TU WIEN Informatics

Predicting Tumor-Infiltrating Lymphocytes for Glioblastoma using Radiomics and Deep Learning Approaches

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Markus Zvonek, BSc.
Matrikelnummer 01525726

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr. Robert Sablatnig
Mitwirkung: Univ.Prof. Dipl.-Ing. Dr. Georg Langs

Wien, 6. Oktober 2022

Markus Zvonek	Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Predicting Tumor-Infiltrating Lymphocytes for Glioblastoma using Radiomics and Deep Learning Approaches

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

Markus Zvonek, BSc.
Registration Number 01525726

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr. Robert Sablatnig
Assistance: Univ.Prof. Dipl.-Ing. Dr. Georg Langs

Vienna, 6th October, 2022

Markus Zvonek | Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Markus Zvonek, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 6. Oktober 2022

Markus Zvonek

Danksagung

Mit dem Ende dieser Arbeit ist es an der Zeit, mich bei denen zu bedanken, die diese erst möglich gemacht und unterstützt haben.

Als Erstes möchte ich mich bei Georg Langs von der Medizinischen Universität Wien bedanken, der diese Arbeit von Anfang an betreut hat. Danke für all die Anregungen und die ganze Unterstützung, die zur Fertigstellung dieser Arbeit beigetragen haben, sowie für die Möglichkeit in dem spannenden Umfeld der medizinischen Bildverarbeitung eine Arbeit schreiben zu können.

Auch Robert Sablatnig von der Technischen Universität Wien möchte ich an dieser Stelle meinen Dank. Er hat nicht nur meine Arbeit offiziell betreut und mir mit wertvollem Feedback weitergeholfen, sondern auch für sein Seminar *Scientific Presentation and Communication*, das mir beim wissenschaftlichen Schreiben geholfen hat.

Ein großes Dankeschön an die Kollegen vom CIR Lab an der Medizinischen Universität Wien für die Hilfe bei dieser Arbeit. Besonders Karl-Heinz ist mir mit Rat und Tat von der ersten Stunde an zur Seite gestanden, und auch Christoph hat mir den Weg aus so mancher Sackgasse gezeigt.

Bedanken möchte ich mich auch bei den Ärzten des AKH Wiens, die nicht nur die Daten zur Verfügung gestellt haben, sondern mir auch geholfen haben, wo es nur möglich war. Vielen Dank Anna, Julia und Maximilian für all die Hilfe.

Ein riesengroßer Dank gebührt meiner Familie, allem voran meinen Eltern, die mich während des ganzen Studiums unterstützt, auch dabei diese Arbeit zu absolvieren. Auch meinem Bruder und meinen Freunden gebührt ein großer Dank, die während dem Studium und dieser Arbeit immer wieder für Motivation und auch Ablenkung gesorgt haben.

Acknowledgements

At the end of this work, I would like to thank those who supported me throughout this thesis.

First and foremost, a big thank you to Gerog Langs from the Medical University of Vienna, who has been my advisor for this thesis. Thanks for all the ideas, your support, and the possibility to write my master thesis in the fascinating area of medical image analysis.

I would also like to thank Robert Sablatnig of TU Wien, my official advisor for this thesis. His feedback help me a lot throughout my thesis, while his seminar *Scientific Presentation and Communication* improved my scientific writing skills a lot.

A big thank you to my colleagues at the CIR lab of the Medical University of Vienna for all their support. A special thank you goes to Karl-Heinz who provided a lot of help from the very first hour, as well as to Christoph for guiding me out of some deadlocks.

Big thanks to the clinicians of the University Hospital Vienna for providing the data and their continuous support. Thanks Anna, Julia, and Maximilian.

A giant thank you goes to my family, especially my parents, who supported me throughout my studies including this thesis. I would like to thank my brother and my friends here as well, they did not only support me from the first minute of my studies and this master thesis, but motivated me and provided some necessary distraction along the way.

Kurzfassung

Das Glioblastoma multiforme ist der tödlichste primäre Hirntumortyp. Obwohl Immuntherapien eine Behandlungsmethode für andere Krebsarten sind, gibt es sie nicht für das Glioblastom. Die Forschung hat gezeigt, dass die immunologische Mikroumgebung eine entscheidende Rolle bei erfolgreichen Immuntherapien spielt. Tumor-infiltrierende Lymphozyten sind ein wesentlicher Bestandteil der Mikroumgebung und für Immuntherapien von Bedeutung.

Diese Arbeit beschreibt zwei Ansätze zur Vorhersage solcher Tumor-infiltrierender Lymphozyten aus in vivo Magnetresonanz-Volumina. Die verwendeten Methoden extrahieren Informationen aus manuell segmentierten Regions-of-Interest, um maschinelle Lernmodelle zu erstellen, die die extrahierten Merkmale mit Markern der Tumor-infiltrierenden Lymphozyten in Verbindung bringen. Der erste Ansatz nutzt radiomische Merkmale und verwendet elastische Netze und Random Forests. Der zweite Ansatz verwendet ein modifiziertes ResNet50 als Deep-Learning-Komponente.

In den Experimenten werden 56 bis 88 Sätze von Magnetresonanz-Volumina und Tumor-Infiltrations-Lymphozyten-Markern verwendet, um die Methoden zu trainieren und zu bewerten. In einer quantitativen Analyse werden die Korrelationen zwischen den Werten der Grundwahrheit der Tumor-Infiltrations-Lymphozyten-Marker und den vorhergesagten Werten untersucht. Die qualitative Analyse bewertet die Stabilität der ausgewählten prädiktiven Merkmale und die Herkunft der am besten prädiktiven Merkmale.

Die Ergebnisse zeigen, dass der Radiomics-Ansatz einige Tumor-Infiltrations-Lymphozyten-Marker auf Grundlage der Magnetresonanz-Volumina vorhersagen kann, aber nicht alle. Die Auswahl der prädiktiven Merkmale ist stabil, während sich einige der prädiktiven Merkmale auf bestimmte Teile des Glioblastoms konzentrieren. Der Deep-Learning-Ansatz hingegen kann die Zielwerte für die Testdaten nicht vorhersagen.

Abstract

Glioblastoma multiforme is the most lethal primary brain tumor type. Although immunotherapies are a treatment method for other cancer types, they are not available for glioblastoma. Research has revealed that the immune microenvironment plays a crucial role in successful immunotherapies. Tumor-Infiltrating Lymphocytes are an essential part of the microenvironment and are of high significance for immunotherapies.

This thesis describes two approaches for predicting such Tumor-Infiltration Lymphocytes from in vivo magnetic resonance imaging data. The methods used extract information from manually segmented Regions-of-Interest to build machine learning models that associate the features extracted from images with markers of the Tumor-Infiltrating Lymphocytes. To this end, the first approach utilizes radiomics features, elastic nets regression, and random forests. The second approach uses a modified ResNet50 as a deep learning component for prediction.

The experiments use 56 to 88 sets of magnetic resonance volumes and Tumor-Infiltration Lymphocytes markers to train and evaluate the methods. A quantitative analysis investigates the correlations between the ground truth values of the Tumor-Infiltration Lymphocytes markers and the predicted values. The qualitative analysis evaluates the stability of the predictive features chosen and the origin of the most predictive features.

Results show that the radiomics approach can predict some Tumor-Infiltration Lymphocytes markers based on the magnetic resonance volumes, but not all of them. The choice of the predictive features is stable, while some of the predictive features' origins focus on particular parts of the glioblastoma. The deep learning approach cannot predict the target values for the test data in our experiments.

Contents

Acronyms

10-FCV	10-Fold Cross-Validation
CART	Classification And Regression Trees
CD3+	Cluster of Differentiation 3
CD8+	Cluster of Differentiation 8
CNN	Convolutional Neural Network
CT	Computer Tomography
FLAIR	FLuid Attenuated Inversion Recovery
GBM	GlioBlastom Multiforme
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
GradCAM	Gradient-weighted Class Activation Mapping
LASSO	Least Absolute Shrinkage and Selection Operator
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
MRI	Magnetic Resonance Imaging
NGTDM	Neighboring Gray Tone Difference Matrix
PD1+	Programmed cell Death-1
PET	Positron Emission Tomography
ROI	Region-Of-Interest
T1c	T1 Contrast-enhanced
TIL	Tumor-Infiltrating Lymphocyte

Mathematical Notations

n	Number of samples/target values
j	Sample index $j \in 1, \ldots, n$
$\mathbf{I}_j^{acquired}$	Image acquired for sample j
$\mathbf{I}_j^{preproc}$	Image preprocessed for sample j
$\vec{\mathbf{x}}$	All samples
$\vec{\mathbf{x}}_{train}$	Training samples
$\vec{\mathbf{x}}_{test}$	Test samples
y_j	Ground truth target value for sample j, $y \in \mathbb{R}$
y_j^{den}	Ground truth value for a *Density* TIL marker, e.g. *CD3+ Density*, of sample j
y_j^{pos}	Ground truth value for a *Positive* TIL marker, e.g. *CD3+ Positive*, of sample j
y_j^{neg}	Ground truth value for a *Negative* TIL marker, e.g. *CD3+ Negative*, of sample j
y_j^{per}	Ground truth value for a *Percentage* TIL marker, e.g. *CD3+ Percentage*, of sample j
$\vec{\mathbf{y}}$	All ground truth target values
$\hat{y}_j$	Predicted target value for sample j
$\vec{\hat{\mathbf{y}}}$	All predicted target values
$\mathbf{D} = (x_1, y_1), \ldots, (x_n, y_n)$	Data set containing pairs of samples and target values
$\mathbf{D}_k^{train}$	Training set for fold k of cross-validation
$\mathbf{D}_k^{val}$	Validation set for fold k of cross-validation
$\mathbf{D}_{NN}^{train}$	Deep Learning training set
$\mathbf{D}_{NN}^{test}$	Deep Learning test set
f_{pp}	Image preprocessing steps for sample j
f_{fe}	Radiomics feature extraction
$\vec{\mathbf{x}}_j$	Features extracted from sample j
$\vec{\mathbf{x}}_j^{radiomics}$	Radiomics features extracted for sample j
m	Number of features in a feature vector
h	Feature index $h \in 1, \ldots, m$
$x_{j,h}$	Feature at index h of sample j, $x_{j,h} \in \vec{\mathbf{x}}_j$
$f_{ed}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_l)$	Euclidean distance between the feature vectors of samples j and l, with $j \neq l$

