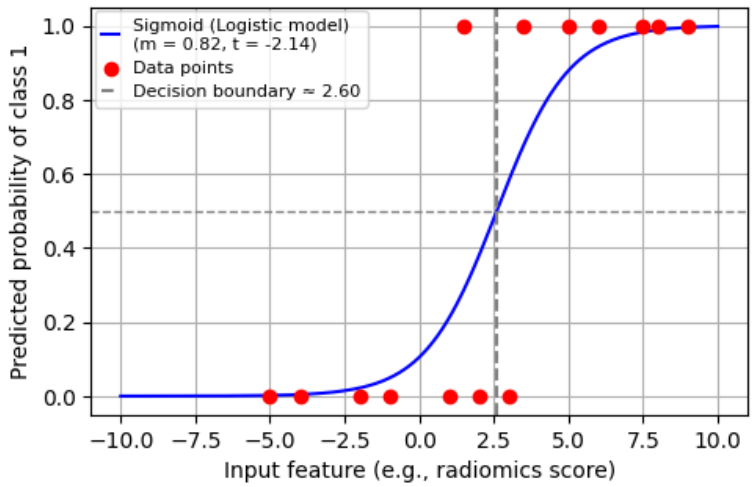

Figure 9: Example data with sigmoid function of a trained model

It is noticeable that the sigmoid function is now shifted to the right, and the decision boundary is now at x ≈ 2.60. Out of the 14 observations 12 are predicted correctly, giving it an accuracy of 0.86. With the now optimally trained model, we can also make predictions for new observations based on their variable *x*.

The next steps are to validate the model and such to assess its performance and reliability, especially when using it on data that was not used for model training. This can for example be performed by k-fold cross validation, where the training data is split, the model is trained on one part and tested with the other. The process can be repeated multiple times. Another method is the leave-one-out validation, where one observation is singled out, the model is trained on all others and applied to the single observation. This method is rather computationally expensive since this is repeated for all observations but resembles the use in a medical setting where eventually the focus is on a single patient to whom a model is applied.

Logistic regression is not limited to binary classification. It can be extended to multiclass problems using strategies such as *one-vs-rest* (OvR) or *multinomial logistic regression*. In OvR, a separate model is trained for each class against all others. In the multinomial approach, all classes are modeled simultaneously. Scikit-learn's LogisticRegression can handle both strategies and automatically switches to the appropriate solver when *multi_class='multinomial'* is selected. This enables logistic regression to be applied to more complex classification tasks with three or more categories.

In view of the relatively small dataset with only 134 patients, logistic regression as a relatively simple linear model is a natural choice. A further benefit is interpretability, since the coefficients can be interpreted as feature importance.

2.3.4. Random Forest

Decision trees are the building blocks of the random forest algorithm.[13,14] A decision tree is a multilayered sequence of decision rules called nodes. Each node splits all incoming observations (or patients) into two subgroups until there is a path of nodes sorting all observations into a terminal node. For a random forest (RF), a large number of decision trees are trained each on a random subset of patients. When using the model for making predictions, an ensemble of decision trees is used to perform a majority vote.

In case of a complex dataset commonly present in radiomics, it is often possible to create a single decision tree that sorts every observation exactly to the desired category via a complex path of many nodes. This tree usually does not generalize well and is the way a decision tree overfits. To prevent overfitting, various rules are introduced , such as limiting the maximum

number of layers (maximum depth), specifying the minimum number of samples required in a group after a split, or setting the minimum number of samples in a node at which the algorithm stops creating further splits, resulting in a terminal node. This reduces the complexity of the trained decision tree and thus reduces overfitting. To understand how decision trees are trained, we need to introduce a measure of dispersion. The default measure implemented in the random forest model in scikit-learn[81] is Gini Impurity *G* with the formula:

$$G = 1 - \sum_{k=1}^{n} {p_k}^2$$

n is the number of classes, *k* the individual class and p_k the probability of that class. If two classes have a similar frequency in the data like in an example *e1* where class 1 has a ratio of *0.48*, we have a Gini impurity of $G_{e1} = 1 - (0.48^2 + 0.52^2) = 0.50$. If the data mostly consists of one class like in an example *e2* where class 1 has a ratio of *0.95*, we have a Gini impurity of $G_{e2} = 1 - (0.95^2 + 0.05^2) = 0.095$. The lower the impurity, the lower the dispersion.

An alternative measure is the entropy *H* with the formula:

$$H = -\sum_{k=1}^{n} p_k \log_2(p_k)$$

In our examples *e1* and *e2* the respective entropies would be $H_{e1} = -(0.48 * \log_2(0.48) + 0.52 * \log_2(0.52)) = 1.00$ and $H_{e2} = -(0.95 * \log_2(0.95) + 0.05 * \log_2(0.05)) = 0.29$. Analogously to the Gini impurity, the lower the entropy, the lower the dispersion. In practice both measures are usually equivalent. Gini impurity is preferred because it is less computationally expensive.

If we want to evaluate how splitting a group in two subgroups changes the dispersion, we can use the difference between the initial Gini impurity and the weighted average of the resulting Gini impurities of the resulting groups after the split. The split with the highest difference is the one that 'organizes' the data the most, so in terms of classification it is the best. To find an optimal way to split the data, we need to calculate the threshold that gives the best difference in Gini impurity for each feature we have and then choose the feature with the best optimal difference. This happens once in the first layer in the so-called *entry node* for the whole data sets and then subsequently for the two subgroups in the second layer and their subgroups in the third layer and so on until the maximum depth or stopping parameter is reached.

When dealing with a large dataset and or a large sample size, it is challenging to create a decision tree that models the complexity of the data without being too extensive and prone to overfitting. To overcome this difficulty, Ho[83] in 1995 proposed to use multiple decision trees

that each are trained in a random subset of the features and to perform a majority voting among the many decision trees. Building on this approach, Breiman in 2001[84] introduced the concept of using random subsets at each individual node, creating the random forest algorithm as is in use today.[85] Multiclass classification is implemented and automatically applied in scikit-learn.

2.4. Objectives and Research Questions

The hypothesis of this research is that histological properties like vascularization, necrosis, and distribution of Antoni A and B areas that are determined by tumor and patient genetics as well as by physiological properties of a tumor such as oxygenation, influence the tumor response and ultimately the outcome after SRS by impacting radiosensitivity, inflammation and regeneration capability, and that the same histological properties influence the radiographic appearance of the tumor in terms of texture and shape.[7–9] Please find a graphical version of this hypothesis in Figure 10.

Depending on the presence of these proposed relations, the prediction of the response to radiosurgery could be possible in vestibular schwannoma, leading to the following main research question:

Does the indirect relationship between genetics, composition of micro-environment or histological structure and the radiographic appearance allow for predictive modelling of tumor response in VS treated with SRS?

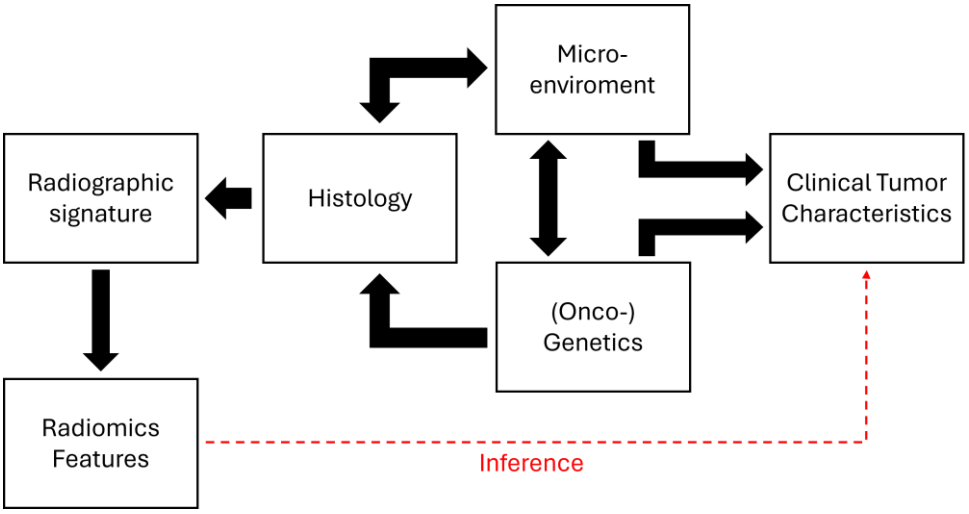

Figure 10: The radiomics hypothesis

This is especially relevant in case of progression after therapy. In a pretherapeutic context, a high likelihood of a progressive disease might facilitate a decision making away from SRS and

most likely towards microsurgery. At the same time, in case of a predicted stable disease, remission or even pseudoprogression will reassure the choice of SRS. Especially in the realm of shifting paradigms (see 2.1.10) this could contribute to a degree of clarity by introducing individualized prospective evidence.

Therefore, the main objective is to create and evaluate radiomics-based machine learning models using the pre-therapeutic planning MRI for the prediction of the volumetric tumor response to facilitate the decision making in a pretherapeutic setting. This prediction could also facilitate assessment of tumor response in case of volume increase after radiosurgery.

3. Material and Methods

3.1. Ethics Statement

This research is purely retrospective and is based on patient records from the Department of Stereotactic and Functional Neurosurgery at the University Hospital of Cologne. All data was available at the start of the project, and no additional data collection was performed. Researchers have lawful access through a contractual relationship with the institution. According to § 6 (2) GDSG NRW, patient consent is not required. Data was pseudonymized, and all data protection regulations were fulfilled.

3.2. Inclusion Criteria

This study included cases of primary progressive VS that were treated at the Department of Stereotactic and Functional Neurosurgery, University Hospital of Cologne (UKK) between 2012 and 2022 with the Cyberknife® who had at least 2 follow up MRI of sufficient quality spanning at least 24 months, and where the proposed response criteria by Rueß et. al. [4] were applicable. Sufficient image quality was defined as a slice thickness of ≤ 2mm for the axial plane or ≤ 3mm if an additional sagittal or coronal plane was available. If both were not applicable, T2-DRIVE images with a slice thickness of ≤ 2mm were used if present and allowed for sufficient discrimination of tumor and bone.

3.3. Stereotactic Radiosurgery

High-resolution contrast enhanced MRI and CT scans were obtained prior to SRS. For MRI, a 1.5 or 3 Tesla scanner (Philips, Hamburg, Germany) was used. For noninvasive fixation of the

patients' head, a thermoplastic mask was fitted to the head. The MR sequences included in the planning MRI protocol are listed in Table 2.

Sequence name	Weighting	Type	Contrast enhancement (CE)	Slice thickness [mm]
T1FFE3D	T1	Fast Field Echo (FFE)	-	1.2
T1FFE3D CE	T1	Fast Field Echo (FFE)	Gadolinium	2
T2TSE3D	T2	Turbo Spin Echo (TSE)	-	2
T2DRIVE3D	T2	Driven Equilibrium (DRIVE)	-	1

Table 2: Modalities of the planning MRI

For the 1-mm slice thickness contrast enhanced planning-CT, a wide-bore tomograph (Toshiba Aquilion, 2011) was used. The MR images were registered to the CT using the Accuray Precision® software. The VS as well as the critical structures were outlined by an experienced neurosurgeon. The treatment planning was performed by a medical physicist. All plans were evaluated in an interdisciplinary conference by a radiation oncologist, a neurosurgeon experienced in SRS and a medical physicist. For the irradiation with the CyberKnife®, the patient was again immobilized by the thermoplastic mask and positioned by the robotic couch of the Cyberknife®. The planned radiation dose was applied by multiple beams converging on the desired target, and intra-fractional patient movements were compensated by image-guidance through orthogonal plain X-rays acquired at a regular pace during treatment. The dose amounted to 12-13Gy referred to the 70-80% isodose.

3.4. Follow-up

The imaging and clinical data were obtained from the electronic medical records stored in the hospital information system, the available imaging studies from the UKK Picture Archiving Systems (PACS) and the CyberKnife® database. Routine clinical and MR-tomographic follow-up examinations were scheduled after 6 months, 12 months and annually in the following.

3.5. Imaging Endpoints

The response criteria proposed by Rueß et al.[4] were used for response assessment as detailed below. The initial volumes at the time of SRS were obtained from the planning contours on the

2mm contrast enhanced T1-FFE MRI images in the planning software Accuray Precision®. The follow-up volumetry was performed using Horos.[86]

Progressive disease (PD), pseudoprogression (PP) and local control (LC) were defined as imaging endpoints as listed in Table 3. This definition is referencing the volumetric aspect of the proposed outcome criteria as introduced by Rueß et al.[4] (see 2.1.9). PD is used in case of a volumetric increase of >40% or change of Koos Grade from I or II to III or IV, PP summarizes early and late pseudoprogression defined as a >20% volume increase followed by a decrease to baseline within 24-36 months regardless of the timing, and LC comprises all other classes.

Category	Definition
Progressive Disease (PD)	≥40% increase in volume relative to baseline and/or change of Koos grade I or II to III or IV
Pseudoprogression (PP)	>20% increase in volume relative to baseline followed by decrease within 24–36 months
Local Control (LC)	<20% increase in volume relative to baseline or decrease in volume

Table 3: Definition of imaging endpoints

This simplification was made to enable model development on a study population of n=125 and is focused on clinical implications. While PD may require salvage SRS or surgery, PP stands out as a PD mimic usually not requiring an intervention, while complete response, partial response and stable disease summarized as LC do not require any intervention.

ML models were trained for each class vs. a combination of the remaining classes as binary model, and one multi-class model was trained for all 3 classes (3-C).

3.6. Software

All analyses were performed using Python[87] if not otherwise specified. Core libraries used were NumPy[88] for numerical operations, pandas[89] for data manipulation, scikit-learn[90] for machine learning and PyRadiomics[20] for feature extraction. For segmentation of the VS in planning MRI and radiation planning, the dose planning software for the Cyberknife^{R}, Accuray Precision^{R} Version 3.3.1.3 was used. For manual segmentation of the follow up images, Horos v4.0.0 RC5 was applied.

Software/Library	Purpose
dcmrtstruct2nii[91]	RTSS conversion
dicom2nifti[92]	Image conversion
joblib[93]	parallelization
matplotlib[94]	Visualization
NumPy[88]	numerical operations
pandas[89]	data manipulation
pycm[95]	confusion matrix calculations
Pydicom[96]	DICOM metadata
PyRadiomics[20]	feature extraction
Python[87]	programming language
scikit-learn[90]	machine learning
SciPy[97]	statistics
seaborn[98]	visualization

Table 4: Used Python libraries

3.7. Image Preprocessing

Medical imaging research and clinical workflows frequently rely on standardized data formats that contain image information and metadata. The most widely used clinical format is the Digital Imaging and Communications in Medicine (DICOM) standard. DICOM was developed to ensure interoperability between medical imaging devices and information systems, and it defines both the image data structure and associated metadata, such as acquisition parameters, patient identifiers, and modality. The individual 2D-slices of a 3D-image are stored in a directory as distinct files. The metadata is kept in the DICOM header of each file.[99,100]

For research applications, DICOM files are often converted into the Neuroimaging Informatics Technology Initiative (NIFTI) format. NIFTI was designed for efficient handling of multidimensional neuroimaging data and has become the de facto standard in computational imaging research. In contrast to most of the DICOM formats, the image as a whole is represented in a 3D-matrix.[101]

In radiation oncology, the delineation of anatomical structures and treatment volumes is typically stored in DICOM-RT Structure Set (RTSS) objects. An RTSS file stores contours of e.g. an VS as part of an irradiation plan as polygonal regions of interest (ROIs).[102]

The CT and MRI images as well as the planning target volumes (PTV) of the radiosurgery plans created by an SRS-experienced neurosurgeon were exported from Accuray Precision®

as DICOM files and stored in a dedicated directory. The directory was scanned, and the DICOM header of all files was read using dcmread[103] of the library pydicom[96]. Relevant information such as patient name, patient ID, modality, series and study instance unique identifiers as well as the file location was stored in a CSV file creating a table of metadata for all DICOM images. This table of metadata was used for guiding all downstream operations including conversion to NIFTI format, association with the PTVs and feature extraction. Each patient was assigned a unique, pseudonymized project ID, and a list of names and IDs was stored separately. The DICOM files were converted to NIFTI files, since this format is required by PyRadiomics[104]. The conversion was performed using dicom2nifti[92] while the RTSS files containing the PTVs were converted using dcmrtstruct2nii[91]. In many neuroimaging studies a brain extraction is performed that isolates the brain and omits the surrounding tissue.[105] Since VS are extra-axial tumors, it was not performed here. A bias field correction which compensates for inhomogeneities of the magnetic fields in the MRI[106] did not seem relevant due to the small volume of the VS.