$f^{pred}_{radiomics}$	Radiomics prediction function
f^{learn}	Training a machine learning model
$\mathbf{m}_{trained}$	A trained machine learning model
f^{eval}	Evaluating a machine learning model
f^{learn}_{RF}	Training a Random Forest model
$\mathbf{m}_{RF}$	A trained Random Forest model
f^{eval}_{RF}	Evaluating a Random Forest model
t	Number of trees in a Random Forest
u	Index of tree in a Random Forest, $u \in 1, \ldots, t$
n_u	Number of samples chosen for tree u of a Random Forest
$\hat{y}_u$	Prediction of tree u of a Random Forest
$\vec{\mathbf{x}}_u$	Sample chosen for tree u of a Random Forest
$n_{nodeSize}$	Number of samples where the tree of a Random Forest stops splitting
f^{train}_{EN}	Training an Elastic Net model
$\mathbf{m}_{EN}$	A trained Elastic Net model
f^{eval}_{EN}	Evaluating an Elastic Net model
$\hat{\beta}$	Elastic net coefficients vector
$\hat{\beta}^*$	Corrected elastic net estimator
$\mathbf{m}_{NN}$	Deep Learning model
g	Cost function
b	Point along a function
w	Convolution kernel of a CNN
s	Result of a convolutional layer in a CNN
$\mathbf{I}^{NN}_j$	Input to CNN Layer of sample j
f^{learn}_{NN}	Training a CNN
σ	Standard deviation
μ	Average
σ_{tm}	Standard deviation of ground truth values for TIL marker tm
μ_{tm}	Mean of ground truth values for TIL marker tm
MAE	Mean Absolute Error
r	Correlation coefficient according to Spearman
p	Correlation's significance according to Spearman

CHAPTER 1

Introduction

This first chapter presents the motivation and problem statement of this thesis. Subsequently, this chapter also gives the motivation from a medical point of view and the aims of this thesis. In the end, this chapter provides the structure of this work.

1.1 Motivation

Cancer is a disease [100] that is investigated in a broad range of studies, such as, [21], [46], [55] to name a few. A cell whose growth and spread becomes uncontrollable marks the beginning of cancer [100]. Figure 1.1 depicts the percentage of total death causes worldwide for 2019, based on the data published by the World Health Organization (WHO) [68], and shows that cancer is the second most frequent cause of death.

Cancer can originate from almost every tissue in the body [100], and the brain is no exception to that [21]. The so-called GlioBlastom Multiforme (GBM) is a type of brain cancer with a median survival duration of 12 to 15 months [21]. The tumor's rapid growth and aggressive, neurologically destructive, and highly invasive behavior is a major cause of the short survival duration [61]. Due to their characteristics, GBMs are among the most lethal cancers in humans [61].

Figure 1.2 presents the most lethal cancers worldwide for 2019, based on the data of the WHO [68]. As Figure 1.2 indicates, GBMs (and brain cancers in general) are not among the most common types of malignant neoplasms, but their characteristics cause them to be among the deadliest [61].

The molecular heterogeneity of GBMs is one of the significant obstacles for treatments trying to improve the survival duration [21]. The heterogeneity limits molecular biomarkers from representing all biological activities of the entire tumor since such a biomarker usually originates from a single confined subregion of the GBM [21].

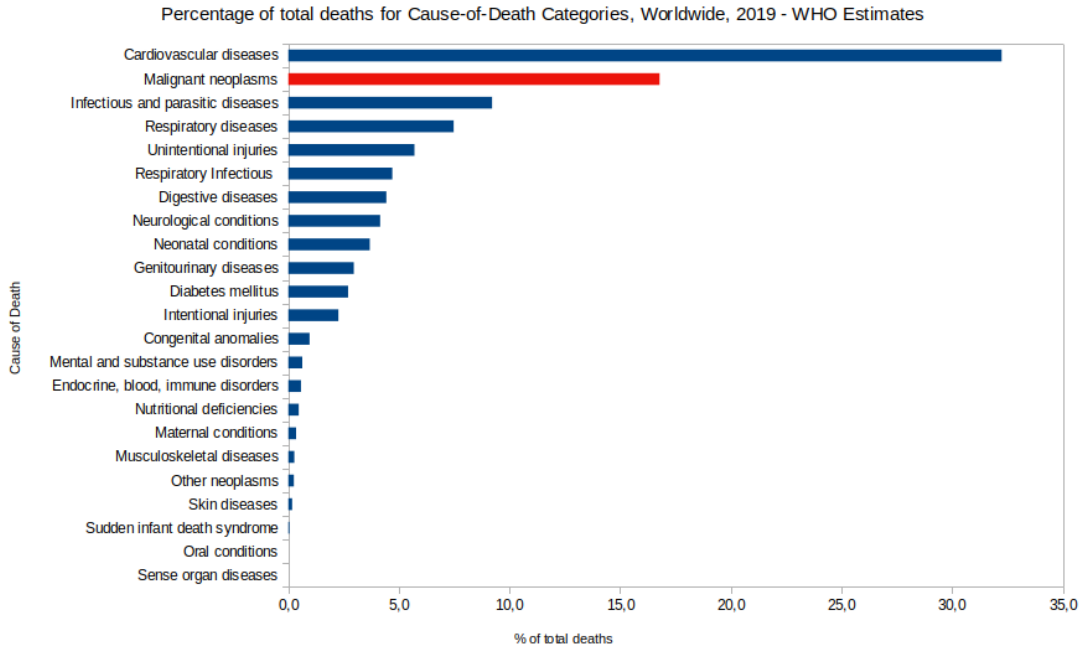


Figure 1.1: This chart shows the worldwide estimated numbers for the death causes categories for 2019 (data published by the WHO [68]). The bar for cancers is highlighted in red since GBMs belong to this category.

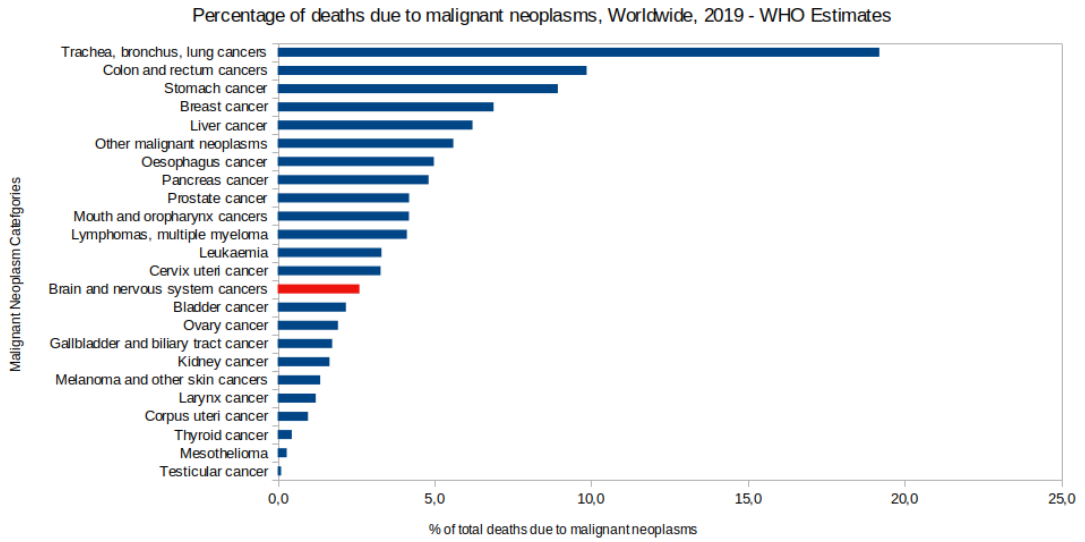


Figure 1.2: This chart shows the worldwide estimated numbers of deaths due to different malignant neoplasms for 2019 (data published by the WHO [68]). The bar for brain cancers is highlighted in red since GBMs belong to this category.

Using medical imaging (or information derived from the images) as a surrogate for the tumor has the potential to solve the challenge of heterogeneity [106]. Medical imaging, usually Magnetic Resonance Imaging (MRI), is the standard clinical practice for diagnosing GBMs and monitoring the patients' treatment [21]. It captures the tumor in vivo, avoiding surgical risks, which makes medical imaging a safer approach [21]. In addition, it is possible to acquire medical images multiple times over the course of the diagnosis and treatment, which allows tracking of the disease's progression [106]. Medical imaging also captures the entire tumor's radiological heterogeneity and the surrounding environment, compared to molecular analysis of the tissue performed on a local sample [106].

1.2 Problem Statement: Determining Tumor Characteristics from Imaging Data

This thesis aims at determining tumor characteristics that are relevant for treatment decisions from in vivo imaging data.