3.8. Segmentation

As VOI for the feature extraction the planning target volume (PTV) of the irradiation planning was used. In VS, a benign tumor without invasive growth, the PTV is largely identical to the gross tumor volume (GTV) which is in this context defined by the visual tumor margin in the imaging.

3.9. MR-Radiomics Feature Extraction

The process of obtaining the radiomics features followed the principles of the radiomics workflow as described in Chapter 2.2 and is once more visualized in more detail in Figure 11. The features were calculated in adherence to the IBSI standard.[68]

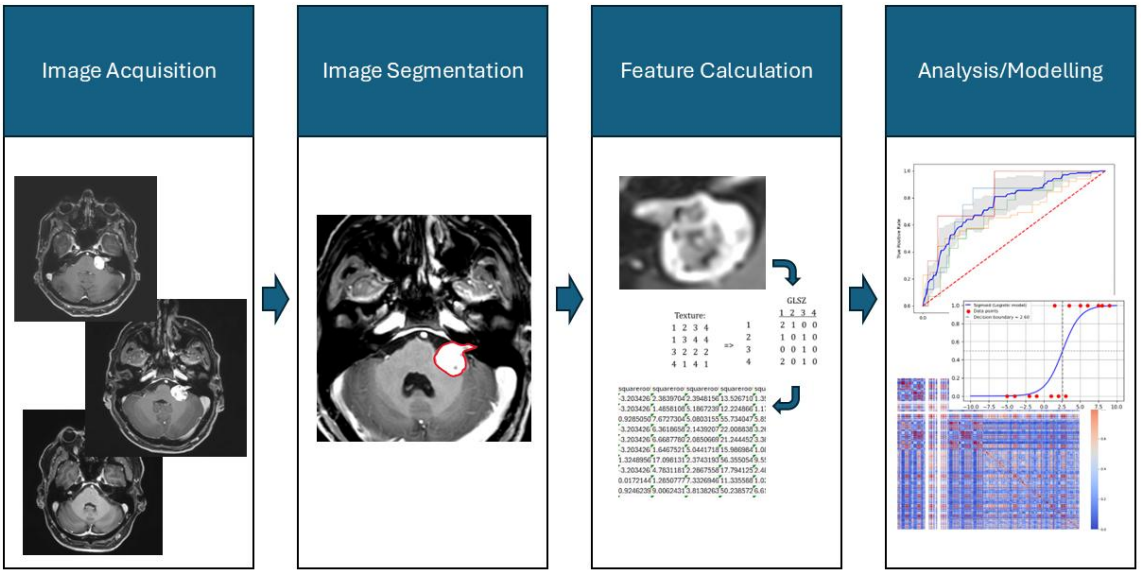

Figure 11: Radiomics Workflow

Before feature extraction, the VOI of the MRI underwent intensity normalization (Z-score normalization). The features were then calculated from the original images as well as from images after the application of a transformation such as application of various wavelet filters and the Laplacian of Gaussian filter, or by mathematical operations on the gray levels with followed normalization, such as squaring, calculating the square root, the exponential, the logarithm and the local binary pattern.

From each PTV and each MR modality (original image and filtered versions), 107 features of different feature classes (see 7.2 in the Appendix) were extracted using pyradiomics.[68]

First-order-statistics radiomics features are derived from the application of descriptive statistics to the numeric distribution of voxel intensities. This includes for example the maximum, minimum, standard deviation, skewness, entropy and others.[20]

3D shape features describe the size and shape of the VOI. For most of the 3D-features, a triangle mesh is generated to represent the surface of the VOI. The mesh is used to calculate features like e.g. volume, surface area, sphericity, and surface area to volume ratio. *2D shape features* work similarly for 2D ROIs.[20]

The *gray level co-occurrence matrix* (GLCM) maps the frequency by which two voxels of two respective gray levels are adjacent or have another specific distance to each other. As described by Haralick et al. in 1973, from the GLCM, textural features such as entropy or contrast can be calculated. The GLCM features were originally intended for the classification of satellite imaging of different terrain types like water or grassland.[107]

In the *gray level run length matrix* (GLRLM), the length and frequency of voxels with the same or similar grey level following each other in one direction are mapped in a matrix. The x-axis of the matrix describes the length, the y-axis the gray scale. The values in the matrix represent the number of lines with the corresponding length and grey scale. Four matrices result for an analyzed image, as a matrix can be calculated in vertical, horizontal and oblique perspective. Features *like short run emphasis* and *gray level non-uniformity* are derived from the matrix. Introduced by Galloway in 1975, it was also proposed to be used for the classification of satellite imaging, especially for capturing patterns such as streaks, uniform textures, and directional structures.[108]

Similarly, the *gray level size zone matrix* (GLSZM) described by Thibault et al. in 2014 does not record the length of distances with the same grey level, but the size and frequency of two-dimensional contiguous zones of voxels with the same gray level. In contrast to GLRLM and GLCM, GLSZM was developed with a medical application in mind. Thibault et al. described the classification of the shape of nuclei in skin fibroblasts in progeria disease.[109]

The *Neighboring Gray Tone Difference Matrix* (NGTDM), introduced by Amadasun et al. in 1989, quantifies the difference between a voxel and the average gray level of its surroundings. From it features like coarseness, contrast, busyness, complexity, and strength can be derived.[110]

The *gray level dependence matrix* (GLDM) quantifies the degree to which gray-level values in an image depend on their neighbors, capturing the size and strength of local gray-level clusters. Unlike GLRLM or GLSZM, which focus on linear or zone structures, the GLDM counts the number of connected voxels (within a defined neighborhood) that are dependent on a central voxel, where dependence is defined by the difference in gray level being less than or equal to a specified threshold. Introduced by Sun et al. in 1982, it was – once again – aimed at the identification of terrain in satellite imaging. A full list of features is provided in 7.2 in the Appendix.

Eventually the radiomics features of the different modalities with or without contrast enhancement (CE) were combined to allow for features of different modalities find entry into one model. Groups used are displayed in Table 5.

individual	**combined**
T1 FFE	T1FFE, T1FFE CE
T1 FFE CE	T1 FFE CE, T2TSE
T2TSE	T1 FFE, T1 FFE CE, T2 TSE

Table 5: Individual and combined modality groups

3.10. Machine Learning Workflow

A conceptual overview over the basic machine learning workflow used here is displayed in Figure 12

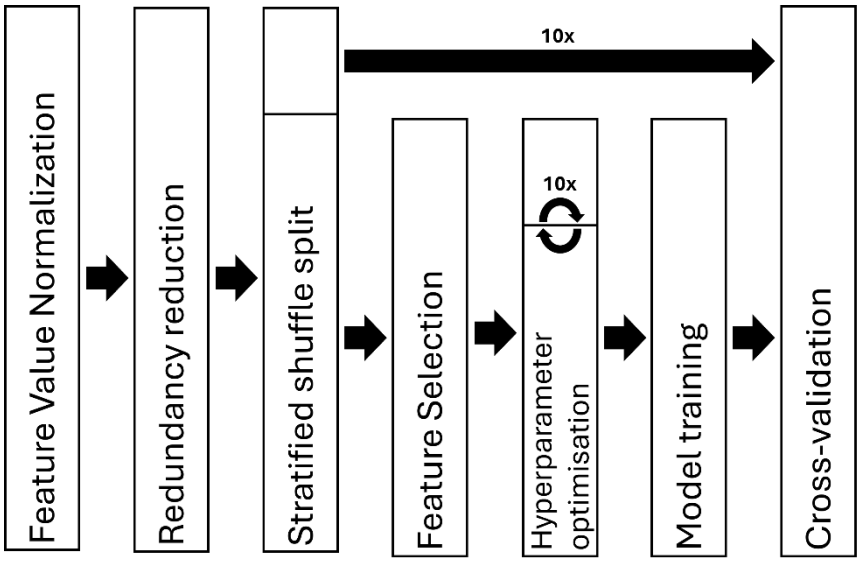

Figure 12: Machine learning workflow

Normalization ensures that the features are scaled. In redundancy reduction, for features that are highly correlated with each other, only one is kept. For the purpose of cross-validation, the data is then randomly split ten times to create different test and training sets. In the next step, features are selected that are significantly correlated with the clinical endpoints. This is followed by the optimization of the hyperparameters facilitated by another, inner loop of cross-validation. In a next step the respective models are trained and then tested on the test set created earlier, constituting the outer cross-validation.

To account for the particularities of the logistic regression and the random forest algorithm, lightly different workflows were created for each. For logistic regression an optional step for feature selection was introduced using an elastic net penalty (Figure 13)

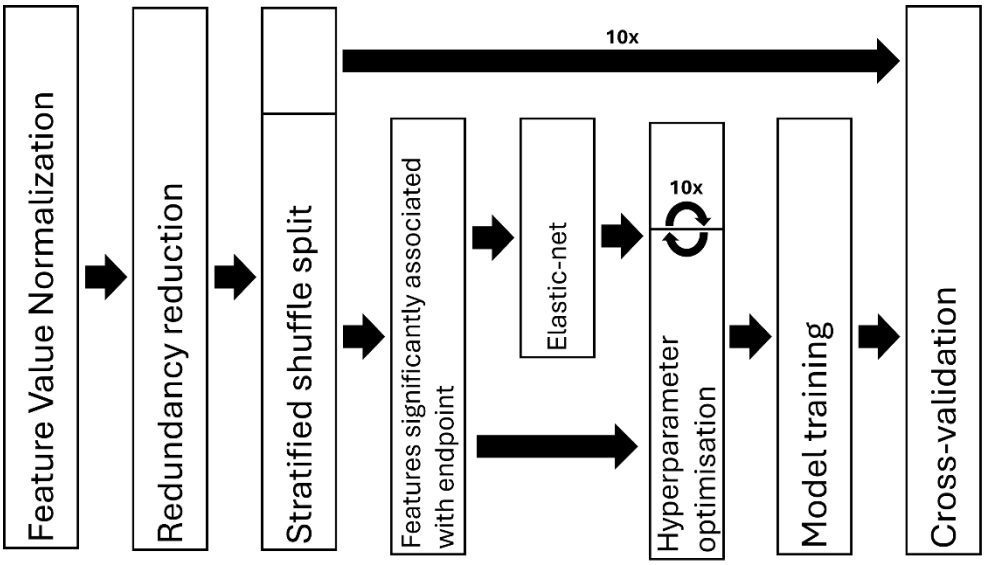

Figure 13: Machine learning workflow for logistic regression

Recursive feature elimination and cross validation (RFECV) was used as an optional step before hyperparameter optimisation as displayed in Figure 14. In an alternative path not depicted here, hyperparameter optimisation was performed first, followed by RFECV using the already optimized hyperparameters.

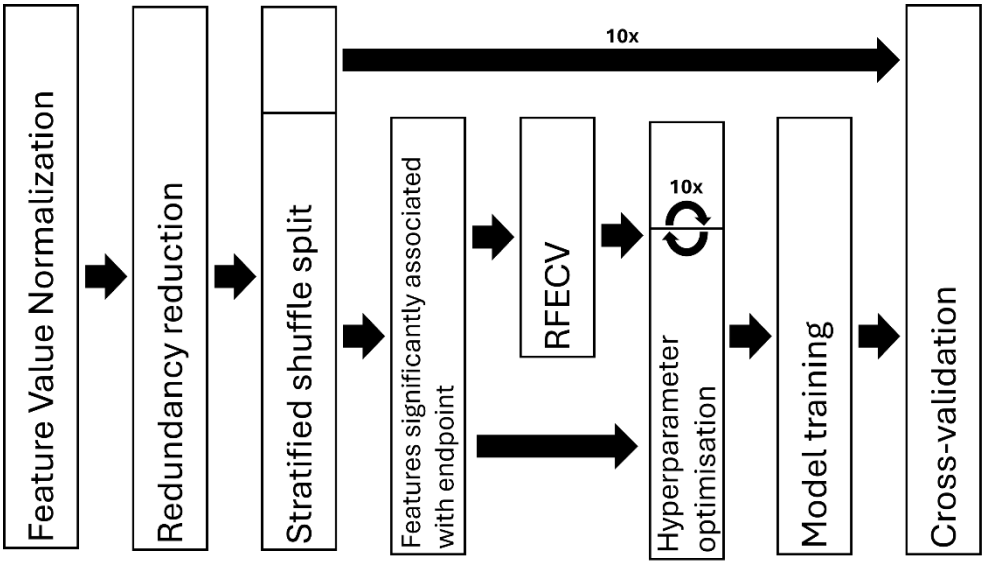

Figure 14: Machine Learning Workflow for Random Forest

The process is repeated for each clinical endpoint and combination of MRI modalities. The workflow is explained in detail in the following chapters.

3.10.1. Feature Value Normalization

For machine learning algorithms in general it is best if the different features used have the same order of magnitude. If that is not the case the algorithms might not converge or result in badly performing estimators.[111] A commonly used approach is the z-score normalization.[112] From each value of each feature the mean of all values of that feature is subtracted and divided by the standard deviation of the values of that feature. For using the model on future or different data it is key to normalize the features from new data with the same means and standard deviations, so they are stored in a normalization matrix, which is not only applied once here but can be used for additional data further on.

3.10.2. Feature Selection Based on Redundancy

Feature selection is an essential step in radiomics-based prediction modeling as it reduces dimensionality, decreases redundancy, and helps to prevent overfitting, particularly when the number of extracted features is high compared to the number of available cases. From the initial set of features, highly inter-correlated variables or irrelevant features are filtered out before model training.[64] For the latter purpose, various algorithms exist, e.g. univariate statistical tests such as the Mann-Whitney-U test or the Kruskal-Wallice-Test that explore the distribution of feature values in the different classes, recursive feature elimination (RFE)[113], or embedded methods such as LASSO that inherently shrink feature coefficients toward zero.[105]

It is customary to perform a layered approach, starting with the deselection of inter-correlated, redundant features, then selecting features that are statistically correlated with the clinical endpoints and then applying further selection methods such as RFE,[114–117] performed inside the outer loop of cross validation in this workflow.

Many machine learning algorithms do not handle features well that are highly correlated to each other. To fix that problem, out of any 2 correlated features only one is kept. For this purpose, a correlation matrix is calculated that contains all correlation coefficients of all features to each other. For visualization, a heatmap of the correlation matrix for all features on the T1 MRI with contrast enhancement is displayed in Figure 15. The red regions represent features that are highly correlated, the blue regions features with low correlation. The red diagonal from the left top to the right bottom represents the correlation of features with themselves resulting in a correlation coefficient of 1. The white regions represent features where the calculation failed, because the feature has zero variance (and is subsequently set to 0 by normalization) or where values are missing due to failure of feature calculation.

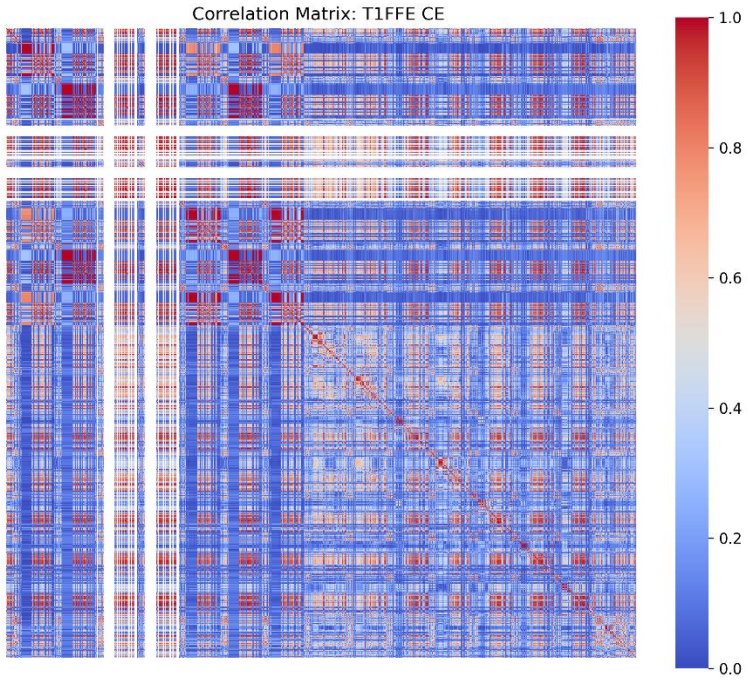

Figure 15: Correlation matrix for features of the contrast enhanced T1 sequence

The next step is to iterate through the matrix and only keep non-correlated features, defined by various thresholds. Redundancy reduction can be performed before the initiation of the outer cross-validation since it is endpoint independent.