Clinicians treat GBMs with the commonly known treatment methods for cancers, i.e., surgery, radiation therapy, and chemotherapy [90]. However, Thomas et al. point out that these treatment methods do not boost the survival prediction [90]. Immunotherapy is a different mechanistic approach, and immunotherapy treatments are successful for other cancer types [90]. However, studies about immunotherapies for GBMs indicate that patients do not benefit from immunotherapies treatment [64]. But such studies also reveal that this treatment method is successful for a small subpopulation [76], [102]. However, the exact reasons for the existence of these subpopulations remain undetermined [76], [102]. This fact only emphasizes that further research regarding immunotherapies as a treatment for GBMs is needed. A first step can be taken by investigating and identifying the receptive subpopulation.

A relation between Tumor-Infiltrating Lymphocytes (TILs) and the information derived from the medical images may identify the small subpopulation. If they are related, TILs could be predicted based on the information derived from the medical imaging data, similar to how Wu et al. [106] predict the TILs for breast cancer. Later research could focus on the relations between the TILs and the success of immunotherapies. In the ideal case, identifying the small subpopulation and separating them from the other patients will be possible with the findings of this work. In this case, TILs based on the medical imaging information are valid biomarkers. In the long run, a relation between the medical imaging information of GBMs and TILs could be the first step to successful immunotherapy for such a lethal cancer type as GBMs.

TILs are of interest since they are a significant factor in the immune microenvironment, which is important for successful immunotherapy [62]. In addition, TILs are widely tested as biomarkers [106]. As Wu et al. point out, the guideline commonly used to determine TILs states that pathologists count within the stromal regions to obtain the

TIL values [106]. Despite the guideline, there is no way to optimally assess TILs under specific clinical scenarios as biopsy inevitably introduces a sampling bias, and estimating TILs remains subjective [106]. An evaluation of a one-time point biopsy hardly represents the dynamic evolution of the microenvironment, but repeated biopsies are impractical and put the patients at risk of complications [106]. "*Therefore, a non-invasive biomarker that can assess and monitor the tumor immune contexture by in vivo imaging would overcome the aforementioned hurdles and be invaluable for patient management*" ([106], p. 311).

1.3 Aims of the Work

This study's has four aims:

1. Prediction of TIL values from imaging data containing GBMs and evaluation of accuracy for different TIL values.
2. Comparison of radiomics and deep learning approaches for the prediction of TIL values.
3. Identification of visual signatures of TILs that have an association with the imaging data.
4. Comparison of features extracted from different regions of interest.

All in all, this study's general aim is to investigate if it is possible to predict values of TILs, specifically Cluster of Differentiation 3 (CD3+), Cluster of Differentiation 8 (CD8+), and Programmed cell Death-1 (PD1+), with information derived from the imaging data. This study investigates the TILs separately. Due to this, a possible outcome can be that CD8+ values are predictable, while values of PD1+ are not.

1.4 Structure of the Work

The outline of the thesis is as follows:

- **Chapter 1** introduces the field of this thesis and provides an overview of the motivation, problem statement, and thesis' aims.
- **Chapter 2** provides background information on the medical aspects of this work, focusing on GBMs and TILs.
- **Chapter 3** introduces the concept of radiomics and discusses the steps of the method in detail.

- **Chapter 4** reviews relevant basic machine learning methods and introduces the ElasticNet and Random Forest methods in detail. This chapter also discusses validation approaches.
- **Chapter 5** discusses the background information about deep learning. The explanation focuses on neural networks and Convolutional Neural Networks (CNN)s more specifically.
- **Chapter 6** provides the related work of this thesis. This chapter focuses on studies about radiomics with GBMs, deep learning with GBMs, and the prediction of TILs.
- **Chapter 7** describes the main contribution of this thesis, methods to estimate values of different TILs from MRI images.
- **Chapter 8** reports the experiments' results for the radiomics and the deep learning approach.
- **Chapter 9** discusses the findings and results of the experiments. In addition, the research question is answered in this chapter as well.
- **Chapter 10** gives the conclusion of this work and provides insight into future work.

CHAPTER 2

Medical Background

This chapter describes the medical background of this thesis. At the start, the description focuses on GBMs. Afterward, this chapter provides the relationship between cancers and the immune system. The end of the chapter gives the medical background of TILs, including the types of TILs used in this work.

2.1 Glioblastoma Multiforme

Cushing introduced the term GBM in the second half of the nineteenth century, while in Vienna in 1904, the first operation on a patient suffering from this tumor was performed [96]. The term GBM describes a primary brain neoplasm, which consists of a phenotypically and genetically heterogeneous group of tumors [43], [65], [96]. As Urbańska et al. [96] describe, over 90% of diagnosed GBM cases develop as primary GBM, arising through multistep oncogenesis from normal glial cells. Secondary neoplasm, developing from low-grade tumors through progression [41], [93], make up the remaining cases of GBM [96]. Even though the genetic basis and molecular pathways underlying the development of primary and secondary GBM differ [43], they are not showing significant morphological differences [47], [96]. Figure 2.1 shows an exemplary case of a GBM.

Urbańska et al. state that the etiology of GBMs has yet to be fully elucidated [96]. GBM is a spontaneous tumor, despite the medical history describing the occurrences of GBMs in related people [34], [96]. However, only 1% of the cases make up the familial form [83], and the genetic background for development differs from spontaneously arising ones [23]. In the course of genetic diseases such as tuberous sclerosis [69], Turcot syndrome [31], multiple endocrine neoplasia type IIA [89], or neurofibromatosis type I [16] GBMs may also develop [96]. Head injuries might predispose the development of GBMs as well [66], [107]. Viruses (e.g., the human cytomegalovirus) and ionizing radiation are potential etiologic agents that increase the probability of developing a GBM [96].

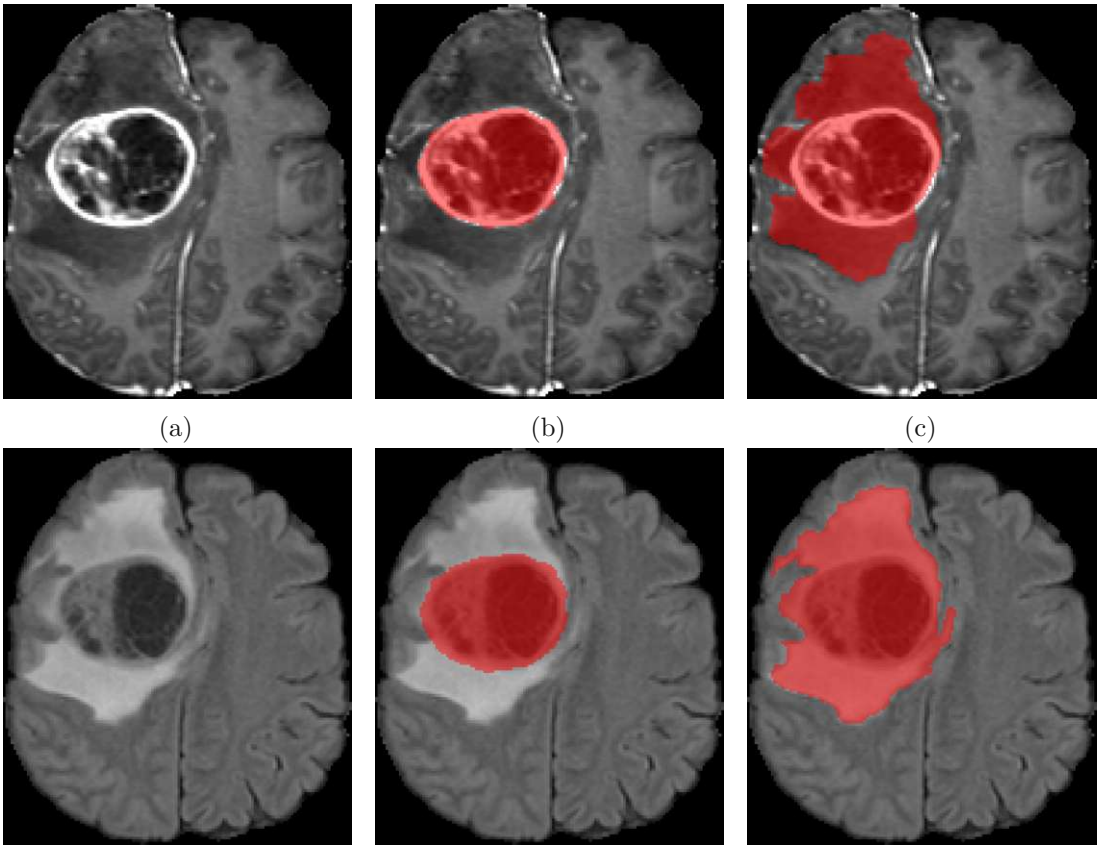


Figure 2.1: Example of a GBM, (a) shows the T1 contrast-enhanced (T1c) sequence of this example, in (b) the GBM is segmented/highlighted, while in (c) the GBM and the caused edema is segmented; (d) depicts the same GBM but with the FLuid Attenuated Inversion Recovery (FLAIR) sequence, (e) and (f) show again the segmentation for the tumor, and the edema plus tumor, respectively. (Source: The GBM images and segmentations displayed are from the data set used in this thesis.)

Infiltrating growth is a characteristic of GBMs, making it difficult to distinguish normal tissue clearly from the tumor mass [43], [107]. Increasing intracranial pressure can be caused by a growing tumor [17], as well as occasionally hydrocephaly [39]. Metastases of GBMs are rare and occur by blood [59] or cerebrospinal fluid [12], but not by lymphatic vessels, since the brain is devoid of those [77]. Reported targets of metastases are spleen, lungs, liver, bones, lymph nodes, pleura, pancreas, and small intestine [2], [36], [67], [77], [98], [103]. Researches hypothesize that the barrier created by cerebral meninges, the rapid tumor growth, and the short course of GBMs are reasons for the low metastatic likelihood [94].