3.10.3. Train/Test-Set Split for Initiation of Nested Cross-Validation

The usefulness of a trained machine learning model or an estimator is defined by its ability to generalize, i.e. to make accurate predictions on new, unseen data that were not used for model training. A simulation of the performance on new data can be done by splitting the data into a train and a validation set. Model training is performed on the training set, and the performance is measured by applying the trained model to the validation set. In order to reduce the impact of an arbitrary split, this procedure is repeated multiple times, and the average of the performance measures of the individual validation folds is used as performance measure.[118]

Cawley and Talbot[119] make the case that over-fitting is not only a concern of model training, but also of model selection including the optimization (tuning) of hyperparameters. According to them, even if only the model selection and the hyperparameter optimization are performed on the entire data, the model might be selected in a way that makes it better in predicting test sets of that specific dataset, even though that data was not used in the fitting. This implies that the performance estimate of the trained models would be overly optimistic, while the capability of predicting new data would be decreased.

To overcome this, they propose to implement methods like nested cross-validation as for example described by Stone in 1974.[15] While model selection and hyperparameter optimization is performed in an inner loop, the model is then tested in an outer loop with data that was not used for model selection, hyperparameter optimization or model fitting. The final model could then be the best estimator from the outer cross-validation or the best model with the best hyperparameters, that is trained on the entirety of the data.

While this type of over-fitting undoubtedly applies to complex models, Wainer and Cawley[120] later described that it is only relevant to a limited extent for classifiers with few hyperparameters as were used here, and another downside is that the nested approach comes with an increased computational cost. Analogous to hyperparameter optimization, it is possible to introduce bias and an overly optimistic performance measure by feature selection before cross validation as described in an example of gene-expression data by Ambroise and McLachlan.[121] Interestingly, Demircioğlu demonstrated that the AUC-ROC is positively biased by up to 0.15 in published radiomics related papers when feature selection is performed on the entire dataset.[122] Therefore, the authors advise to include the feature selection in the inner fold of the cross-validation.

Since the next step of feature selection will be depending on the correlation with the clinical endpoints it is necessary to initiate the outer cross-validation first. A typical design for nested cross validation is a 10x10 approach meaning having 10 folds each in the inner and outer cross validation.[120,123,124] Due to the sample size of only up to ~120 cases depending on the used modalities and clinical endpoint, the unequal frequency of different classes of clinical endpoints and the necessity of all classes being present in some scoring methods of the hyperparameter tuning, we decided to apply a test split size of 0.2. The split was performed using StratifiedShuffleSplit.[125]

3.10.4. Feature Selection Based on Correlation with Clinical Data

To filter out relevant features further, the next step is to select those with statistically significant correlation with the clinical endpoints. Since the radiomics data are non-parametric, for the classification tasks the Mann-Whitney-U-test (MWU) were used for the binary (e.g. PP versus no PP), and the Kruskal-Wallis-test (KW) was used for the multivariate classification (e.g. proposed response-criteria). The features were then selected based on p-value-cutoffs. For the calculation of the p-values, manwhitneyu[126] and kruskal[127] of the library scypy[97] were used.

3.10.5. Feature Selection Based on Regularization of Logistic Regression

Another optional step that further reduces the number of features for logistic regression is based on the magnitude of the coefficients of a trained regression model. The coefficients can be interpreted as feature importance, since values with low absolute coefficients or coefficients set to zero have no impact on the result of the predictions while features with high absolute coefficients are most influential. The number of features to select was set to the square root of the number of patients. It is controversial if Ridge Regression (L2) or Lasso (L1)[ii] is the preferred regularization for coefficient-based feature selection[105]. We decided to use a mix of both, using ElasticNet[128] with a L1-ratio of 0.5, so both regularization methods are used and are equally weighted.

3.10.6. Feature Selection with Recursive Feature Elimination and Cross-Validation (RFECV)

Another feature selection method that is used optionally here is recursive feature elimination and cross validation (RFECV) first described by Guyon in 2002[113]. Here, repeatably models are trained and cross-validated. After each repetition one (or more) least important features are removed to find an optimal number and the optimal features. RFECV uses a machine learning model with fixed hyperparameters. So either RFECV is performed using standard hyperparameters or hyperparameter optimization is performed before RFECV. Here, both ways were implemented. Since RFECV did not perform well with logistic regression, its use was limited to the workflow of random forest models.

3.10.7. Hyperparameter Optimization

Besides selecting a model, also selecting its hyperparameters that define its properties independently of the training data like e.g. the regularization strength in Logistic regression or the maximum depth in random forest is crucial. For each hyperparameter of interest, a range

---

[ii] L1 and L2 describe different regularization terms (Ω). Both penalize large coefficients (ω). L1 adds the sum of the magnitudes of the coefficients to the cost function of the logistic regression algorithm:

$$\Omega_{L1} = \sum |\omega|$$

while L2 overly penalizes large coefficients by adding the sum of the squared coefficients:

$$\Omega_{L2} = \sum \omega^2$$

of all values that shall be tested is created. An algorithm then tests all possible combinations. The hyperparameters of the estimator that performs best in a cross-validation are then used for further model training. Hyperparameter optimization was performed using GridSearchCV[129] of scikit-learn[81]. As metric for the best hyperparameters, ROC-AUC, accuracy and the F1-score were used for the classification by scikit-learn's[81] LogisticRegression[82] and RandomForestClassifier[85] models.

3.10.8. Model Training for Cross-Validation

With each set of best hyperparameters, another estimator was trained on the entire train split of the outer cross-validation with the best hyperparameters of GridSearchCV. Decision boundaries or thresholds were defined for each class individually based on maximizing the F1-score (see below) on all cases excluding the respective validation set of the outer cross-validation.

3.10.9. Evaluation

When making predictions using logistic regression or random forest, the estimator returns a score from 0 to 1. If it exceeds a threshold (that is yet to defined), the observation is predicted to be in category 1. For multinomial logistic regression a score is given for every category. By default, the threshold or decision boundary is 0.5 as described above. This might not necessarily be the best decision boundary, especially in case of unequal proportions of the classes. A custom threshold can be chosen to maximize certain performance measures such as the F1-score (see below).

The estimators are evaluated by comparing their predictions to the true values on the held-out set. Independent from the model type, the predictive performance of classifiers can be determined by counting the number of correct and incorrect classifications with respect to the underlying ground truth, called accuracy, and further analyses can be made using Receiver Operating Characteristic (ROC) by applying a range of decision boundary values to the confusion matrices, resulting in metrics such as sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), F1-score, and the area under the ROC curve (AUC).

The confusion matrix reports the number of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) see Figure 16.

	Predicted positive	**Predicted negative**
Real positive	True positive (TP)	False negative (FN)
Real negative	False positive (FP)	True negative (TN)

Figure 16: Confusion matrix

From that Table the true positive rate (TPR) commonly called sensitivity in medical context is calculated as

$$TPR = \frac{TP}{TP + FN}$$

The true negative rate (TNR) or specificity is calculated as

$$TNR = \frac{TN}{TN + FP}$$

The positive predictive value (PPV) is the rate of a predicted positive being a real positive and is calculated as

$$PPV = \frac{TP}{TP + FP}$$

Analogously the negative predictive value (NPV) is the rate of a predicted negative being a real negative and is calculated as

$$NPV = \frac{TN}{TN + FN}$$

A score commonly used for reporting the predictive performance in a single value is the F1-score. It is calculated as the harmonic mean of PPV and TPR:

$$F1 = \frac{2\,PPV \times TPR}{PPV + TPR}$$

The ROC curve plots the true-positive rate against the false-positive rate across all possible decision thresholds, providing a threshold-independent evaluation of the classifier's

discriminative ability. The AUC summarizes this performance into a single value between 0 and 1, where 0.5 equals a random classifier, 1 defines a perfect classifier.[130–133] The AUC was calculated using scikit-learn's[81] roc_auc_score[134]. The other metrics were calculated using ConfusionMatrix[135] of the library pycm[95]. The evaluation was repeated in each of the 10 folds and averaged.

4. Results

The objective of this research was to explore the predictability of the long-term volumetric outcome of VS after SRS using pretherapeutic MRI by radiomics based machine learning models. This approach is thought to support the general radiomics hypothesis which postulates a correlation between radiographic signature, histology and clinical tumor characteristics also in the case of vestibular schwannoma, and in addition could facilitate predictive modeling of the treatment response of VS to SRS especially with regard to progressive disease after therapy.

4.1. Patient Characteristics

This single-center retrospective study included 122 cases (63 female, 59 male) of primary progressive VS which were treated by robotic-guided SRS utilizing a Cyberknife® at the Department of Stereotactic and Functional Neurosurgery of the Uniklinik Köln (UKK) between 2012 and 2022. The age ranged from 20 years to 82 years with an average of 58.0±13.5 years, and the average follow up duration was 56.8±21.6 months. The mean pretherapeutic tumor volume was 1.44±1.43mL, while the median was 0.96mL (IQR 1.36mL) indicating a left skewed distribution as visualized in Figure 17 (left).

Number of patients		**122**
Female : Male		**63:59**
Age [years]		**58.0±13.5**
Follow-up duration [months]		**56.8±21.6**
Tumor Volume before SRS [mL]		
Mean		1.44±1.43
Median		0.96
IQR		1.39
Koos Grade:		
I	20	16.4%
II	65	53.2%
III	35	28.7%

IV	2	1.6%
Imaging Endpoints		
Progressive Disease (PD)	10	8.2%
Pseudoprogression (PP)	29	23.8%
Local Control (LC)	83	68.0%

Table 6: Patient characteristics

More than two thirds (69.6%) appeared to be small intra-canalicular tumors or tumors with extra-canicular extension without contact to the brainstem classified as Koos Grade I and II, while 28.7% were categorized as Koos grade III indication brainstem contact and only 2 cases were classified as Koos Grade IV indicating brainstem compression. The distribution is visualized in Figure 17 (right).

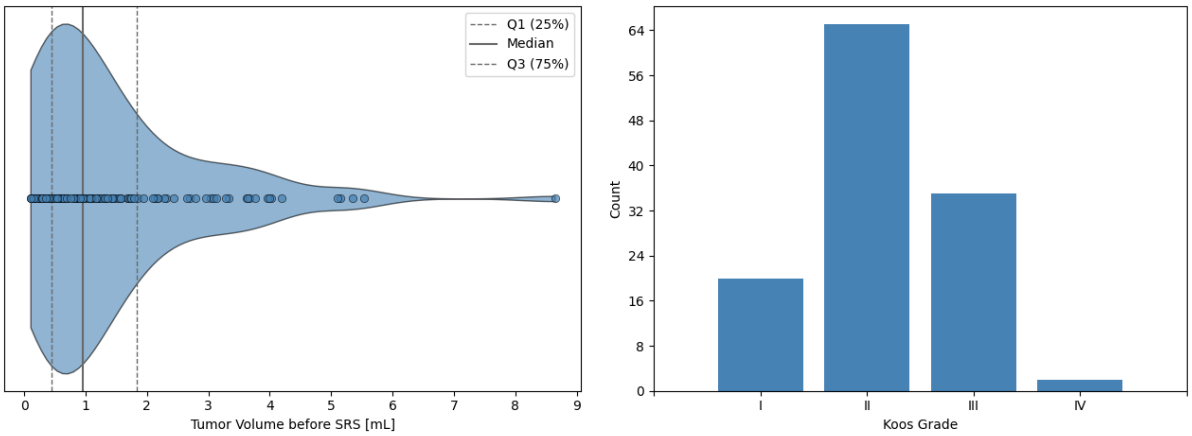

Figure 17: Pretherapeutic tumor volume distribution (left) and Koos Grade (right)

Following SRS, 68.0% of the VS had a stable volumetric course or underwent partial remission, summarized as local control (LC), while 23.8% presented with an early or late pseudoprogression. This indicates a radiographic success rate of the single session SRS of 91.8%. In total, 8.2% of the patients had true progression triggering either salvage SRS or microsurgical resection. The characteristics of the patient cohort are listed in Table 6. In Figure 18 the course of the relative changes in volume of all individual cases up until end of follow-up or salvage treatment is plotted on the left and the average volume changes and their standard deviation are shown on the right.

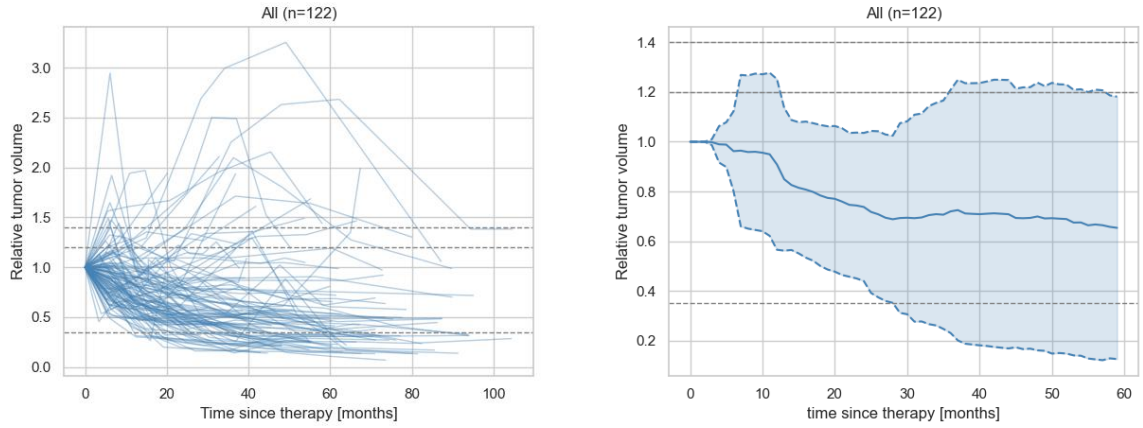

Figure 18: Follow-up of volume changes of individual cases (left) and mean follow up volume changes ± standard deviation (right) of all included cases

4.2. Progressive Disease (PD)

As described above, particular attention was directed towards the prediction of progressive disease (PD) as this event might necessitate salvage SRS or microsurgery (MS). A prediction of a high probability of PD might even be used for indicating primary MS.

The 10 cases of PD included in this study appeared to exhibit a heterogenous distribution of pretherapeutic tumor volume and Koos Grade (Figure 19).

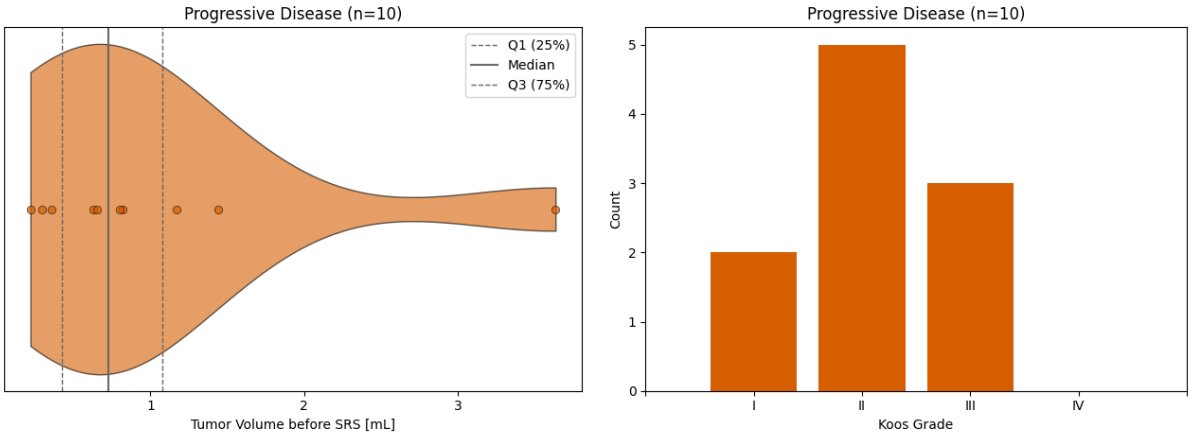

Figure 19: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in cases that later developed PD

As the plot of individual follow-up volume changes shown in Figure 20, also the timepoint of progression proves to be heterogenous.