The most common symptoms of GBMs are headaches, ataxia, dizziness, vision disturbances, and frequent syncope, depending on the location and the increasing intracranial

pressure caused by the GBM's clinical stage [54], [80]. Potential misdiagnoses of GBMs are inflammations, infections, and immunological and circulatory diseases because the symptoms are unspecific [54]. Further indications for GBMs are seizures in people not suffering from epilepsy [80]. MRI is the primary tool for diagnosing a GBM [96], while a histopathological examination of the tumor removed (or its parts) is the basis for the definitive diagnosis [44]. If removing the tumor via surgery is impossible, a fine-needle aspiration biopsy is performed [82]. Criteria of the WHO are the basis for morphological diagnosis, whereas the staging of central nervous system tumors includes a proliferative index, survival time, response to treatment, assessment of their morphology, and grade of malignancy (grade I-IV) [96]. While grade I includes non-malignant tumors, and grade II relatively non-malignant tumors, grade III classifies low-grade malignancy tumors, and grade IV describes the most malignant tumors; GBM is a grade IV tumor [48].

Complete resection is impossible since GBM infiltrates the surrounding tissue, and radiotherapy is not always efficient [43]. The blood-brain barrier makes treatment difficult, and tumor cells in the areas of hypoxia are resistant to radiotherapy [19]. Feasible surgical resection followed by radiotherapy and chemotherapy is the mainstay of the treatment for GBMs [87].

2.2 Cancer and the Immune System

The immune system can detect external threats (e.g., bacteria or viruses) and internal threats such as malignant cells [101]. The vaccines' success in preventing diseases indicates that the immune system has a kind of memory to protect the human body [101]. The innate and adaptive immunities form the protective memory [101].

Throughout vertebrate evolution, the innate immunity has been present [97], [101]. Even though the innate immune system is primitive, it can respond within minutes to hours [97], [101]. However, the innate immune system is not built against any specific organism since it does not possess an immunological memory [97], [101]. Macrophages, dendritic cells, neutrophils, mononuclear phagocytes, and natural killer cells are among the innate immune system components [101].

Compared to that, the adaptive immune system rearranges antigen receptors on T- and B-cells [101]. With this, specific structures on antigens (so-called epitopes), which trigger immune responses, can be recognized [97], [101]. Immunoglobulin M and G antibodies are relevant to cancer immunity as they remove pathogens and clear circulating antigens [97], [101]. In addition, these immunoglobins affect complement fixation, antibody-dependent cellular cytotoxicity, and target cell signaling [97], [101].

Cancer induces inflammatory and immune responses when it invades healthy tissue and forms metastasis [70]. These responses can eliminate a tumor with the so-called *immune surveillance* which is a hypothesis that proposes that surveying the body for malignant cells and tumors, as well as recognizing and eliminating those cells, is a significant role of the immune system [70], [97]. If the immune surveillance and response are successful,

the immune system can eliminate tumors at the early stages [70]. Developing tumors need to use immune tolerance induction or immune evasion processes to survive [70], [97]. Immune tolerance induction describes a process where tumors adapt to their immunological environment in a way that prevents immune responses, which can be harmful [70]. Creating a local microenvironment that inhibits immune cell activity, which targets the tumors, describes the immune evasion process [70].

Cancer immunotherapy aims to resurrect the suppressed immune system so that it attacks the tumor cells again and ideally eradicates cancer [49], [97]. The dominant mechanism of immune evasion taken by the tumor provides a potential Achilles' heel [97], [101]. Therapeutically attacking this potential Achilles' heel can restore immune control [97], [101]. Multiple mechanisms may be present, but many cancer types are likely to use similar immune evasion mechanisms [97], [101]. Some cancer immunotherapies cause a narrow activation range of the immune system, while others result in a broad activation [3]. Some immunotherapies, e.g., monoclonal antibodies, are available commercially, while others require personalization with genetic engineering [3].

2.3 Tumor-Infiltrating Lymphocyte

Part of the immune system are TILs which penetrate tumor defenses to attack malignant cells [97]. Cancers can use immune checkpoints to evade elimination by deactivating TILs - immune checkpoints are ligand and receptor pairs that are part of immune response modulations [97]. As an example, when the immune checkpoint ligand Programmed Death Ligand-1, expressed by malignant cells, engages the corresponding immune checkpoint receptor (PD1+) on the surfaces of activated T-cells, the T-cells become ineffective and adopt an "exhausted" phenotype [97], [101].

CD3+ and CD8+ are receptor glycoproteins on mature T-lymphocytes, where they act as antigens [75]. Rathore et al. [75] report that CD3+ cell density is already correlating with oropharyngeal, colon, or cervical cancer, and CD8+ correlates with colon and ovarian carcinoma [75]. In addition, Rathore et al. [75] mention that CD8+ TILs may inhibit tumor growth and that a higher number of CD8+ TILs links with disease-free survival and overall survival in particular [75].

Another promising immunotherapy target is the PD1+ receptor present on activated T-cells [3], [70]. PD1+'s main role is preventing autoimmunity during an inflammatory response by limiting T-cell activity [3], [70]. Binding PD1+ to its ligands (PD-L1 or PD-L2) on CD8+ T-cells, leads to apoptosis [3], [97]. The binding also leads to decreased T-cell proliferation and cytokine production [3], [97]. Tumor cells and the tumor microenvironment overexpress the ligands PD-L1 and PD-L2 [3], [70]. As a result, many cancer types and a large proportion of many tumor types' TILs express PD-L1 [3], [70]. Consequently, a treatment attacking the PD1+/PD-L1 pathway can cause a more prolonged and active antitumor immune response [33].

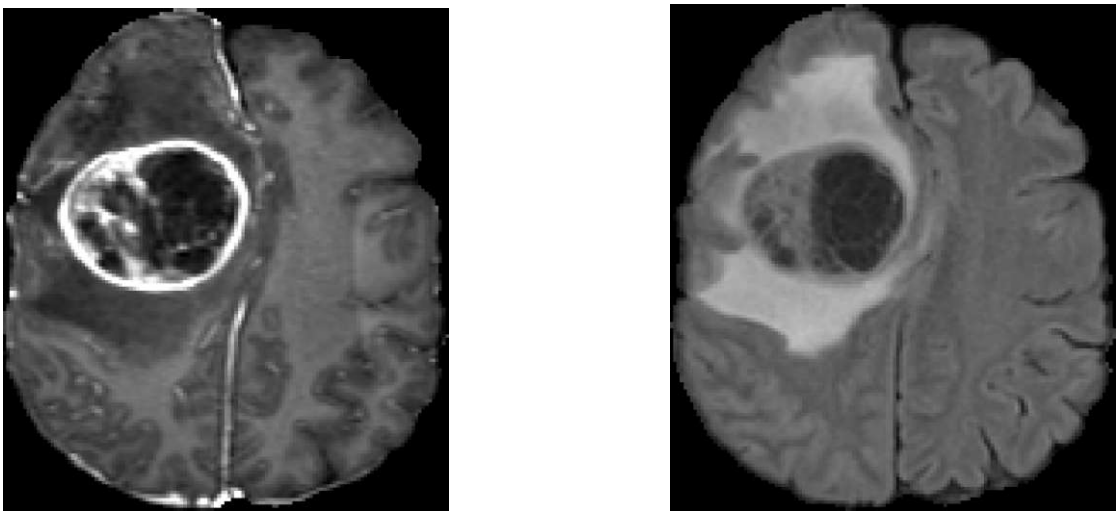
2.4 Cancer Imaging

Medical imaging is a significant factor that has informed medical science and treatment [1]. Clinicians use imaging for oncologic diagnosis and treatment guidance by non-invasively assessing tissue's characteristics [1]. Medical institutions have access to image acquisition and reconstruction methods, such as MRI, modern Computer Tomography (CT), Position Emission Tomography (PET), and combined PET/CT [28]. Imaging has the potential to guide therapy since it provides a comprehensive view of the tumor and can be used to monitor progression [1]. In addition, clinicians acquire images during treatment repeatedly in routine practice, and imaging is considered less invasive than surgery or biopsies [1].

Compared to that, methods relying on surgeries or biopsies do not allow a complete characterization of the entire tumor, as these methods extract only small samples of the tissue for analysis [1]. Clinicians use MRI routinely for diagnosis, characterization, and clinical management of GBMs [24]. MRI is a non-invasive and powerful diagnostic imaging tool that allows a global assessment of a GBM and its interaction with the local environment [24]. Images acquired with MRI capture multidimensional and in vivo snapshots of GBMs [24]. MRI can extract functional, physiological, compositional, and structural information [24].

MRI utilizes the body's natural magnetic characteristics to produce detailed images [7]. Imaging methods use the hydrogen nucleus (a single proton) due to its abundance in fat and water [7]. The protons spin on their axis, behaving like a bar magnet while these axes are aligned randomly [7]. However, the protons' axes line up when the body is in a strong magnetic field like an MRI scanner [7]. Adding additional energy (with a radio wave) to the magnetic field causes the magnetic vector to deflect [7]. Switching off the radiofrequency's source causes the magnetic vector to return to its resting state and emit a signal (another radio wave) [7]. This emitted signal is detected and used to create the images [7]. There are two ways to measure the time it takes for the protons' complete relaxation [7]. The first's name is T1 relaxation time, which is the time the magnetic vector needs to return to its resting state [7]. T2 relaxation is the name of the second time which the axial spin takes to return to its resting state [7]. Relaxation times build the basis for the differentiation between the (e.g., fat and water)), as different tissues have individual relaxation times [7]. Unlike x-ray and CT, MRI uses radiation in the radiofrequency range, which does not harm the tissue and is found all around us, and consequently, researchers consider MRI as non-invasive [7].

FLAIR and T1c are MRI sequences that can be useful for brain tumors such as GBMs [53]. In detecting subtle changes in various brain areas, FLAIR sequences are useful since they display a high sensitivity to many diseases [22]. Figure 2.2 displays an example of a GBM patient's FLAIR and T1c sequences. Figure 2.2a shows the patient's T1c sequence, where the bright area depicts the contrast-enhanced part of the tumor. Figure 2.2b displays the FLAIR sequence where the bright area in the brain depicted shows the edema caused by the tumor, while the dark areas within the bright area show the brain tumor itself.



(a) (b)

Figure 2.2: Example of T1c and FLAIR sequences acquired of a GBM patient. (a) displays the T1c sequence, here the bright area is the contrast-enhanced part of the tumor. (b) shows the patient's FLAIR sequence. The dark area within the brain is the tumor itself, while the brighter area shows the edema caused by the tumor. (Source: The GBM images displayed are from the data set used in this thesis.)

2.5 Summary

This chapter describes the medical background focusing on GBMs, the immune system, TILs, and methods for imaging cancer. Regarding GBMs, this chapter describes what they are, how they can occur, what is done to treat them, and why they are so lethal. In addition, this chapter provides an overview of the general workings of the immune system and its interaction with cancers. In detail, an introduction to TILs and their importance in immune therapy is described. Methods for cancer imaging are presented, focusing on the MRI sequences FLAIR and T1c used in this thesis.

CHAPTER 3

State-of-the-Art: Radiomics

Kumar et al. [53] point out that the term *radiomics* describes a process of the extraction and analysis of large amounts of imaging features obtained from medical images, which have a high throughput [53]. The general hypothesis of radiomics is that the heterogeneity in the medical images is a result of the underlying genetic heterogeneity of the tissue [108]. With this, radiomics aims to identify quantitative imaging indicators which predict clinical outcomes, such as response to a specific cancer treatment [108]. Radiological practice for cancer is usually qualitative [53], such as *a peripherally enhancing mass in the upper right lobe*. Quantitative radiologic measurements are limited to the tumor size, but such quantitative measurements do not reflect the complexity of the tumor morphology or behavior [53]. Changes in these measures do not provide predictive therapeutic benefit in many cases [53]. Radiomics aims to transform images into mineable data with high throughput and fidelity [53]. Radiomics is an accumulation - a pipeline - of multiple steps when looked at it in detail [53]. The most common parts of the radiomics pipeline are:

- **acquisition of the medical images.** Typical medical imaging modalities are MRI, CT, diffusion-weighted imaging, or PET [42], [53].
- **Region-of-Interest (ROI)'s segmentation.** Segmenting ROIs in images, e.g., tumor tissue, necrotic tissue, or normal tissue, is an important, required step for further analysis [53].
- **extraction of radiomics features.** Radiomics features are extracted from the ROIs defined and describe characteristics such as shape or texture patterns [53].
- **analysis of the features extracted.** Statistical modeling highlights associations between the clinical characteristics and features extracted [18].

By extracting quantitative features from clinical images, radiomics aims to improve the understanding of biology and treatment [108]. As cancer imaging analytic methods

have produced novel insights, computational models based on imaging are becoming an important technology that allows identification, analysis, and validation of quantitative features extracted [108]. Though radiomics primarily origins from basic research and is most well-developed in oncology, applications in many scenarios are possible [28].

Figure 3.1 displays a typical architecture of the radiomics pipeline, consisting of the four main modules: image acquisition, ROI segmentation, feature extraction, and feature analysis.

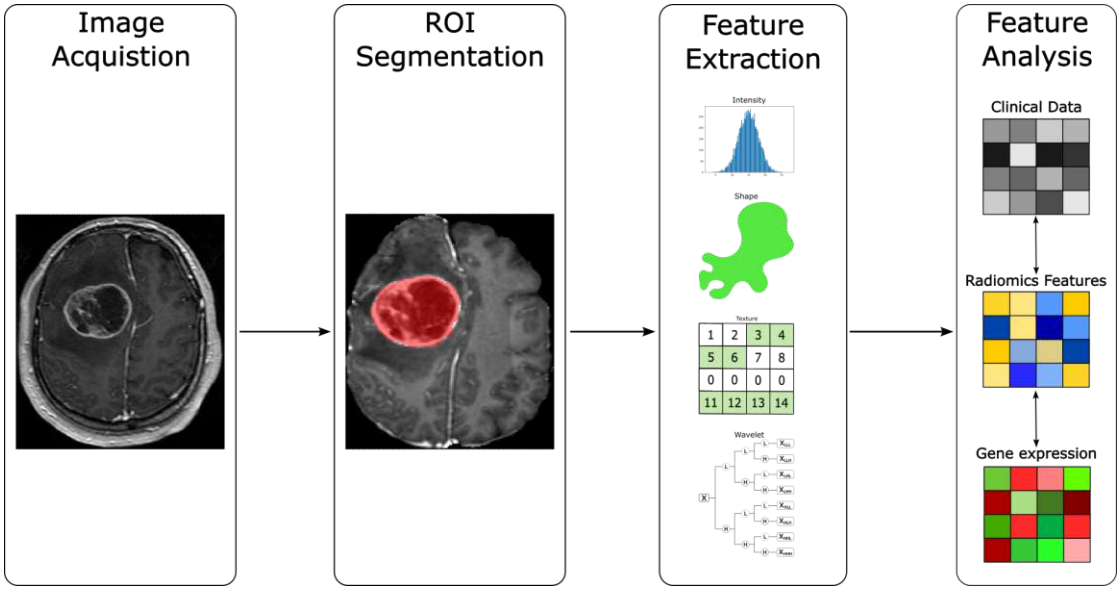


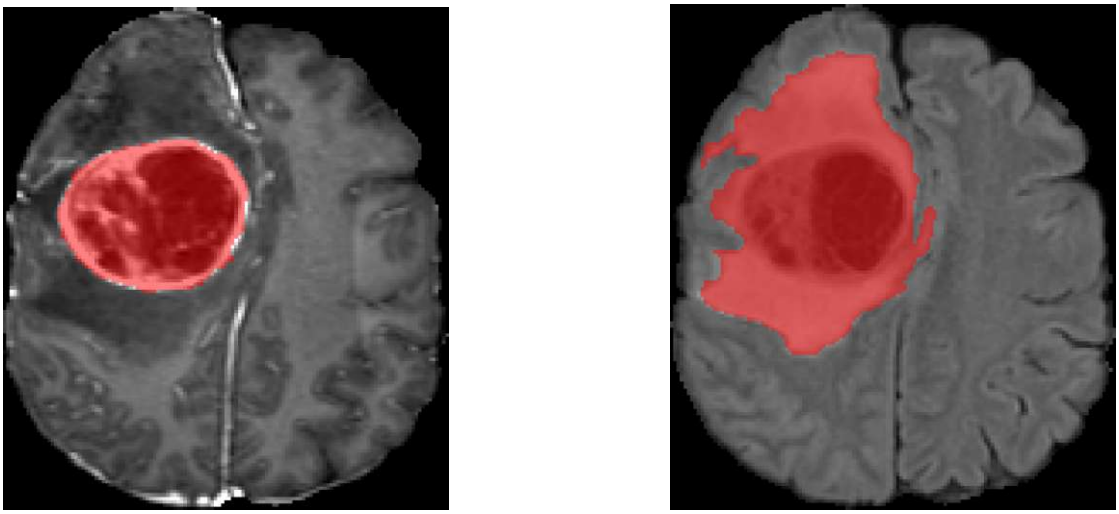
Figure 3.1: Illustrates a typical architecture of a radiomics pipeline. The illustration contains the four main modules of a radiomics pipeline: image acquisition, ROI segmentation, feature extraction, and feature analysis. (Source: Adapted from [1]; The GBM images and the segmentations displayed are from the data set used in this thesis.)

3.1 Region-of-Interest Segmentation

A requirement for radiomic image analysis is accurate labeling of the ROI [18]. Tumor volume, edema volume, normal tissue, and other anatomical structures need to be defined to extract radiomic features from them [18], [53]. The terms *segmentation* or *labeling* describe an act of employing pathological, clinical, and imaging features to mark out the ROI on two-dimensional MRI images [18]. Clinicians - typically an oncologist or radiologist - segment the ROI [18]. This manual segmentation is treated as ground truth [53], but it suffers from interreader variability since the ROI definitions differ between clinicians [18]. To overcome this issue, each clinician generates their own ROIs, and the true ROI is considered the common area [18], or the clinicians find a consensus [72]. Later on, the final segmentation is matched with the corresponding MRI images to extract

the radiomic features [18]. Figure 3.2 shows examples of ROIs segmented by clinicians for this work. In Figure 3.2a an example of a ROI segmenting the contrast-enhanced tumor is displayed, while a ROI labeling the edema and tumor region is shown in Figure 3.2b.

Even though different anatomical regions and imaging modalities require ad hoc segmentation approaches, they share some common requirements [53]. A segmentation method should be time-efficient, provide accurate and reproducible boundaries, and be as automatic as possible [53]. Alternatives to manual segmentation are semiautomatic and automatic segmentation [53]. Across different imaging modalities and various anatomical regions like the lung or brain, semiautomatic and automatic segmentation methods can be used [53]. Automatic segmentation methods are preferable for their time efficiency and precision [58]. However, a significant signal difference between the background and the lesion is needed for an automatic method to be feasible [58]. Due to this, semiautomatic methods are preferable when tumors are surrounded by relatively homogenous structures, as an experienced clinician is required to correct the ROIs automatically segmented [58].