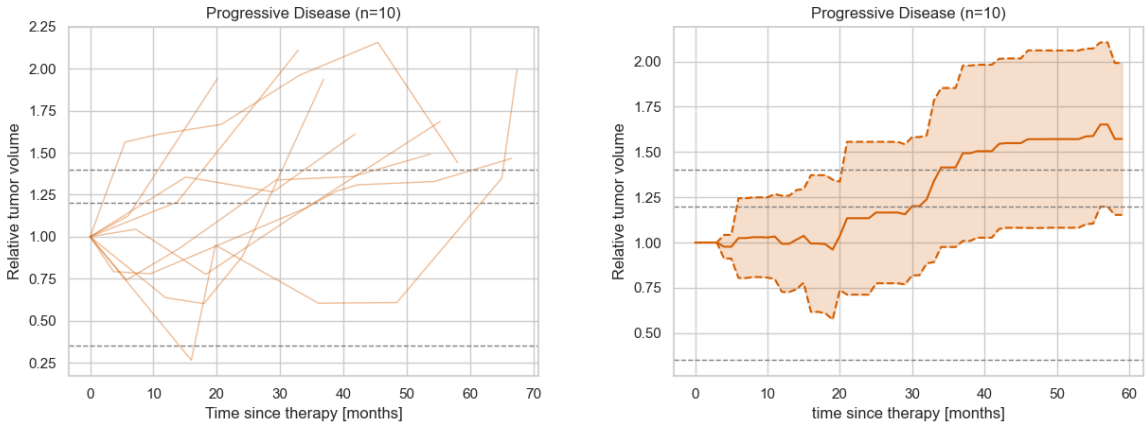

Figure 20: Follow-up volumes of individual cases (left) and mean follow up volume ± standard deviation (right) of PD

With regard to the predictability of PD using MR-radiomics based ML-models, the best random forest estimator scored a ROC-AUC of 0.83 ± 0.16 with an F1-score of 0.60 ± 0.21. It was trained on features of the T1FFE only. The feature selection path included the optional layer of RFECV which selected 3.4 ± 2.8 features. The ROC- curve is displayed in Figure 21.

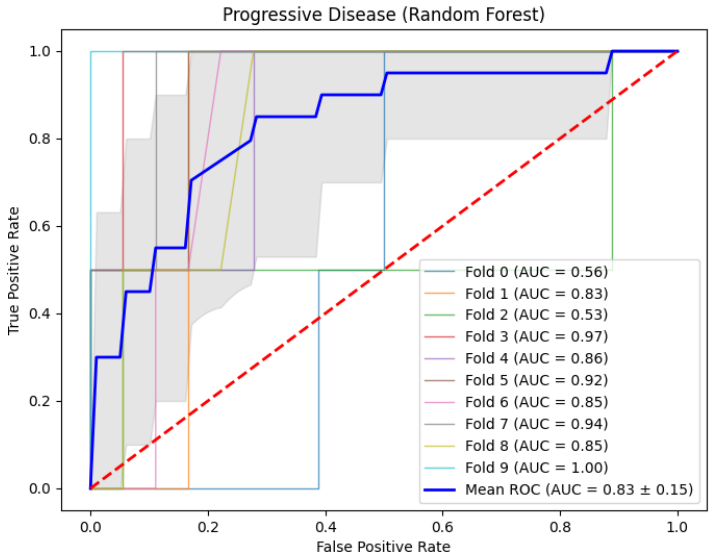

Figure 21: ROC AUC curve of the PD binary response model using random forest

As described in material and methods, all metrics are arithmetic means of the outer 10-fold cross-validation. Further metrics as well as the metrics of the best logistic regression estimator are displayed listed in Table 7.

Clinical Endpoint	Progressive Disease	
Model	Logistic Regression	Random Forest
AUC	0.80 ± 0.08	0.83 ± 0.16
F1-score	0.52 ± 0.13	0.60 ± 0.21
Accuracy	0.82 ± 0.13	0.86 ± 0.16
Sensitivity	0.80 ± 0.26	0.80 ± 0.26
Specificity	0.82 ± 0.17	0.87 ± 0.18
PPV	0.49 ± 0.29	0.59 ± 0.31
NPV	0.98 ± 0.03	0.98 ± 0.03

Table 7: Results of the outer cross-validation for binary classification models for PD

The most commonly selected features among the ten folds are listed in Table 8. First order kurtosis in the gradient transformed MRI of the native T1 sequence was selected in all cases. The gradient transformation maps the local intensity gradients in the MRI and such emphasizes contrast rich regions. The feature is a measure of peakedness in the intensity distribution. A high value reflects a distribution with rich tails and a narrow distinct peak, while a low value reflects a distribution with a wide peak and not so prominent tails.[10]

feature	modality	frequency
gradient_firstorder_Kurtosis	T1FFE	10
wavelet-HLL_firstorder_Skewness	T1FFE	5
exponential_firstorder_Skewness	T1FFE	3
wavelet-HHL_firstorder_Median	T1FFE	2
wavelet-LHL_glcm_ClusterProminence	T1FFE	2
squareroot_firstorder_Range	T1FFE	1
wavelet-HLH_glcm_ClusterProminence	T1FFE	1
wavelet-LHL_firstorder_Entropy	T1FFE	1
wavelet-LLH_glcm_SumEntropy	T1FFE	1
wavelet-HLL_firstorder_Kurtosis	T1FFE	1

Table 8: The ten most frequently selected features of the reported binary PD model

The first order skewness in a wavelet transformed native T1 image was selected in half the folds. The HLL wavelet transformation emphasizes elongated edges or anisotropic structures, while the skewness describes the asymmetry of the intensity distribution. A high positive value reflects a long right tail, while high negative values reflect a long left tail. Values close to zero indicate close to equal tails.[10]

4.3. Pseudoprogression (PP)

As pseudoprogression (PP) is defined by a transient volume enlargement, the clinical challenge is the differentiation between PP and PD. The study cohort contained 29 cases with a volume increase of more than 20% with a consecutive decrease within 24-36 months, thus fulfilling the proposed definition of PP according to Rues et. al.[4]. Like with PD, the subgroup with PP presented with diverse volume and Koos Grade distributions as displayed in Figure 22.

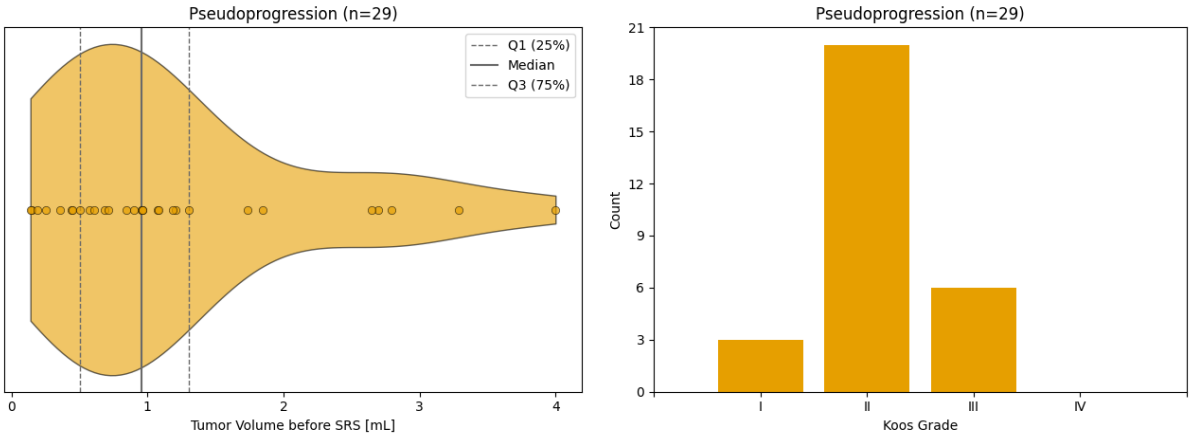

Figure 22: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in included cases of PP

PP is subclassified in early and late pseudoprogression with a cutoff of the volume peak occurring before or after 12 months following SRS. In the plots of the follow up volume changes (see Figure 23), both the group of early PP (n=19) and late PP (n=10) form distinguishable peaks.

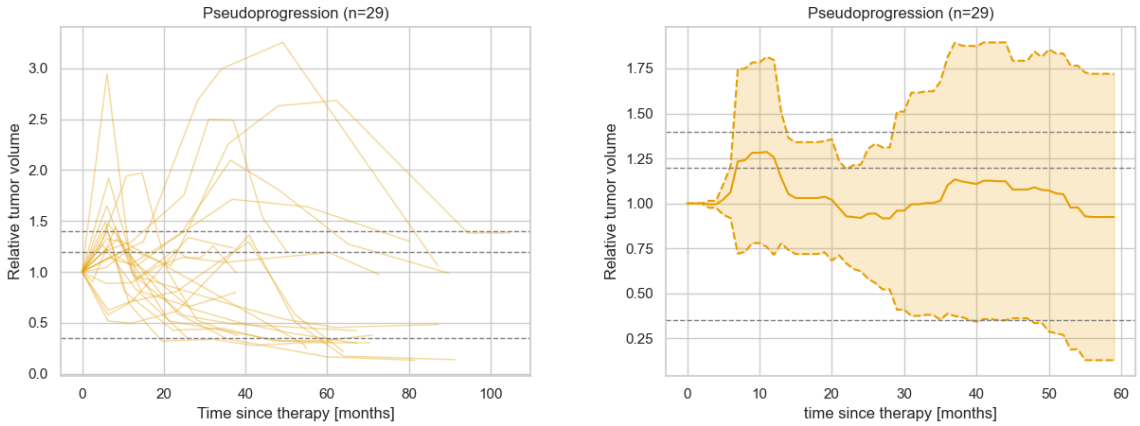

Figure 23: Follow-up volume changes of individual cases (left) and mean follow up volume changes ± standard deviation (right) in cases of PP

The best estimator for the prediction of PP proved to be a random forest model with a mean ROC-AUC of 0.69 ± 0.10 and a mean F1-score of 0.51 ± 0.11. The AUC-curve is displayed in

Figure 24. It likewise was trained exclusively on T1FFE without CE only and utilized a feature selection path including RFECV resulting in 31.9 ± 18.1 features. Further metrics and the metrics of the best logistic regression model are listed in Table 9.

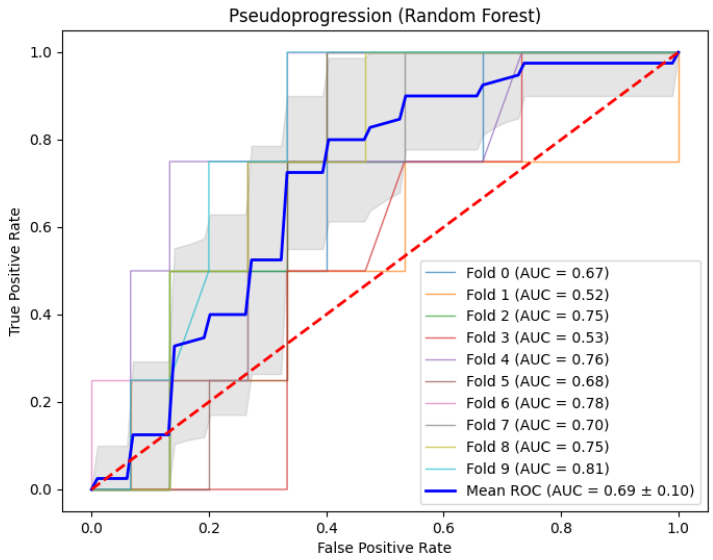

Figure 24: ROC AUC curve PP binary response model using random forest

Clinical Endpoint		**Pseudoprogression**
Model	Logistic Regression	Random Forest
AUC	0.67 ± 0.10	0.69 ± 0.10
F1-score	0.46 ± 0.10	0.51 ± 0.11
Accuracy	0.55 ± 0.23	0.59 ± 0.21
Sensitivity	0.82 ± 0.17	0.90 ± 0.13
Specificity	0.47 ± 0.33	0.51 ± 0.28
PPV	0.35 ± 0.15	0.37 ± 0.12
NPV	0.92 ± 0.05	0.96 ± 0.05

Table 9: Results of outer cross-validation for binary classification models for PP

The two features selected in all ten folds were the 90th Percentile of the intensity distribution in a wavelet transformed T1FFE with CE and the minimum value in the intensity distribution in another wavelet transformed T1FFE with CE. Further features are displayed in Table 10.

feature	**modality**	**frequency**
wavelet-LLH_firstorder_90Percentile	T1KM	10
wavelet-LHH_firstorder_Minimum	T1KM	10
original_gldm_LargeDependenceLowGrayLevelEmphasis	T1FFE	9
wavelet-HLL_firstorder_Skewness	T1KM	9
wavelet-LLH_glszm_GrayLevelNonUniformityNormalized	T1FFE	9
original_firstorder_Minimum	T1KM	8
wavelet-LHL_firstorder_Skewness	T1FFE	8

wavelet-LHH_firstorder_10Percentile	T1KM	8
wavelet-HLH_glszm_SmallAreaEmphasis	T1KM	8
original_firstorder_10Percentile	T1KM	7

Table 10: The ten most frequently selected features of the reported binary PP model

4.4. Local Control (LC)

The majority (n=68) of the cohort consisted of patients with local control. The group comprises all cases fulfilling the definitions of stable disease (n=48) and partial remission (n=35) which usually does not require further intervention. The pretherapeutic volume distribution and Koos Grades are displayed in Figure 25. Of note, both cases classified as Koos Grade IV fell into this favorable category.

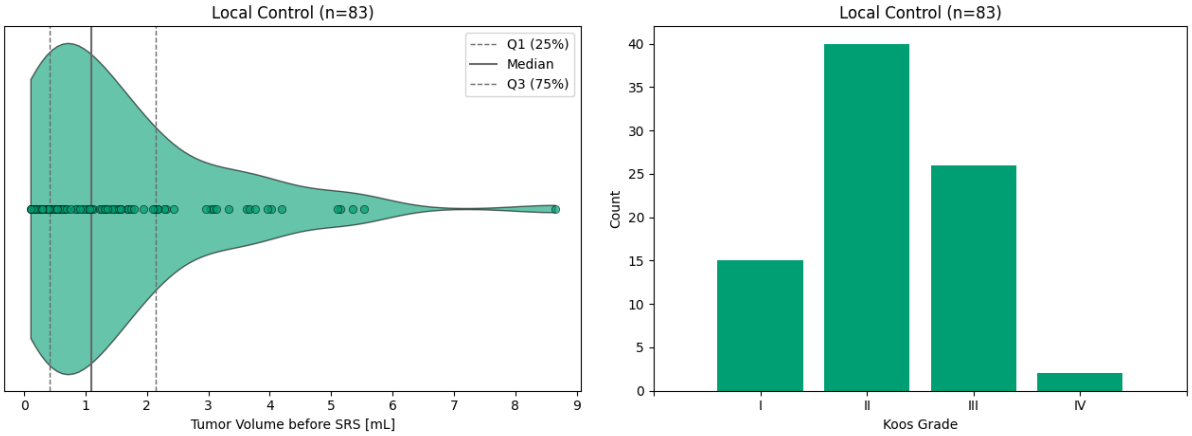

Figure 25: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in included cases of LC

The follow-up volume changes of individual cases and the mean follow up volume changes including standard deviation are displayed in Figure 26.