(a) (b)

Figure 3.2: These images show examples of ROI segmentation on the FLAIR and T1c sequences of a GBM patient. The red areas are the ROIs segmented, although they have different meanings. The ROI displayed in (a) segments the contrast-enhanced tumor, while the complete affected area consisting of edema and tumor is the ROI segmented in (b). (Source: The GBM images and the segmentations displayed are from the data set used in this thesis.)

3.2 Extraction of Radiomics Features

The extraction of the radiomics features, which quantitatively describe the ROI, is the core of the radiomics pipeline [28]. The features extracted can be separated into two types

- "semantic" and "agnostic" features [28]. While features commonly used in radiology are semantic features, agnostic features capture the lesion's heterogeneity [28]. These agnostic features are descriptors of the ROI mathematically extracted and are usually not within the radiologists' lexicon [28]. First-order, second-order, and higher-order statistics are among agnostic features [28].

First-order statistical outputs describe the distribution of individual voxel's values without concern for any spatial associations [28]. Typical first-order statistics are methods that reduce the image intensities of an ROI to single values like minimum, maximum, mean, median, entropy, skewness, as well as histogram-based methods [28]. Second-order statistics describe statistical interrelationships between voxels with non-similar or similar contrast values; hence second-order statistics are generally referred to as "texture" features [28]. Texture analysis can measure the lesion's heterogeneity in radiomics [28]. To extract (non-)repetitive patterns, higher-order statistics impose filter grids on the image [28]. Among these are wavelets, which are filter transforms that multiply a matrix of complex-linear or radial "waves" with the image [28]. Laplacian transforms of Gaussian bandpass filters, which extract areas with increasingly coarse texture patterns, are another example of higher-order statistical methods [28].

PyRadiomics is a software library that extracts the radiomics features from medical images [30]. Equation 3.1 describes mathematically the extraction of radiomics features from a (preprocessed) image:

$$f_{fe}(\mathbf{I}_j^{preproc}) = \vec{\mathbf{x}}_j^{radiomics}, \tag{3.1}$$

where $f_{fe}()$ describes the feature extraction process, $\mathbf{I}_j^{preproc}$ is the preprocessed image of sample j, and $\vec{\mathbf{x}}_j^{radiomics}$ is the resulting radiomics feature vector for sample j.

3.3 Feature Analysis and Prediction

Statistical modeling can highlight relationships between a given feature's scope and a clinical characteristic once the radiomic features are extracted [18]. Feature analysis can use statistical methods, machine learning, or artificial intelligence, such as neural networks, random forests, and Bayesian networks [28]. Since, in practice, not all information is available for each patient, models should be able to handle sparse data [28]. The size of the data set and the data quality determine the power of the statistical model entirely [28].

Supervised machine learning methods (e.g., Least Absolute Shrinkage and Selection Operator (LASSO), random forest, neural network, Bayesian network) place varying numbers of the pre-determined predictive features into groups [18]. To determine the most reliable combination, the features' relative contribution to the model's predictability is changed [18]. These methods make no assumptions about the meaning of individual features, despite using a priori knowledge through a training set [28]. A random forest is a simple classifier that automatically selects predictive features [18].

Feature analysis can be done *univariate* or *multivariate* [18]. Univariate analysis determines if a single feature is a predictor for the clinical characteristic on its own, whereas significance is typically defined as a $p < 0.05$ or $p < 0.01$ [18]. Among univariate analysis are methods for assessing the correlation (like Pearson, Spearman rank) and significance (like Wilcoxon test or log-rank) [18]. The multivariate analysis separates seemingly relevant features on univariate analysis from likely independent predictors [18]. It is crucial to limit non-predictive features from influencing the final statistical model [18]. Equation 3.2 describes the prediction step of radiomics mathematically:

$$f^{pred}_{radiomics}(\vec{\mathbf{x}}^{radiomics}_j) = \hat{y}_j, \tag{3.2}$$

where $f^{pred}_{radiomics}()$ describes the process of predicting a new outcome $\hat{y}_j$ based on the radiomics feature vector $\vec{\mathbf{x}}^{radiomics}_j$ for sample j.

3.4 Summary

This chapter describes state-of-the-art radiomics. First, this chapter provides information about what radiomics is and introduces the most common steps. Afterward, ROI segmentation is described, and different methods to tackle this issue are presented. After segmenting the ROI, the radiomics features can be extracted, whereas different categories of radiomics features are described. In the end, the radiomics features extracted can be analyzed and used for the prediction model - this chapter outlines methods for that step as well.

CHAPTER 4

State-of-the-Art: Machine Learning

Jordan and Mitchell describe that machine learning addresses how a computer can automatically improve through experience [38]. The term *learning problem* defines the problem of improving a performance measure through training experience when executing a task [38]. Figure 4.1 provides an non-exhaustive taxonomy of different machine learning methods. A machine learning problem deals with data(sets), which are composed of multiple data points (or so-called samples) [4]. Each data point represents what should be analyzed, e.g., a patient in a group of patients [4].

The properties of each data point are described as *features*, which can be categorical (e.g., sex of the patient), ordinal (e.g., tumor stage), or numerical (e.g., the diameter of the tumor) [4]. A data point can consist of features with different categories [4], e.g., a female patient suffering from a stage II tumor with a diameter of 3 cm. The combination of all features makes up the *feature space*, as each feature represents one dimension of it [4]. The value of a feature determines the placement of the data point along each dimension [4]. Combining all values of all features creates the so-called *feature vector* [4].

A similarity or distance measure needs to be defined to compare two feature vectors [4]. Simple similarity measures would be the Euclidean distance (see Equation 4.1, taken from [4]) between the feature vectors of two samples ($\vec{\mathbf{x}}_i$ and $\vec{\mathbf{x}}_l$) for features $h = 1 \dots m$ [4]. However, depending on the data type, much more complex similarity measures can be used [4].

$$f_{ed}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_l) = \sqrt{\sum_{h=1}^{m}(x_{i,h} - x_{l,h})^2} \quad with\ i\ and\ l \in 1, \dots, n,\ i \neq\ l, \tag{4.1}$$

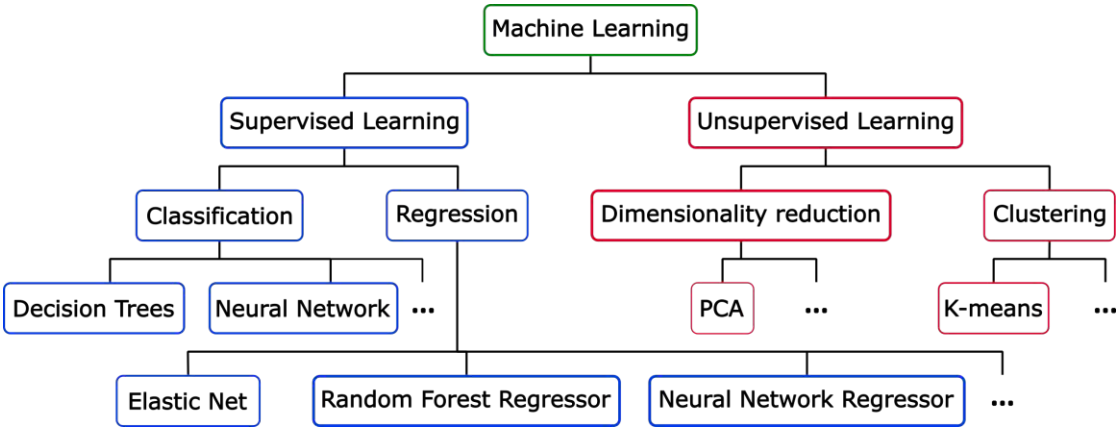


Figure 4.1: Non-exhaustive taxonomy of different machine learning methods (adapted from [4])

where $x_{i,h}$ is the feature at position $h, h \in 1, \ldots, m$ of the feature vector $\vec{\mathbf{x}}_i$ of sample i. Likewise, $x_{l,h}$ is the feature at the same position of feature vector $\vec{\mathbf{x}}_l$ of sample l. The results euclidean distance is described with $f_{ed}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_l)$.

4.1 Supervised Learning

As Figure 4.1 indicates, the term *supervised learning* describes a subset of machine learning methods [4]. Supervised machine learning methods consist of approaches where the data points are labeled, which means that their data points' outcome variable is known and present during the learning process [4], [92]. Supervised learning aims to learn generalized rules (or a generalized model) that maps the data points to their respective labels [4], [92]. A supervised machine learning model trained correctly should be able to predict unseen, new data points without knowing their associated label [4]. Supervised learning consists of two main categories, on one hand *classification* where the labels are qualitative, and on the other hand *regression* where the labels are quantitative [4], [92]. This work focuses on regression, as the target values (the TILs) are quantitative.

Mathematically speaking, a learning set $\mathbf{D}$ consists of data pairs $(x_j, y_j), j = 1 \ldots n$ where x_j represents the input for data point j and y_j is the target value [13]. In regression tasks, the y_js are quantitative values $y_j \in \mathbb{R}, j = 1 \ldots n$ [92]. Equation 4.2 describes the training $f^{learn}()$ and evaluation process of a machine learning model, where a machine learning model $\mathbf{m}_{trained}$ trained with $\mathbf{D}$ and the features of a sample $\vec{\mathbf{x}}_j$ are used during the prediction step to predict $\hat{y}_j$.

$$f^{learn}(\mathbf{D}) = \mathbf{m}_{trained} \rightarrow \mathbf{m}_{trained}(\vec{\mathbf{x}}_j) = \hat{y}_j \tag{4.2}$$

Based on the observed data points, the model parameters are estimated [4]. This procedure is the so-called *model fitting* [4]. A similarity measure between the model and

the data is defined [4]. Minimizing the similarity measure derives the optimal values of the model parameters [4]. This similarity measure, which is also called *loss function* or *cost function*, should always be minimized regardless of the chosen similarity measure, as this is the aim of model fitting [4]. Badillo et al. [4] state, that this minimization has two requirements:

- The values predicted from the model should be close to the observed values of the data points to avoid *underfitting*, and the model has a high *bias* [4].
- The predictive model should generalize beyond the observed data points [4]. When a model predicts well on the training dataset but poorly on an unseen test dataset, the model *overfits* [4]. An overfitting model is often too complex, causing the predictive model to have a high *variance* [4].