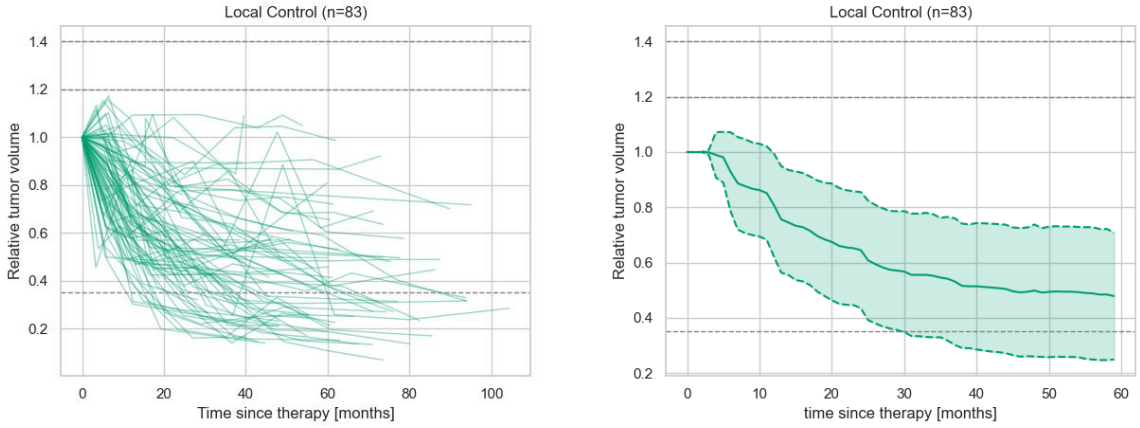

Figure 26: Follow-up volumes of individual cases (left) and mean follow up volume ± standard deviation (right) of LC

The best estimator, again a random forest model, resulted in a ROC AUC of 0.64 ± 0.09 and an F1-score of 0.83 ± 0.01. The ROC-curve is displayed in Figure 27. It was trained on a combined feature pool auf T1FFe with and without CE and used a feature selection protocol including RFECV selecting 38.9 ± 12.9 features.

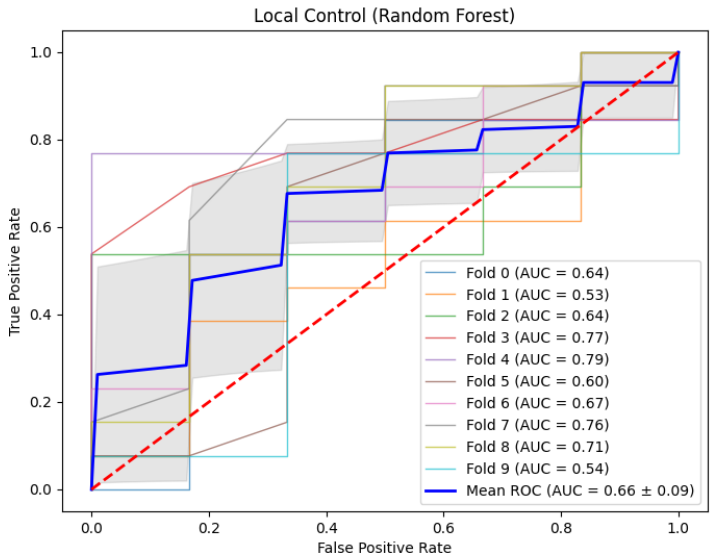

Figure 27: ROC AUC curve LC binary response model using random forest

Further metrics and metrics of the best logistic regression estimator are displayed in Table 11.

Clinical Endpoint		**Local Control**
Model	Logistic Regression	Random Forest
AUC	0.64 ± 0.09	0.66 ± 0.09
F1-score	0.83 ± 0.01	0.83 ± 0.02
Accuracy	0.72 ± 0.03	0.72 ± 0.04
Sensitivity	1.00 ± 0.00	0.99 ± 0.02
Specificity	0.10 ± 0.09	0.12 ± 0.16
PPV	0.71 ± 0.02	0.71 ± 0.04
NPV	1.00 ± 0.00	0.95 ± 0.11

Table 11:Results of outer cross-validation for binary classification models for LC

Here, also 2 features were selected by all folds, both from a wavelet transformed T1FFE with CE: the entropy and the kurtosis. The latter, as described above, reflects the peakedness of the distribution. The entropy, already mentioned in the description of the random forest algorithm, describes the randomness of voxel values. High entropy indicates a heterogeneous and irregular patter, whereas low entropy corresponds to a more regular or homogeneous texture.[10]

feature	modality	frequency
wavelet-LLH_firstorder_Entropy	T1KM	10
wavelet-HLL_firstorder_Kurtosis	T1KM	10
wavelet-HHL_firstorder_Skewness	T1KM	9
wavelet-HLH_firstorder_Entropy	T1FFE	9
wavelet-LHH_gldm_DependenceNonUniformityNormalized	T1KM	8
wavelet-LHL_glcm_ClusterProminence	T1FFE	8
wavelet-LLH_glszm_GrayLevelNonUniformityNormalized	T1FFE	8
wavelet-HLH_firstorder_Mean	T1KM	8
gradient_firstorder_Kurtosis	T1FFE	8
wavelet-LLH_glrlm_HighGrayLevelRunEmphasis	T1KM	8

Table 12: The ten most frequently selected features of the reported binary LC model

4.5. Multiclass Classification

Besides the binary ML-models differentiating between PD and no PD, PP and no PP as well as LC and no LC, also a 3-way multiclass classifiers (3-C) were trained.

The best estimator for this classification was a logistic regression estimator. The ROC-AUC was 0.72 ± 0.07 and the F1-score was 0.25 ± 0.01. The ROC-curve is displayed in Figure 28. The feature selection from for this model resulted in features from T1FFE with and without CE as well as the T2TSE, included the optional layer of selection via an elastic net regularized logistic regression model, and resulted in 8 selected features which was the predefined number to select.

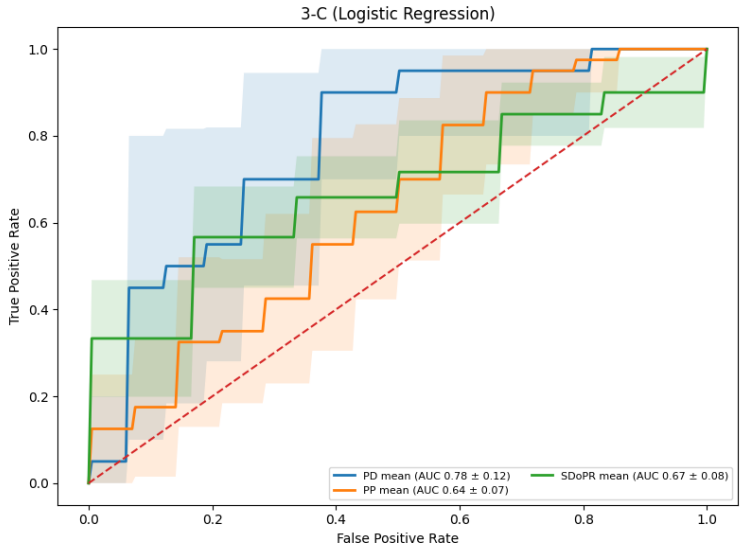

Figure 28: ROC AUC curve for the 3-C multiclass response model using logistic regression

Further metrics as well as the metrics of the best random forest estimator are displayed in Table 13.

Metric	Logistic Regression	Random Forest
AUC	0.72 ± 0.07	0.70 ± 0.08
F1-score	0.25 ± 0.01	0.33 ± 0.13
Accuracy	0.73 ± 0.03	0.73 ± 0.06
Sensitivity	0.29 ± 0.02	0.37 ± 0.15
Specificity	0.63 ± 0.02	0.67 ± 0.08
PPV	0.23 ± 0.01	0.50 ± 0.18
NPV	0.58 ± 0.04	0.66 ± 0.10

Table 13: Results of outer cross-validation for 3-C

It is noticeable that like in the binary classification, the prediction of PD performed with a comparably high AUC score of 0.78 ± 0.12.

Like the binary classification random forest PD estimator, here also the first order feature kurtosis in the gradient transformation of the T1FFE without was selected in all folds. The second most frequently feature chosen one was the 10th percentile of the intensity distribution in a wavelet transformed T1FFE with CE. The ten most frequently chosen features are listed in Table 14.

feature	modality	frequency
gradient_firstorder_Kurtosis	T1FFE	10
wavelet-HLH_firstorder_10Percentile	T1KM	7
wavelet-LLH_firstorder_Skewness	T1KM	6
wavelet-LLL_glszm_HighGrayLevelZoneEmphasis	T1KM	5
wavelet-LHL_firstorder_Skewness	T2TSE	4
wavelet-HLL_glszm_ZoneEntropy	T1KM	3
wavelet-LLH_firstorder_90Percentile	T1KM	3
wavelet-HLL_glszm_SizeZoneNonUniformityNormalized	T1KM	3
wavelet-LHH_glcm_SumEntropy	T1KM	3
wavelet-HLL_firstorder_Kurtosis	T1KM	3

Table 14: The ten most frequently selected features of the reported 3-C model

5. Discussion

Addressing the question whether MR-radiomics based prediction of volumetric tumor response of VS after SRS is feasible, we observed that it was possible to build moderate to good performing models especially for the prediction of progressive disease in a relatively small cohort of 122 patients (see Figure 29).

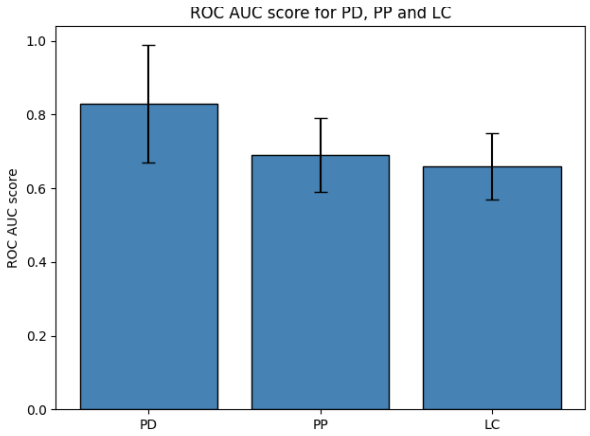

Figure 29: ROC AUC scores for Progressive Disease (PD), Pseudoprogression (PP) and Local Control (LC)

This strongly supports the hypothesis that MRI appearance is linked to the histology of the tumor and probably also to the genetics, biodynamics and radiosensitivity which ultimately determine tumor response. While the level of predictive power does not yet meet the requirements for clinical deployment, these findings are encouraging and motivate to explore and develop this concept further.

5.1. Progressive Disease (PD)

The most significant finding result is the good performance of the prediction of progressive disease with a mean ROC-AUC of 0.83 ± 0.16 and an F1-score of 0.60 ± 0.21 (see Figure 29). This is especially encouraging, since PD usually requires further treatment like salvage SRS or microsurgery. This result needs to be evaluated in view of the fact that only 10 out of the 122 patients experienced progressive disease. For the nested cross-validation, two layers of test sets with a test size of 0.2 were created. This means that features were selected and hyperparameters were optimized on less than 80 patient train-set, containing six to seven cases of PD, with a test-set size of around 20 patients containing one or two patients with PD. The tuned model of the inner loop was then trained on around 100 cases and tested on roughly 20 in the outer cross validation containing 2 cases of PD.

These findings demonstrates that, even when based on a small number of positive cases, radiomics-based modeling can achieve useful predictive performance. It seems reasonable to

assume that larger datasets could enable the development of more precise and generalizable prediction models.

A challenge in the model development is the dispersion of the timepoint of PD (see Figure 30). While two cases reached the 40% mark within 10 months, one case underwent progression after over 5 years. Since the inclusion criteria only required a follow up of 24 months, misclassification of cases with shorter follow-up duration seems inevitable, both impeding the model development and limiting the scope of the prediction.

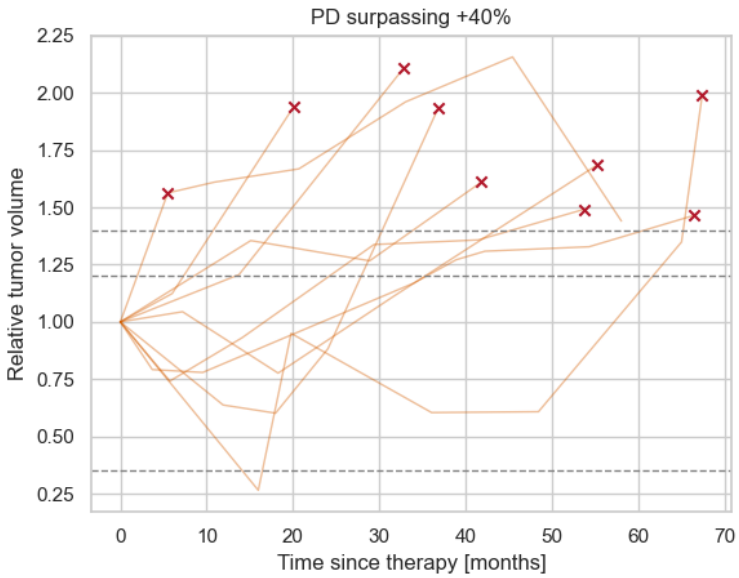

Figure 30: Time points when individual cases of PD first surpassed +40%

Another point worth noticing is the heterogeneity of VS resulting in PD in terms of volume and Koss Grade. Also, both cases classified as Koos Grade IV did not undergo progression within the timeframe of the follow-up. Even though the examination of non-radiomics factors does not fall in the scope of this study, a rule of thumb like 'SRS fails in large VS' would not be supported by the cohort data.

5.2. Pseudoprogression (PP)

The ROC-AUC of 0.69 ± 0.10 based on 28 out of 122 cases is a solid finding suggesting the predictability of PP from other courses to some degree. Even though prediction of PP is not as critical as the prediction of PD, such models could contribute to the evaluation of the general expected clinical course of a patient case. The lower-than-PD-ROC-AUC is certainly partially explained by the heterogeneity of PP by definition – early and late PP – which is also reflected by the courses of the volumetric changes observed here and visualized once again in Figure 31.

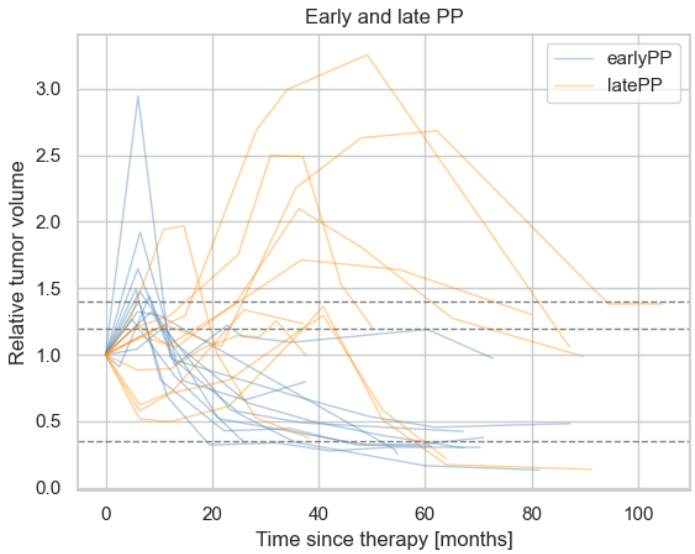

Figure 31: Follow-up volumes of individual cases of early PP (blue) and late PP (orange)

While early PP clearly forms a peak within the first 12 months, late PP do not present as adjacent continuum but rather accumulate around month 40. In this small cohort with only ten cases of late PP this could be coincidental. Nevertheless, a deeper understanding of the differences between early and late PP in terms of pathophysiology and its influence of timing would be interesting.

A limiting factor and a potential source of misclassification here is that some cases presented with a pronounced volume increase before returning to pretherapeutic levels. This volume increase might have been interpreted as PD in other cases, causing secondary treatment and causing the case being classified as PD in this study, despite eventual volume decrease and the classification as PP which would have followed.

5.3. Local Control (LC)

While the ROC-AUC 0.64 ± 0.09 is solid but not overwhelming, the F1-score of 0.83 ± 0.01 looks promising at first glance, but this figure is misleading. The F1-score is defined as the harmonic mean of precision (= TP / (TP + FP)) and recall (= TP / (TP + FN)). In the case of imbalanced classes with a high proportion of real positives, this is also reflected by the high sensitivity (true positive rate) of 0.99 ± 0.02 while the specificity (true negative rate) is as low as 0.12 ± 0.16. This means positives get identified correctly almost always, while negatives mostly are falsely classified as positives. The estimator almost always guesses LC and is right many times thus scoring a high F1 score. Of note, this is not a flaw of the model or model training itself, but results from the definition of the decision boundary. Here, the threshold is set to achieve the maximum F1-score which gives this false incentive to always score LC if

this is the dominant class. Since the ROC-AUC is threshold independent it does serve as a reliable performance estimate.

There are several aspects that render this shortcoming less relevant. First, the identification of LC is not the primary clinical interest. The clinical goal is not to identify future LC from a large group that should be treated by SRS, but the goal is to identify those with low probability to achieve LC where different treatment options might or might not be favorable and secondly, thresholds can be set manually ad libitum. The maximization of the F1-score was chosen for this research because it proves to be useful in vision of PD and PP, which are primary clinical interests. When tuning thresholds for clinical practice a balanced sensitivity and specificity might be desired, or false negatives and false positives could be tied to a certain ethical or financial cost which could be minimized. While this would result in a model with lower F1-score and higher usefulness in clinical practice, this was not addressed in the present work. Another approach avoiding this problem in the first place is multiclass classification.

5.4. Three-class Model (3-C)

In practice, having three individual estimators bears the difficulty that they could be contradictory. Fortunately, both logistic regression and random forest offer a multiclass classification with the correct definition of thresholds and decision-making algorithms (see 3.10.9) that provide an unambiguous prediction.

The ROC-AUC proved to be reasonably good with a score of 0.72 ± 0.07 and the F1 fell short with a score of 0.25 ± 0.01 which is unsurprising for a multiclass classification and such small cohort size. Also, unsurprisingly in view of the good performance in the binary models, the subgroup with the highest ROC-AUC score of 0.78 ± 0.12 was PD.

5.5. Methodical Considerations

Limitations that need to be addressed are the single center and even single scanner design used in a retrospective setting. The generalization to an external validation group with MRI from other scanners as well as further prospective studies would be a valuable next step. Another opportunity would be examination of transferability of the models to cases treated with other modalities of SRS such as the Gamma Knife.

5.6. Comparison to Literature

In comparison to the three prior publications on prediction of volumetric response of VS after SRS using radiomics (as of 04.05.2025[136,137]), a strength of the present research is the usage of the volumetric response criteria as proposed by Rueß et al..[4] Bossi Zanetti et al.[6] used the tumor volume after 24 and 34 months after treatment, but did not differentiate between PD and PP, while Langenhuizen et al.[21] defined any volume increase within 2 years as PP and any later volume increase as PD, such that cases with a volume increase within 2 years resulted in treatment. Here cases of late PP were not appreciated as such and cases of early progression were excluded. Yang et al.[22] defined a non-responding tumor as an increase of tumor volume or a decrease of less than 10% in the last image of an observation period of at least two years. The clinical relevance of distinguishing between a substantial increase in tumor volume, which may warrant surgical intervention, and an unchanged tumor volume that requires no treatment remains unclear..

Another methodical strength of the present investigation is usage of nested cross-validation where all endpoint-related procedures including feature selection and hyperparameter optimization were accounted. Bossi Zanetti et. al.[6] reported a similar procedure. Yang et al.[22] reported the usage of a random 30% holdout set for feature selection followed by an inner loop of 10-fold cross-validation for hyperparameter optimization and model training. While this procedure does test the model on data that was not involved in the creation of the model, a non-iterative approach lacks stability, as by chance a particularly beneficial or non-beneficial holdout-set might have been chosen.

Langenhuizen et al.[21] reported the usage of cross-validation only for model training. As discussed in 3.10.3, feature selection can largely contribute to overfitting and can substantially increase the resulting validation metrics. While laying valuable groundwork as pioneers in outcome prediction of vestibular schwannoma after stereotactic radiosurgery and suggesting general feasibility of such modeling, the reported AUC scores of up to 0.99 are thus unlikely to represent a reliable indicator of generalizability beyond the study cohort.

It is also worth mentioning that the features defined by the Image Biomarker Standardization Initiative[68] and made available in the Python library PyRadiomics[20] were used in the present study whereas the features used by Bossi Zanetti et. al.[6]. Langenhuizen et al.[21] and Yang et al.[22] relied on in-house developed sets. In terms of cohort size, Yang et al.[22] included 336 patients, Bossi Zanetti et al.[6]. 108 patients and Langenhuizen et al.[21] 85 patients, leaving the present research's cohort of 134 patients outnumbered by only one study.

Regarding model performance , leaving aside Langenhuizen et al.[21] due to the renunciation of a holdout set or a nested cross-validation, Yang et al.[22] reported an excellent AUC score of 0.91 for the binary classification of tumor response, where the classes were built on the rather arbitrary criterion whether the tumor volume at the end of the follow up period of at least two years amounted to more or less than 90% of the pretherapeutic tumor volume. As discussed, the stability of the evaluation with a holdout set without iteration is limited. Repetition of the analysis with different random holdout sets might have led to divergent results. The second binary model for the prediction of PP within the group of patients with an eventual tumor volume of 90% or less compared to before therapy was reported to score an AUC of 0.88. Being still very high, the interpretability in a clinical context appears to be difficult, since the prediction of PP relies on the premise of tumor regression. A Bayesian analysis of both models or a multiclass model accounting for all possible cases would facilitate the interpretation in terms of clinical applicability. A possible explanation of the overall high scores is the compatible large cohort of 336. Bossi Zanetti et. al.[6] did report accuracy, sensitivity and specificity but no AUC scores, what limits the interpretability of the results in case of an presumably imbalanced dataset[iii]. The main metrics of the three mentioned publications are displayed in Table 15.

	Cohort size	**accuracy**	**sensitivity**	**specificity**	**AUC**	**F1-score**
Langenhuizen et al.[21]	85	0.77	0.71	0.83	0.93	-
Langenhuizen et al.[21] (Tumors >5mL)	29	0.83	0.83	0.82	0.99	-
Yang et al.[22] Response vs Nonresponse	336	0.88	0.86	0.91	0.91	0.88
Yang et al.[22] Response with PP vs without PP	336	0.85	0.83	0.88	0.88	0.85
Bossi Zanetti et. al.[6] 24 months	108	0.73	0.6	0.85	-	-
Bossi Zanetti et. al.[6] 36 months	108	0.65	0.47	0.83	-	-

Table 15: Reported classifier metrics in prior publication regarding volumetric response of VS to SRS. Not reported is indicated by '-'.

[iii] E.g. in case of a dataset with 80% being class one, an estimator always predicting class one has an accuracy of 80%, while still not being informative. Sensitivity and specificity do help in imbalanced datasets. A single score metric like AUC or F1 would nevertheless help with interpretation.

Comparing the reported results of studies with nested cross-validation of a holdout set, it seems that the better performing models of the present research surpass the performance of those reported by Bossi Zanetti et. al.[6], and reach a similar performance by means of the present 3-C model, but cannot reach the reported performance of Yang et al.[22], what may in part be due to the 2.5 times smaller cohort size.

Further limitations to acknowledge in the present study are the imbalance in some of the clinical endpoint groups somewhat limiting the stability of the evaluation even when using nested-cross-validation approaches, the single center design and the absence of an external validation set, limiting the demonstration of generalizability. A larger cohort size would also certainly contribute to the development of a more robust model. Since radiomics features might vary in case of use of different scanners or protocols, this one scanner, one protocol study, comes with the benefit of stability of radiomic features within the dataset, which also limits the transferability of the findings to different or mixed constellations.

5.7. Outlook

The estimators trained here generate categorical predictions. With larger numbers and better models also probabilistic predictions would be possible. A future procedure could be, that if the probability of PD surpasses a certain threshold, say 95%, the patient is recommended for primary microsurgery instead of SRS. If this level of confidence could be reached for half of all cases with PD, and assuming the proportion of PD of 8% as observed here is representative, by recommending microsurgery to 4 out of 100 patients, the treatment failure rate could be cut by half and the rate of successful treatment in this hypothetical scenario could be increased from 92% to 96%. This benefit would be counterweighted by 2 out of 1000 cases (0.2%) undergoing microsurgery despite SRS being an effective treatment. For many centers this would affect less than 2 patients per decade.

Besides increasing the numbers, another way to increase the model performance could be the addition of non-radiomics features (already piloted by Bossi Zanetti et. al.[6]). This could include genetic information, which is already performed in metastatic brain tumors where in addition, ontogenetic information from the primary tumor often available. [138]

Another opportunity is the development of models that make predictions for all available management options including scan and wait, microsurgery and systemic treatment, so that for each case a profound, evidence-based, and most important, personalized outlook for each treatment option could be provided. Further endpoints like the short- and long-term functional outcome in terms of hearing preservation or vertigo could be included in such modeling. While

it is unclear if the realization of this vision will be technically possible, the findings of this research advocate for the exploration of such concepts. The next immediate step taken by the author will be the development of models for the prediction of functional outcome regarding hearing and vertigo in cases of VS treated with SRS. Since the development of the ML-pipeline required substantial effort (~9 months), a future project will be to make the code reusable and accessible for the public to facilitate the development of new models for other clinical questions.

6. Bibliography

1 Kumar V, Abbas AK, Aster JC, editors. Schwannomas. In: Robbins & Cotran pathologic basis of disease, Tenth edition. Philadelphia, PA: Elsevier, 2021: p.1237.

2 Lin EP, Crane BT. The Management and Imaging of Vestibular Schwannomas. *AJNR Am J Neuroradiol* 2017; **38**: 2034–43.

3 Goldbrunner R, Weller M, Regis J, *et al.* EANO guideline on the diagnosis and treatment of vestibular schwannoma. *Neuro-Oncology* 2020; **22**: 31–45.

4 Rueß D, Schütz B, Celik E, *et al.* Pseudoprogression of Vestibular Schwannoma after Stereotactic Radiosurgery with Cyberknife®: Proposal for New Response Criteria. *Cancers* 2023; **15**: 1496.

5 Rueß D, Pöhlmann L, Grau S, *et al.* Outcome and toxicity analysis of single dose stereotactic radiosurgery in vestibular schwannoma based on the Koos grading system. *Sci Rep* 2020; **10**: 9309.

6 Bossi Zanetti I, De Martin E, Pascuzzo R, *et al.* Development of Predictive Models for the Response of Vestibular Schwannoma Treated with Cyberknife®: A Feasibility Study Based on Radiomics and Machine Learning. *J Pers Med* 2023; **13**: 808.

7 Aerts HJWL, Velazquez ER, Leijenaar RTH, *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014; **5**: 4006.

8 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016; **278**: 563–77.

9 Lambin P, Leijenaar RTH, Deist TM, *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017; **14**: 749–62.

10 Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - Reference Manual. *Radiology* 2020; **295**: 328–38.

11 Berkson J. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* 1944; **39**: 357–65.

12 Cramer JS. The Origins of Logistic Regression. *SSRN Journal* 2003. DOI:10.2139/ssrn.360300.

13 Decision Trees. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp.195-208.

14 Ensemble Learnig and Random Forests. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp.211-235.

15 Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 1974; **36**: 111–33.

16 Fernández-Méndez R, Wan Y, Axon P, Joannides A. Incidence and presentation of vestibular schwannoma: a 3-year cohort registry study. *Acta Neurochir* 2023; **165**: 2903–11.

17 Matthies C, Samii M. Management of 1000 Vestibular Schwannomas (Acoustic Neuromas): Clinical Presentation. *Neurosurgery* 1997; **40**: 1.

18 Reznitsky M, Petersen MMBS, West N, Stangerup S-E, Cayé-Thomasen P. Epidemiology Of Vestibular Schwannomas – Prospective 40-Year Data From An Unselected National Cohort. *Clin Epidemiol* 2019; **11**: 981–6.

19 Gupta VK, Thakker A, Gupta KK. Vestibular Schwannoma: What We Know and Where We are Heading. *Head Neck Pathol* 2020; **14**: 1058–66.

20 Van Griethuysen JJM, Fedorov A, Parmar C, *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* 2017; **77**: e104–7.

21 Langenhuizen PPJH, Zinger S, Leenstra S, *et al.* Radiomics-Based Prediction of Long-Term Treatment Response of Vestibular Schwannomas Following Stereotactic Radiosurgery. *Otol Neurotol* 2020; **41**: e1321–7.

22 Yang H-C, Wu C-C, Lee C-C, *et al.* Prediction of pseudoprogression and long-term outcome of vestibular schwannoma after Gamma Knife radiosurgery based on preradiosurgical MR radiomics. *Radiotherapy and Oncology* 2021; **155**: 123–30.

23 Kumar V, Abbas AK, Aster JC, editors. Robbins & Cotran pathologic basis of disease, Tenth edition. Philadelphia, PA: Elsevier, 2021.

24 Wippold FJ, Lubner M, Perrin RJ, Lämmle M, Perry A. Neuropathology for the Neuroradiologist: Antoni A and Antoni B Tissue Patterns. *AJNR Am J Neuroradiol* 2007; **28**: 1633–8.

25 Koos WT, Day JD, Matula C, Levy DI. Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas. *Journal of Neurosurgery* 1998; **88**: 506–12.

26 Mulkens TH, Parizel PM, Martin JJ, *et al.* Acoustic schwannoma: MR findings in 84 tumors. *American Journal of Roentgenology* 1993; **160**: 395–8.

27 Jeng CM, Huang JS, Lee WY, Wang YC, Kung CH, Lau MK. Magnetic resonance imaging of acoustic schwannomas. *J Formos Med Assoc* 1995; **94**: 487–93.

28 Gomez-Brouchet A, Delisle MB, Cognard C, *et al.* Vestibular schwannomas: correlations between magnetic resonance imaging and histopathologic appearance. *Otol Neurotol* 2001; **22**: 79–86.

29 Charabi S, Mantoni M, Tos M, Thomsen J. Cystic vestibular schwannomas: neuroimaging and growth rate. *J Laryngol Otol* 1994; **108**: 375–9.

30 Spickler E, Lufkin R, Teresi L, *et al.* MR of hemorrhagic acoustic neuromas. *Comput Med Imaging Graph* 1991; **15**: 333–7.

31 George-Jones NA, Chkheidze R, Moore S, Wang J, Hunter JB. MRI Texture Features are Associated with Vestibular Schwannoma Histology. *The Laryngoscope* 2021; **131**. DOI:10.1002/lary.29309.

32 Paldor I, Chen AS, Kaye AH. Growth rate of vestibular schwannoma. *J Clin Neurosci* 2016; **32**: 1–8.

33 Stangerup S-E, Caye-Thomasen P, Tos M, Thomsen J. The Natural History of Vestibular Schwannoma. 2006; **27**.

34 Germano IM, Sheehan J, Parish J, *et al.* 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. *Neurosurgery* 2018; **82**: E49–51.

35 Dhayalan D, Tveiten ØV, Finnkirk M, *et al.* Upfront Radiosurgery vs a Wait-and-Scan Approach for Small- or Medium-Sized Vestibular Schwannoma: The V-REX Randomized Clinical Trial. *JAMA* 2023; **330**: 421–31.

36 Greenberg MS, Abel N, Agazzi S. Handbook of Neurosurgery. Georg Thieme Verlag, 2020 DOI:10.1055/b-006-161147.

37 Harati A, Scheufler K-M, Schultheiss R, *et al.* Clinical features, microsurgical treatment, and outcome of vestibular schwannoma with brainstem compression. *Surg Neurol Int* 2017; **8**: 45.

38 Won S-Y, Kilian A, Dubinski D, *et al.* Microsurgical Treatment and Follow-Up of KOOS Grade IV Vestibular Schwannoma: Therapeutic Concept and Future Perspective. *Front Oncol* 2020; **10**. DOI:10.3389/fonc.2020.605137.

39 Ferrara N, Hillan KJ, Gerber H-P, Novotny W. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. *Nat Rev Drug Discov* 2004; **3**: 391–400.