4.2 Overfitting and Underfitting

A learning algorithm aims to learn a model that describes the observed training data and can generalize to new unseen data while avoiding overfitting and underfitting [4]. If the model is too simple or there are not enough informative features extracted from the training data, underfitting can occur [4]. Overfitting can occur when the model is too complex or by extracting too many features over a small set of training samples [4]. "*This underfitting/overfitting issue is also often referred to as the bias/variance trade-off, which comes from the expression of the expected prediction error, including both bias and variance terms*" ([4] p. 876).

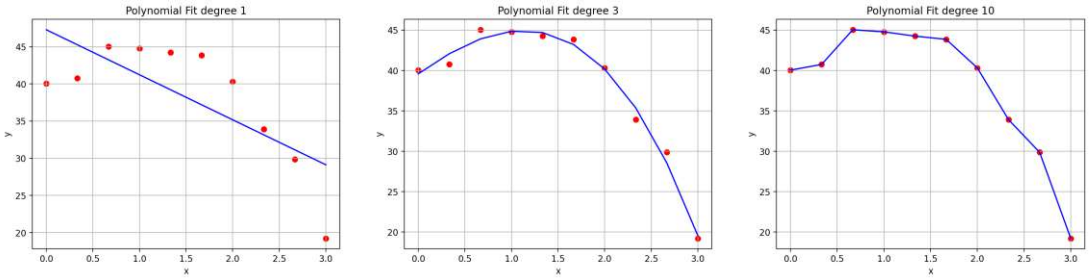
The bias reflects the model's average error for different training sets [4]. The variance indicates the model's sensitivity to the training data set [4]. Increasing the model complexity decreases the bias but increases the variance, so a trade-off between minimizing bias and variance is needed [4]. Figure 4.2 illustrates an example of overfitting and underfitting while also providing a good fit for these data points. Figure 4.2a illustrates an example of underfitting. Figure 4.2b presents an example of a good fit and a fitting bias/variance trade-off. Figure 4.2c illustrates an example of overfitting.

4.3 Regularized Linear Regression

A linear regression model builds on the assumption that the regression function is linear regarding its inputs [92]. Such a linear regression model has h predictors $x_{j,1}, \ldots, x_{j,h}$ for sample j [109]. The prediction of the linear regression model's response $\hat{y}$ is given in Equation 4.3:

$$\hat{y}_j = \hat{\beta}_0 + x_{j,1}\hat{\beta}_1 + \cdots + x_{j,h}\hat{\beta}_h, \tag{4.3}$$

where the coefficients vector $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_h)$ have to be produced by the model fitting process [109]. Depending on the circumstances, the evaluation criteria for the model's quality will differ [109]. Usually the two aspects *prediction accuracy on new, unseen data*



(a) Underfitting (b) Good fit (c) Overfitting

Figure 4.2: Examples of overfitting, a good fit, and underfitting. (a) presents an example of underfitting, (b) provides an example of a good fit, while (c) illustrates an example of overfitting (adapted from [4]).

and *interpretability of the model* are of importance [109]. The ordinary least squares estimates are an example of fitting the coefficients vector, but they often perform poorly in both aspects [109]. This is especially the case if a high number of predictors exist [109]. Penalization techniques can improve ordinary least squares [109]. Minimizing the residual sum of squares where the coefficients are subject to a bound on the L_2-norm, ridge regression is an example of an improvement [109].

Tibshirani [91] proposed another technique – the so-called LASSO [109]. The LASSO uses the L_1-penalty for simultaneous shrinkage and variable selection [109]. Zou and Hastie state that the LASSO has some limitations, despite success displayed in several situations [109]. The following three scenarios are of special interest:

1. The LASSO selects n variables at most before it saturates when the number of predictors is larger than the number of samples since the optimization problem is convex [109]. Zou and Hastie point out that this seems to be a limiting feature for the variable selection procedure and that the LASSO is not well defined when the bound on the L_1-norm is not below a certain threshold. [109].
2. The LASSO tends to select only one variable from a group of variables, among which the pairwise correlations are high, and it pays no attention to which variable it selects [109].
3. If the predictors are highly correlated ridge regression dominates the prediction performance of LASSO [91], [109].

In the early 2000s, Zou and Hastie proposed a new variable selection and regularization method - the so-called elastic net [109]. It is a generalization of the LASSO, which fixes the issues mentioned but also performs well when the LASSO performs well [109]. Like the LASSO, the elastic net simultaneously performs a continuous shrinkage and automatic variable selection [109]. In addition, the elastic net can select groups of

correlated variables [109]. Experiments with real-life data examples and simulation studies show that the elastic net often outperforms the LASSO regarding prediction accuracy [109].

Zou and Hastie define their so-called *naïve elastic net* [109] as follows. Let the data set have n observations with m predictors, $\vec{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ be the response, and $\mathbf{EN} = (\mathbf{m}_{EN,1}, \ldots, \mathbf{m}_{EN,n})$ the model matrix, whereas $\vec{\mathbf{x}}_j = (x_{1,h}, \ldots, x_{n,h}), h = 1, \ldots, m$ are the predictors [109]. The response is centered, and the predictors are standardized after a scale and location transformation (see (4.4)) [109].

$$\sum_{j=1}^{n} \hat{y}_j = 0, \qquad \sum_{j=1}^{n} x_{j,h} = 0 \qquad \text{and} \sum_{j=1}^{n} x_{j,h}^2 = 1, \qquad \text{for } h = 1, \ldots, m \tag{4.4}$$

Equation 4.5 defines the naïve elastic net criterion for any fixed non-negative λ_1 and λ_2 [109]. Zou and Hastie [109] point out, that the naïve elastic net estimator $\hat{\beta}$ is the minimizer of Equation 4.5 (see Equation 4.6).

$$L(\lambda_1, \lambda_2, \beta) = |y - EN\beta|^2 + \lambda_2 \sum_{h=1}^{m} \beta_h^2 + \lambda_1 \sum_{h=1}^{m} |\beta_h| \tag{4.5}$$

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\} \tag{4.6}$$

The *naïve* elastic net tackles the issues described in the first and second scenarios since it is an automatic variable selection method [109]. However, Zou and Hastie call it the *naïve* elastic net since empirical evidence shows that it only satisfactorily performs when it is close to LASSO or ridge regression [109]. The (corrected) elastic net estimator $\hat{\beta}^*$ is a rescaled version of the naïve elastic net estimator, the definition is given in Equation 4.7 [109].

$$\hat{\beta}^* = (1 + \lambda_2)\hat{\beta} \tag{4.7}$$

The corrected version preserves the variable selection property of the naïve elastic net, while the shrinkage is undone [109]. Zou and Hastie [109] empirically found that the elastic net performs well compared to ridge regression and the LASSO. As the elastic net outperforms the LASSO, it also tackles the issue described in the third scenario mentioned [109]. Equation 4.8 describes the training process of elastic net model $\mathbf{m}_{EN}$, where an elastic net trained with $\mathbf{D}$. Equation 4.9 describes the prediction step of elastic net model $\mathbf{m}_{EN}$ where it predicts the response $\hat{y}_j$ for the input sample x_j.

$$\mathbf{m}_{EN} = f_{EN}^{learn}(\mathbf{D}) \tag{4.8}$$

$$\hat{y}_j = \mathbf{m}_{EN}(x_j) \tag{4.9}$$

All in all, a trained elastic net model $\mathbf{m}_{EN}$ provides a model to predict outcomes for new, unseen data points [109]. In addition, the elastic net estimator $\hat{\beta}^*$ has only a subset of non-zero variables which are the variables selected [109].

4.4 Random Forest

Breiman devised random forests in the early 2000s [14]. They belong to the most successful machine learning methods currently available for general-purpose classification and regression tasks [10]. Random forests belong to the supervised learning methods and are based on randomized decision trees and the "divide and conquer" principle [10]. Biau and Scornet describe the "divide and conquer" principle applied as follows: "*sample fractions of the data, grow a randomized tree predictor on each small piece, then paste (aggregate) these predictors together*" ([10] p. 2).

Random forests have few parameters for tuning and they can be applied to numerous prediction problems, which is why this method is so popular [10]. In addition, random forests can deal with high-dimensional feature spaces and small sample sizes, and they are recognized for their accuracy [10]. Due to the concept used, random forests are simple to parallelize and can potentially deal with large real-life systems [10].

Even though random forests are popular, properly analyzing them is difficult due to their black-box behavior [10]. Bagging [13] and the Classification And Regression Trees (CART)-split [15] are crucial components of random forests [10]. "*Bagging (a contraction of bootstrap-aggregating) is a general aggregation scheme, which generates bootstrap samples from the original data set, constructs a predictor from each sample, and decides by averaging*" ([10] p. 3). Each node for each tree selects the best cut with the optimized CART-split criterion [10]. The so-called Gini impurity builds the basis for the CART-split criterion in classification tasks, while the prediction squared error is the basis for regression tasks [10].