40 Lu VM, Ravindran K, Graffeo CS, *et al.* Efficacy and safety of bevacizumab for vestibular schwannoma in neurofibromatosis type 2: a systematic review and meta-analysis of treatment outcomes. *J Neurooncol* 2019; **144**: 239–48.

41 Leksell L. The stereotaxic method and radiosurgery of the brain. *Acta Chir Scand* 1951; **102**: 316–9.

42 RADIOCHEMICAL METHODS | Gamma-Ray Spectrometry. In: Encyclopedia of Analytical Science. Elsevier, 2005: 72–9.

43 Meeks SL, Pukala J, Ramakrishna N, Willoughby TR, Bova FJ. Radiosurgery technology development and use. *J Radiosurg SBRT* 2011; **1**: 21–9.

44 Chang SD, Adler JR. Robotics and Radiosurgery – The Cyberknife. *Stereotactic and Functional Neurosurgery* 2002; **76**: 204–8.

45 Yang I, Udawatta M, Prashant GN, *et al.* Stereotactic Radiosurgery for Neurosurgical Patients: A Historical Review and Current Perspectives. *World Neurosurgery* 2019; **122**: 522–31.

46 Lam FC, Byun J, Guru S, *et al.* Understanding the Radiobiology of Central Nervous System Diseases in the Golden Age of Radiosurgery—Does It Matter? *Brain Sciences* 2025; **15**: 649.

47 Tuleasca C, Tripathi M, Starnoni D, Daniel R, Reyns N, Levivier M. Radiobiology of Radiosurgery for Neurosurgeons. *Neurol India* 2023; **71**: 14.

48 Withers HR. The Four R's of Radiotherapy. In: Lett JT, Adler H, eds. Advances in Radiation Biology. Elsevier, 1975: 241–71.

49 Steel GG, McMillan TJ, Peacock JH. The 5Rs of Radiobiology. *International Journal of Radiation Biology* 1989; **56**: 1045–8.

50 Boustani J, Grapin M, Laurent P-A, Apetoh L, Mirjolet C. The 6th R of Radiobiology: Reactivation of Anti-Tumor Immune Response. *Cancers (Basel)* 2019; **11**: 860.

51 Taghizadeh-Hesary F. “Reinforcement” by Tumor Microenvironment: The Seventh “R” of Radiobiology. *International Journal of Radiation Oncology, Biology, Physics* 2024; **119**: 727–33.

52 Park HJ, Griffin RJ, Hui S, Levitt SH, Song CW. Radiation-Induced Vascular Damage in Tumors: Implications of Vascular Damage in Ablative Hypofractionated Radiotherapy (SBRT and SRS). *rare* 2012; **177**: 311–27.

53 Mangoni M, Borghesi S, Aristei C, Becherini C. Radiobiology of stereotactic radiotherapy. *Rep Pract Oncol Radiother* 2022; **27**: 57–62.

54 Ishikawa H, Ma Z, Barber GN. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* 2009; **461**: 788–92.

55 Yovino S, Kleinberg L, Grossman SA, Narayanan M, Ford E. The Etiology of Treatment-related Lymphopenia in Patients with Malignant Gliomas: Modeling Radiation Dose to Circulating Lymphocytes Explains Clinical Observations and Suggests Methods of Modifying the Impact of Radiation on Immune Cells. *Cancer Investigation* 2013; **31**: 140–4.

56 Thust SC, van den Bent MJ, Smits M. Pseudoprogression of brain tumors. *J Magn Reson Imaging* 2018; **48**: 571–89.

57 Huang RY, Bi WL, Weller M, *et al.* Proposed response assessment and endpoints for meningioma clinical trials: report from the Response Assessment in Neuro-Oncology Working Group. *Neuro-Oncology* 2019; **21**: 26–36.

58 Balossier A, Tuleasca C, Delsanti C, *et al.* Long-Term Hearing Outcome After Radiosurgery for Vestibular Schwannoma: A Systematic Review and Meta-Analysis. *Neurosurgery* 2023; **92**: 1130–41.

59 Jakubeit T, Sturtz S, Sow D, *et al.* Single-fraction stereotactic radiosurgery versus microsurgical resection for the treatment of vestibular schwannoma: a systematic review and meta-analysis. *Syst Rev* 2022; **11**: 265.

60 Daloiso A, Cazzador D, Concheri S, Tealdo G, Zanoletti E. Long-Term Hearing Outcome For Vestibular Schwannomas After Microsurgery And Radiotherapy: A Systematic Review and Meta-Analysis. *Otolaryngol Head Neck Surg* 2024; **171**: 1670–81.

61 Stastna D, Urgosik D, Chytka T, Liscak R. Large vestibular schwannomas: long-term outcomes after stereotactic radiosurgery. *Neuro Endocrinol Lett* 2021; **41**: 329–38.

62 Peker S, Samanci Y, Ozdemir IE, Kunst HPM, Eekers DBP, Temel Y. Long-term results of upfront, single-session Gamma Knife radiosurgery for large cystic vestibular schwannomas. *Neurosurg Rev* 2022; **46**: 2.

63 Mayerhoefer ME, Materka A, Langs G, *et al.* Introduction to Radiomics. *J Nucl Med* 2020; **61**: 488–95.

64 Lambin P, Rios-Velazquez E, Leijenaar R, *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* 2012; **48**: 441–6.

65 Tomaszewski MR, Gillies RJ. The Biological Meaning of Radiomic Features. *Radiology* 2021; published online Jan 5. DOI:10.1148/radiol.2021202553.

66 Majumder S, Katz S, Kontos D, Roshkovan L. State of the art: radiomics and radiomics-related artificial intelligence on the road to clinical translation. *BJR Open* 2023; **6**: tzad004.

67 George-Jones NA, Chkheidze R, Moore S, Wang J, Hunter JB. MRI Texture Features are Associated with Vestibular Schwannoma Histology. *Laryngoscope* 2021; **131**: E2000–6.

68 Zwanenburg A, Vallières M, Abdalah MA, *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020; **295**: 328–38.

69 Dutton DM, Conroy GV. A review of machine learning. *The Knowledge Engineering Review* 1997; **12**: 341–67.

70 Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Markets* 2021; **31**: 685–95.

71 Shinde PP, Shah S. A Review of Machine Learning and Deep Learning Applications. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). Pune, India: IEEE, 2018: 1–6.

72 Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems, First edition. Beijing ; Boston: O'Reilly Media, 2017.

73 Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology* 2019; **19**: 64.

74 Sadr H, Nazari M, Khodaverdian Z, *et al.* Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research* 2025; **30**: 418.

75 Shortliffe, Edward H. Computer-Based Medical Consultations: MYCIN. New York: Elsevier, 1976 https://www.shortliffe.net/Shortliffe-1976/MYCIN%20thesis%20Book.htm.

76 Powell A. Risks and benefits of an AI revolution in medicine. *Harvard Gazette* 2020; published online Nov 11. https://news.harvard.edu/gazette/story/2020/11/risks-and-benefits-of-an-ai-revolution-in-medicine/ (accessed April 21, 2025).

77 Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. 2016; published online June 18. DOI:10.48550/arXiv.1606.05718.

78 Litjens G, Kooi T, Bejnordi BE, *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017; **42**: 60–88.

79 Overfitting the Training Data. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp.30-33.

80 Logistic Regression. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp.164-173.

81 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**: 2825–30.

82 LogisticRegression. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Aug 1, 2025).

83 Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995: 278–82 vol.1.

84 Breiman L. Random Forests. *Machine Learning* 2001; **45**: 5–32.

85 RandomForestClassifier. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Aug 1, 2025).

86 Horos Project. Horos. https://horosproject.org.

87 Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

88 Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020; **585**: 357–62.

89 team T pandas development. pandas-dev/pandas: Pandas. 2020; published online Feb. DOI:10.5281/zenodo.3509134.

90 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**: 2825–30.

91 Phil T, Albrecht T, Gay S, Rasmussen ME. Sikerdebaard/dcmrtstruct2nii: v5. 2023; published online March 7. DOI:10.5281/zenodo.7705311.

92 Brys. icometrix/dicom2nifti. 2025; published online July 23. https://github.com/icometrix/dicom2nifti (accessed Aug 4, 2025).

93 The joblib developers. joblib. 2025; published online May 23. DOI:10.5281/ZENODO.15496554.

94 Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 2007; **9**: 90–5.

95 Haghighi S, Jasemi M, Hessabi S, Zolanvari A. PyCM: Multiclass confusion matrix library in Python. *JOSS* 2018; **3**: 729.

96 Darcy Mason, scaramallion, mrbean-bremen, *et al.* pydicom/pydicom: pydicom 3.0.1. 2024; published online Sept 22. DOI:10.5281/ZENODO.13824606.

97 Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**: 261–72.

98 Waskom M. seaborn: statistical data visualization. *JOSS* 2021; **6**: 3021.

99 Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *J Am Med Inform Assoc* 1997; **4**: 199–212.

100 Larobina M. Thirty Years of the DICOM Standard. *Tomography* 2023; **9**: 1829–38.

101 Cox RW, Ashburner J, Breman H, *et al.* A (sort of) new image data format standard: NiFTI-1. 2004.

102 Shrestha A, Watkins A, Yousefirizi F, Rahmim A, Uribe C. RT-utils: A Minimal Python Library for RT-struct Manipulation. *Journal of Open Source Software* 2025; **10**: 7361.

103 pydicom.filereader.dcmread — pydicom 3.0.1 documentation. https://pydicom.github.io/pydicom/stable/reference/generated/pydicom.filereader.dcmread.html#pydicom.filereader.dcmread (accessed Aug 4, 2025).

104 Usage — pyradiomics v3.1.0rc2.post5+g6a761c4 documentation. https://pyradiomics.readthedocs.io/en/latest/usage.html (accessed Aug 4, 2025).

105 Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp* 2002; **17**: 143–55.

106 Kanakaraj P, Yao T, Cai LY, *et al.* DeepN4: Learning N4ITK Bias Field Correction for T1-weighted Images. *Neuroinformatics* 2024; **22**: 193–205.

107 Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Trans Syst, Man, Cybern* 1973; **SMC-3**: 610–21.

108 Galloway MM. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* 1975; **4**: 172–9.

109 Thibault G, Fertil B, Navarro C, *et al.* SHAPE AND TEXTURE INDEXES APPLICATION TO CELL NUCLEI CLASSIFICATION. *Int J Patt Recogn Artif Intell* 2013; **27**: 1357002.

110 Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst, Man, Cybern* 1989; **19**: 1264–74.

111 Feature Scaling. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp.72-75.

112 Kim Y-S, Kim MK, Fu N, Liu J, Wang J, Srebric J. Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. *Sustainable Cities and Society* 2025; **118**: 105570.

113 Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 2002; **46**: 389–422.

114 Irrelavant Features. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: 30.

115 Regularized Linear Models. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp 155ff.

116 Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep* 2015; **5**: 13087.

117 Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 2002; **46**: 389–422.

118 Select and Train a Model. In: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, Third edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023: pp88ff.

119 Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J Mach Learn Res* 2010; **11**: 2079–107.

120 Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications* 2021; **182**: 115222.

121 Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* 2002; **99**: 6562–6.

122 Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights into Imaging* 2021; **12**: 172.

123 Parvandeh S, Yeh H-W, Paulus MP, McKinney BA. Consensus features nested cross-validation. *Bioinformatics* 2020; **36**: 3093–8.

124 Lewis MJ, Spiliopoulou A, Goldmann K, Pitzalis C, McKeigue P, Barnes MR. nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinform Adv* 2023; **3**: vbad048.

125 StratifiedShuffleSplit. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html (accessed Aug 1, 2025).

126 mannwhitneyu — SciPy v1.16.1 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.mannwhitneyu.html (accessed Aug 2, 2025).

127 kruskal — SciPy v1.16.1 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.kruskal.html (accessed Aug 2, 2025).

128 1.1. Linear Models. scikit-learn. https://scikit-learn/stable/modules/linear_model.html (accessed Aug 2, 2025).

129 GridSearchCV. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed Aug 1, 2025).

130 Generative Models for Classification. In: An introduction to statistical learning: with applications in Python. Cham, Switzerland: Springer, 2023: pp 146ff.

131 Powers D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach Learn Technol* 2008; **2**.

132 Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters* 2006; **27**: 861–74.

133 Tharwat A. Classification assessment methods. *Applied Computing and Informatics* 2020; **17**: 168–92.

134 auc. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.metrics.auc.html (accessed Aug 3, 2025).

135 pycm documentation. https://www.pycm.io/doc/index.html (accessed Aug 3, 2025).

136 vestibular schwannoma radiomics - Search Results - PubMed. https://pubmed.ncbi.nlm.nih.gov/?term=vestibular%20schwannoma%20radiomics&sort=date (accessed May 4, 2025).

137 acoustic neuroma radiomics - Search Results - PubMed. https://pubmed.ncbi.nlm.nih.gov/?term=acoustic%20neuroma%20radiomics&sort=date&page=2 (accessed May 4, 2025).

138 Zheng M, Sun X, Qi H, Zhang M, Xing L. Computed tomography-based radiomics and clinical-genetic features for brain metastasis prediction in patients with stage III/IV epidermal growth factor receptor-mutant non-small-cell lung cancer. *Thoracic Cancer* 2024; **15**: 1919–28.

7. Appendix

7.1. Supplementary Tables and Figures

Clinical Endpoint	**Progressive Disease**		**Pseudoprogression**		**Local Control**	
Model	Logistic Regression	Random Forest	Logistic Regression	Random Forest	Logistic Regression	Random Forest
AUC	0.80 ± 0.08	0.83 ± 0.16	0.67 ± 0.10	0.69 ± 0.10	0.64 ± 0.09	0.66 ± 0.09
F1-score	0.52 ± 0.13	0.60 ± 0.21	0.46 ± 0.10	0.51 ± 0.11	0.83 ± 0.01	0.83 ± 0.02
Accuracy	0.82 ± 0.13	0.86 ± 0.16	0.55 ± 0.23	0.59 ± 0.21	0.72 ± 0.03	0.72 ± 0.04
Sensitivity	0.80 ± 0.26	0.80 ± 0.26	0.82 ± 0.17	0.90 ± 0.13	1.00 ± 0.00	0.99 ± 0.02
Specificity	0.82 ± 0.17	0.87 ± 0.18	0.47 ± 0.33	0.51 ± 0.28	0.10 ± 0.09	0.12 ± 0.16
PPV	0.49 ± 0.29	0.59 ± 0.31	0.35 ± 0.15	0.37 ± 0.12	0.71 ± 0.02	0.71 ± 0.04
NPV	0.98 ± 0.03	0.98 ± 0.03	0.92 ± 0.05	0.96 ± 0.05	1.00 ± 0.00	0.95 ± 0.11

Table 16: Results of outer cross-validation for binary classification models for PD, PP and LC

7.2. List of Radiomics Features

7.2.1. Shape

Elongation, Flatness, LeastAxisLength, MajorAxisLength, Maximum2DDiameterColumn, Maximum2DDiameterRow, Maximum2DDiameterSlice, Maximum3DDiameter, MeshVolume, MinorAxisLength, Sphericity, SurfaceArea, SurfaceVolumeRatio, VoxelVolume

7.2.2. Firstorder Statistics

10Percentile, 90Percentile, Energy, Entropy, InterquartileRange, Kurtosis, Maximum, MeanAbsoluteDeviation, Mean, Median, Minimum, Range, RobustMeanAbsoluteDeviation, RootMeanSquared, Skewness, TotalEnergy, Uniformity, Variance

7.2.3. GLCM

Autocorrelation, ClusterProminence, ClusterShade, ClusterTendency, Contrast, Correlation, DifferenceAverage, DifferenceEntropy, DifferenceVariance, Id, Idm, Idmn, Idn, Imc1, Imc2, InverseVariance, JointAverage, JointEnergy, JointEntropy, MCC, aximumProbability, SumAverage, SumEntropy, SumSquares

7.2.4. GLDM

DependenceEntropy, DependenceNonUniformity, DependenceNonUniformityNormalized, DependenceVariance, GrayLevelNonUniformity, GrayLevelVariance, HighGrayLevelEmphasis, LargeDependenceEmphasis, LargeDependenceHighGrayLevelEmphasis, LargeDependenceLowGrayLevelEmphasis, LowGrayLevelEmphasis, SmallDependenceEmphasis, SmallDependenceHighGrayLevelEmphasis, SmallDependenceLowGrayLevelEmphasis

7.2.5. GLRLM

GrayLevelNonUniformity, GrayLevelNonUniformityNormalized, GrayLevelVariance, HighGrayLevelRunEmphasis, LongRunEmphasis, LongRunHighGrayLevelEmphasis, LongRunLowGrayLevelEmphasis, LowGrayLevelRunEmphasis, RunEntropy, RunLengthNonUniformity, RunLengthNonUniformityNormalized, RunPercentage,

RunVariance, ShortRunEmphasis, ShortRunHighGrayLevelEmphasis, ShortRunLowGrayLevelEmphasis

7.2.6. GLSZM

GrayLevelNonUniformity, GrayLevelNonUniformityNormalized, GrayLevelVariance, HighGrayLevelZoneEmphasis, LargeAreaEmphasis, LargeAreaHighGrayLevelEmphasis, LargeAreaLowGrayLevelEmphasis, LowGrayLevelZoneEmphasis, SizeZoneNonUniformity, SizeZoneNonUniformityNormalized, SmallAreaEmphasis, SmallAreaHighGrayLevelEmphasis, SmallAreaLowGrayLevelEmphasis, ZoneEntropy, ZonePercentage, ZoneVariance

7.2.7. NGTDM

Busyness, Coarseness, Complexity, Contrast, Strength

7.3. Code blocks

```
# import the necessary libraries
from sklearn.linear_model import LogisticRegression

# defines the model that is used for training:
model = LogisticRegression()

# fits the model to the example data
model.fit(X, y)
```

Code Block 1: training a logistic regression model in scikit learn

```
from sklearn.linear_model import LogisticRegression

'''
We assume, that a logistic regression model has been trained that we now call
estimator here.
In the prior workflow it has been trained like this:
model = LogisticRegression()
estimator = model.fit(X, y)
'''

threshold = 0.45

# we insert the X of a new observation
y_score = estimator.predict_proba(X)

# check if the y_score exceeds the threshold to make the prediction y_hat
if y_score > threshold:
    y_hat = 'category 1'
else:
    y_hat = 'category 0'
```

Code Block 2: making predictions using a trained logistic regression model with a custom threshold.

```
import pandas as pd
from sklearn.metrics import roc_curve
from sklearn.preprocessing import label_binarize

# calculate the y_score(s) from the estimator
y_scores = estimator.predict_proba(X)

# detect binary classification
if len(list(estimator.classes_)) == 2:
    # roc curve is calculated
    fpr, tpr, thresh = roc_curve(y, y_scores[:,0], pos_label=classes[0])
    # Youden's J is defined
    j_scores = tpr - fpr
    # threshold for maximum Youden's J
    threshold =  thresh[j_scores.argmax()]

# detect multinomial
if len(list(estimator.classes_)) > 2:
    # categories are one-hot encoded/binarized
    y_true_bin = label_binarize(y, classes=classes)  # shape: (n_samples,
n_classes)

    # a threshold is set for every class
    thresholds = []
    for i in range(len(classes)):
        fpr, tpr, thresh = roc_curve(y_true_bin[:, i], y_scores[:, i])
        j_scores = tpr - fpr
        thresholds.append(thresh[j_scores.argmax()])
```

Code Block 3: Set thresholds for binary and multinominal classification for the maximum Youden's J

```
# import the necessary libraries
from sklearn.ensemble import RandomForestClassifier

# defines the model that is used for training:
model = RandomForestClassifier()

# fits the model to the example data
model.fit(X, y)
```

Code Block 4: Training a random forest algorithm with scikit-learn

```
#gridsearch for optimizing hyperparameters in Logistic Regression
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier

# declaring the grid of values for the hyperparameters
param_grid = {
    'n_estimators': [50, 100, 200],     # Number of trees
    'max_depth': [10, 20, 30],          # Control overfitting
    ...,
}

# defining the model
model = RandomForestClassifier()

# setting up GridSearchCV
grid_search = GridSearchCV (
```

```
    estimator   = model,      # inserting the defined model
    param_grid  = param_grid,  # inserting the parameter from above
)

# running the GridSearchCV algorithm
grid_search.fit(X, y)
```

Code Block 5: Usage of scikit-learn's GridSearchCV for a random forest algorithm.

```
import dicom2nifti

# conversion of a dicom image, stored in a directory to a nifti file
dicom2nifti.convert_directory("/dicom_directory", "nifti_filepath")
```

Code Block 6: converting dicom images to nifti using dicom2nifti

```
from dcmrtstruct2nii import dcmrtstruct2nii

# list of possible names for the VOI
structures_of_interest = ["VS", "tumor", ... ]

# converting the mask, aligned to the CT to a nifti file
dcmrtstruct2nii("/dicom_mask_filepath", "dicom_ct_directory", "nifti_filepath",
structures=valid_structs, gzip=True)
```

Code Block 7: converting rtss masks to nifty using dcmrtstruct2nii

```
import radiomics # imports pyradiomics

# set up a radiomics extractor
extractor = radiomics.featureextractor.RadiomicsFeatureExtractor()

# perform feature extraction
result = extractor.execute("/nifti_filepath.gz.nii",
"/nifti_mask_filepath.gz.nii")
```

Code Block 8: Feature extraction using PyRadiomics[20]

```
import numpy as np
import pandas as pd

# creation of empty normalization matrix
normalization_matrix = pd.DataFrame({'feature': X.columns.tolist()})
normalization_matrix = normalization_matrix.set_index('feature')

# calculation of the standard deviation s and mean u for each feature and writing
it to the matrix
for feature in X.columns:
    s = np.std(X[feature])
    u = np.mean(X[feature])
    normalization_matrix.loc[feature,'std'] = s
    normalization_matrix.loc[feature, 'mean'] = u

# applying the normalization for each feature
for feature in X.columns:
    if normalization_matrix.loc[feature, 'std'] != 0:
```

```
        X[feature] = (X[feature] - normalization_matrix.loc[feature, 'mean'] ) /
normalization_matrix.loc[feature, 'std']
```

Code Block 9: Creation and application of a normalization matrix for z-score normalization

```
import pandas as pd

# define a r cutoff
r_cutoff = 0.9

# compute correlation matrix
# with spearman correlation coefficient r
# abs() creates the absolute value
cm = abs(X.corr(method='spearman'))

# lists for selected an unselected features
unselected = []
selected = []

# iterates through all features
for featA in cm.columns:
    # skips if feature A already unselected
    if featA not in unselected:
        # iterates through features again for each feature A
        for featB in cm.columns:
            # skips if feature B already unselected
            if featB not in unselected:
                # checks if features A and B are not the same (r would be 1)
                # and if r cutoff is surpassed
                if featA != featB and cm[featA][featB] > r_cutoff:
                    # if both is true feature B is unselected
                    unselected.append(featB)

# all not unselected features are listed as selected
for f in cm.columns:
    if f not in unselected:
        selected.append(f)
```

Code Block 10: feature selection based on cutoff for spearman correlation coefficient r

```
from sklearn.model_selection import StratifiedShuffleSplit

# creating the split
split = StratifiedShuffleSplit(n_splits=10, test_size=0.2, random_state=42)

# slicing the features X and clinical data y accordingly
for i, (train_index, test_index) in enumerate(split.split(X, y)):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
```

Code Block 11: Usage of StratifiedShuffleSplit

```
import pandas as pd
from scipy.stats import mannwhitneyu, kruskal

def compute_p(X, y):

    # combines X and y to be able to group by the categories of y
```

```
    # X only contains one feature here
    Xny = pd.DataFrame() # X and y
    Xny['X'] = X
    Xny['y'] = y

    # get a list of the classes
    classes = y.unique().tolist()

    #make a list of groups with data for only 1 class
    groups = []
    for c in classes:
        group = Xny[Xny['y']==c].drop('y', axis=1)
        groups.append(group)

    # chooses the right statistical test based on the number of classes
    if len(groups) == 2:
        stat, p = mannwhitneyu(*groups)
    elif len(groups) > 2:
        stat, p = kruskal(*groups)

    return p
```

Code Block 12: Calculation of MWU and KW p-values for feature selection

```
import pandas as pd

# define p-value-cutoff
pco = 0.05

# create dictionary for features and p-values
p_values = {}

# iterate through the columns (=features) in X
for col_name, col_data in X.items():
    p = compute_p(
        X = col_data,
        y = y,
    )
    # add p-value to a dicts with all features and p-values
    p_values[col_name] = p

#select features
sel_feat = []
for f, p in p_values.items():
    if p <= pco:
        sel_feat.append(f)

# create new dataframe with selected features only
X_selected = X [sel_feat]
```

Code Block 13: selecting features based on p-values

```
import pandas as pd
from lifelines import CoxPHFitter

def compute_cox_p(X, y):
    # combines X and y to be able to group by the categories of y
    # X only contains one feature here
    Xny = y[['time','status']].merge(X, left_index=True, right_index=True,
```

```
        how='inner')

    # fitting of the cox’s proportional hazard model
    cph = CoxPHFitter()
    cph.fit(Xny, duration_col='time', event_col='status')

    # extract p-value from the results
    p = cph.summary[['p']].iat[0, 0]

    return p
```

Code Block 14: calculation of p-values from cox's proportional hazard model

```
import pandas as pd
from sklearn.linear_model import LogisticRegression

# defining the number of samples to select:
# square root of number of patients, but at least 4
k = len(y) ** 0.5
k = int(k)
if k<4:
    k=4

#define and fit model
model = LogisticRegression(penalty = 'elasticnet', l1_ratio = 0.5, solver =
‘saga’)
model.fit(X, y)

# rank features by absolute coefficient magnitude
feature_names = X.columns.tolist()
abs_coefs = np.abs(model.coef_[0])
coefs = pd.DataFrame(abs_coefs, index=feature_names).sort_values(ascending=False,
by=0)

# choose the k features with the highest coefficients
selected_features = coefs[:k].index.tolist()

# transform X
X_selected = X[selected_features]
```

Code Block 15: feature selection based on coefficients of logistic regression

```
import pandas as pd
from sklearn.feature_selection import RFECV
from sklearn.ensemble import RandomForestClassifier

# define model
model = RandomForestClassifier()

# set up and fit rfecv
rfecv = RFECV(model, step=1, cv=5)
rfecv.fit(X, y)

# transform X by using the boolean mask returned by rfecv
feature_bool = rfecv.get_support()
for i, feature in enumerate(X.columns):
    if not feature_bool[i]:
        X.drop(feature, axis=1, inplace=True)
```

Code Block 16: RFECV with random forest

```
# used for logistic regression
grid_logistic_regression = {
    'C': [0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100,
300, 1000, 3000],
    'l1_ratio': [0.0, 0.3, 0.5, 0.7, 1.0],
}

# used for cox regression
grid_cox_regression = {
    'penalizer': [0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0],
    'l1_ratio': [0.0, 0.3, 0.5, 0.7, 1.0],
}

# used for both, classification and survival
grid_random_forest = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
}
```

Code Block 17: Hyperparameter grids for logistic regression, random forest classification, cox-regression, and random forest survival

```
import pandas as pd
# import the respective model class e.g. LogisticRegression

y_scores = estimator.predict_proba(X)  # [:1]
# detect binary classification
if len(list(estimator.classes_)) == 2:
    # make prediction, returns 0 or 1
    y_hat_bin = (y_scores[:, 0] >= threshold).astype(int)
    # retranslate in category name
    y_hat = [estimator.classes_[i] for i in y_hat_bin]

# detect multinomial
if len(list(estimator.classes_)) > 2:
    y_hat = []
    for row in y_scores:
        # Check if any class exceeds its own threshold
        above_threshold = [i for i, p in enumerate(row) if p >= thresholds[i]]
        if above_threshold:
            # Pick highest
            best = max(above_threshold, key=lambda i: row[i])
            y_hat.append(estimator.classes_[best])
        else:
            # Fallback to highest score
            y_hat.append(estimator.classes_[row.argmax()])
```

Code Block 18: making predictions from estimator with predefined thresholds

```
from joblib import Parallel, delayed

# list of datasets we want to process
dataset_list = [...]

# function that we want to use for all datasets
```

```
def example_function(ds):
    # some expensive calculations happen here
    ...
    return new_ds

# parallel processing using all available CPU cores is performed
new_data = Parallel(n_jobs=-1)(
    delayed(example_function)(ds) for ds in dataset_list
)
```

Code Block 19: Parallelization using joblib[93]

7.4. Figures

Figure 1: Microscopic appearance of a vestibular schwannoma. A cell-rich Antoni A region is present on the right side, while a comparatively loosened, cell-poor Antoni B region is visible on the left side (Hematoxylin-Eosin, x400)[24] 15

Figure 2: Contrast enhanced T1-weighted MRI depicting the classification of VS as described by Koos[25].....16

Figure 3: Contrast enhanced T1 MRI of a VS treated with SRS. a) MR image for treatment planning, b) manual segmentation, c) radiation planning with isodoses displayed, d) 6 month follow up MRI, e) 20 month follow up MRI 19

Figure 4: Development of the VS volume after SRS in the case of Fig. 3 19

Figure 5: Radiomics workflow.....22

Figure 6: VS of different shapes, sizes and textures.....23

Figure 7: Plot of a sigmoid function.....26

Figure 8: Sigmoid function with added example data.....27

Figure 9: Example data with sigmoid function of a trained model 28

Figure 10: The radiomics hypothesis.....31

Figure 11: Radiomics Workflow 37

Figure 12: Machine learning workflow 39

Figure 13: Machine learning workflow for logistic regression 40

Figure 14: Machine Learning Workflow for Random Forest 40

Figure 15: Correlation matrix for features of the contrast enhanced T1 sequence 42

Figure 16: Confusion matrix.....46

Figure 17: Pretherapeutic tumor volume distribution (left) and Koos Grade (right).....48

Figure 18: Follow-up of volume changes of individual cases (left) and mean follow up volume changes ± standard deviation (right) of all included cases.....49

Figure 19: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in cases that later developed PD 49

Figure 20: Follow-up volumes of individual cases (left) and mean follow up volume ± standard deviation (right) of PD 50

Figure 21: ROC AUC curve of the PD binary response model using random forest 50

Figure 22: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in included cases of PP 52

Figure 23: Follow-up volume changes of individual cases (left) and mean follow up volume changes ± standard deviation (right) in cases of PP 52

Figure 24: ROC AUC curve PP binary response model using random forest 53

Figure 25: Pretherapeutic tumor volume distribution (left) and Koos Grade (right) in included cases of LC 54

Figure 26: Follow-up volumes of individual cases (left) and mean follow up volume ± standard deviation (right) of LC 54

Figure 27: ROC AUC curve LC binary response model using random forest 55

Figure 28: ROC AUC curve for the 3-C multiclass response model using logistic regression 56

Figure 29: ROC AUC scores for Progressive Disease (PD), Pseudoprogression (PP) and Local Control (LC) 58

Figure 30: Time points when individual cases of PD first surpassed +40% 59

Figure 31: Follow-up volumes of individual cases of early PP (blue) and late PP (orange) ... 60

7.5. Tables

Table 1: Incidence and duration of symptoms of VS at time of diagnosis [17] 16

Table 2: Modalities of the planning MRI.....33

Table 3: Definition of imaging endpoints.....34

Table 4: Used Python libraries.....35

Table 5: Individual and combined modality groups 39

Table 6: Patient characteristics.....48

Table 7: Results of the outer cross-validation for binary classification models for PD 51

Table 8: The ten most frequently selected features of the reported binary PD model 51

Table 9: Results of outer cross-validation for binary classification models for PP 53

Table 10: The ten most frequently selected features of the reported binary PP model.....54

Table 11:Results of outer cross-validation for binary classification models for LC 55

Table 12: The ten most frequently selected features of the reported binary LC model.....56

Table 13: Results of outer cross-validation for 3-C 57

Table 14: The ten most frequently selected features of the reported 3-C model 57

Table 15: Reported classifier metrics in prior publication regarding volumetric response of VS to SRS. Not reported is indicated by '-'. 63

Table 16: Results of outer cross-validation for binary classification models for PD, PP and LC 75