How the random forest algorithm grows t different (randomized) trees is described by Biau and Scornet [10] as follows. Before constructing a tree, n_u samples are chosen randomly from the original data set [10]. The tree building takes these - and only these - n_u samples chosen into account [10]. Afterward, each node of every tree performs a split by maximizing the CART-split criterion over the feature subset chosen uniformly at random from the original ones [10]. When all nodes contain less than $n_{nodeSize}$ points, the construction stops [10]. For any $x_u \in \vec{\mathbf{x}}_u$ with $u \in \{1, \ldots, t\}$, each tree predicts $\hat{y}_u$, which are averaged at the end in case of a regression task [10]. Figure 4.3 shows a random forest example for a regression task. Equation 4.10 describes the training $f_{RF}^{learn}()$ and evaluation process of a machine learning model, where a random forest model $\mathbf{m}_{RF}$ trained with $\mathbf{D}$ and the features of a sample $\vec{\mathbf{x}}_j$ are used during the prediction step to predict $\hat{y}_j$.

$$f_{RF}^{learn}(\mathbf{D}) = \mathbf{m}_{RF} \rightarrow m_{RF}(\vec{\mathbf{x}}_j) = \hat{y}_j \tag{4.10}$$

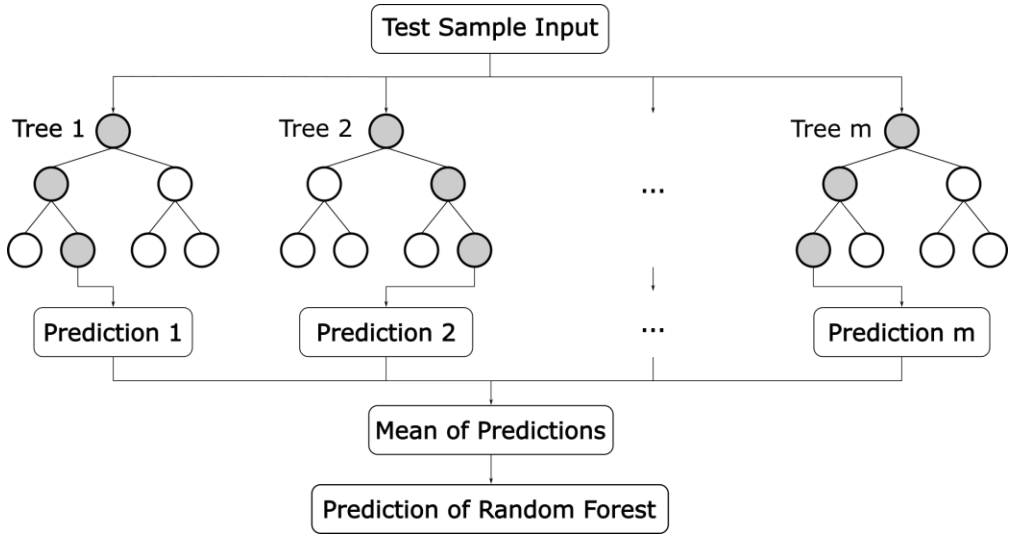


Figure 4.3: Example of a random forest for regression task

4.5 Cross-Validation

Cross-validation is a data resampling method used to assess the generalization ability of predictive models [8]. With that, cross-validation estimates the actual prediction error of predictive models and supports tuning model parameters [8]. The ideal way to assess the generalization ability of a predictive model would be by using new, unseen data which originate from the same population as the training data [8].

However, independent validation studies/data are often not feasible [8]. Berrar continues that estimating the predictive model's performance should happen before investing time and resources for an independent validation [8]. For tuning the model parameter, cross-validation is applied multiple times for different values of tuning parameters, where the final model uses the parameter minimizing the cross-validated error [8]. Due to this, cross-validation addresses the overfitting issue as well [8].

The *single hold-out method* samples some cases for the test set from the learning set at random, while the remaining cases make up the training set [8]. The k-fold random subsampling method generates k pairs of $\mathbf{D}_i^{train}$ and $\mathbf{D}_i^{test}$, $i = 1 \dots k$ by repeating the single hold-out method k times, whereas any pair of training and test set are disjoint, i.e., $\mathbf{D}_i^{train} \cap \mathbf{D}_i^{test} = \emptyset$ [8]. Cross-validation shares characteristics with the repeated subsampling method, but the sampling ensures that no two test data sets overlap [8].

4.5.1 K-Fold Cross-Validation

K-fold cross-validation is a variant of cross-validation where the available training set is split into k disjoint subsets [8]. These disjoint subsets have approximately the same size, whereas the samples are chosen randomly without replacement [8]. The term "fold"

describes the amount of resulting subsets [8]. The model is trained with $k-1$ subsets and is afterward evaluated with the remaining subset, the so-called *validation set*, to measure the performance [8]. Cross-validation repeats this process until every subset has been a validation set [8].

The cross-validation performance is measured by averaging the k performances of the k validation sets [8]. The process of a k-fold cross-validation for $k = 10$, i.e., 10-Fold Cross-Validation (10-FCV) is depicted in Figure 4.4. The first subset is the validation set $\mathbf{D}_i^{val}$ and the remaining nine subsets combined are the training set $\mathbf{D}_1^{train}$ in the first fold, and so on [8].

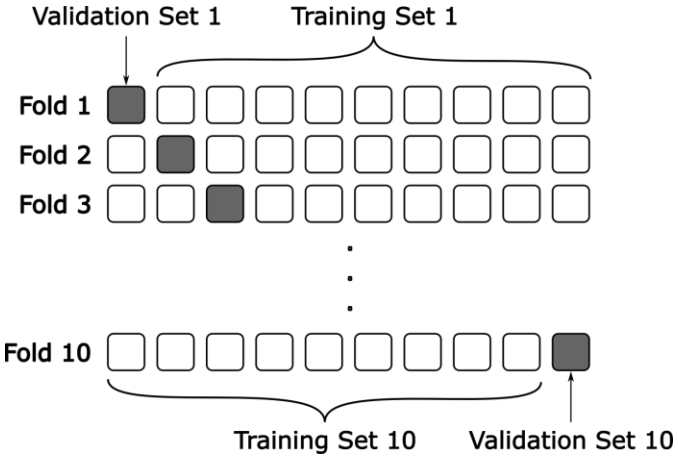


Figure 4.4: Illustrates a 10-FCV as example for k-fold cross-validation. (Adapted from [8])

Berrar further states that cross-validation often involves *stratified* random sampling, which causes the proportions of the targets in the individual subsets to reflect the proportions in the learning set [8]. For example, suppose the learning set consists of $n = 100$ cases whose targets are from 1 to 100, with 80 of the 100 cases having targets below 50 [8]. Without stratification, the random sampling can generate validation sets that only consist of cases with targets above or below 50 [8]. The stratification guarantees that each validation set consists of about 8 cases with values below 50 and 2 cases above 50 when using 10-FCV [8]. Kohavi recommends stratified 10-FCV for real-world data sets [50].

4.5.2 Leave-One-Out Cross-Validation

Leave-One-Out Cross-Validation (LOOCV) is a special case of k-fold cross-validation, where $k = n$ [8]. In LOOCV, each case serves as a validation set in turn, which means the first validation set contains only the first case x_1, the second validation set consists only of the second case x_2, and so on [8]. Figure 4.5 shows an example of the LOOCV procedure. Although the test error of LOOCV is approximately an unbiased estimate of the actual prediction error, its variance is high since two different training sets differ only in one case [92].

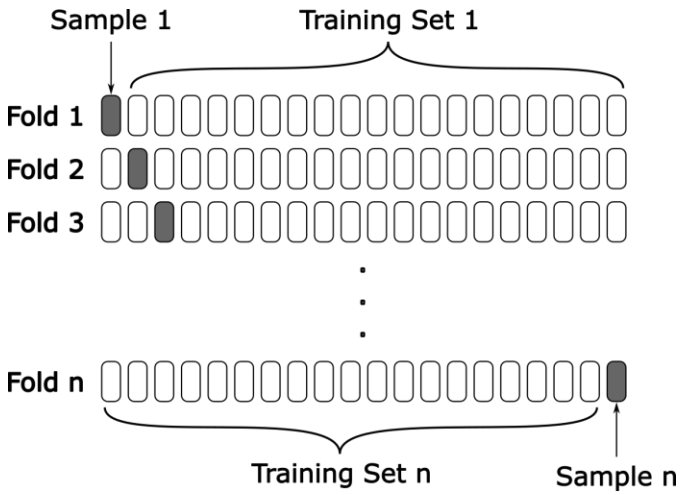


Figure 4.5: Illustrates a leave-one-out cross-validation. (Adapted from [8])

4.6 Summary

This chapter describes state-of-the-art machine learning focusing on subtopics related to this thesis. First, this chapter provides information on machine learning and its major components and workings in general. This chapter describes how supervised learning operates and presents the overfitting/underfitting issue. Afterward, the first machine learning method used for the radiomics approach - elastic nets - is described. Random forests, the second machine learning method used for the radiomics approach, are presented next. In the end, this chapter reports on cross-validation, in detail k-fold cross-validation and LOOCV.

CHAPTER 5

State-of-the-Art: Deep Learning

Linear classifiers, on top of hand-crafted features, are used in many practical applications of machine learning [56]. Such linear classifiers can only split the input space into simple regions [56]. The input-output function needed in problems like image recognition has to be insensitive to irrelevant input variations, e.g., position variations, and they need to be sensitive to minor changes of specific characteristics [56]. Hand-designing a feature extractor is the conventional option, which requires a certain amount of domain expertise and engineering skill [56]. This issue makes it difficult for non-experts to exploit machine learning techniques for their studies, especially in medical-related research [86]. Avoiding the problems mentioned is possible when predictive features can be extracted and learned automatically by a general-purpose procedure, such as deep learning [56].

The architecture of a deep learning procedure consists of multiple layers, each holding a stack of modules [56]. The modules or nodes are subject to learning, and many of them compute non-linear input-output mappings [56]. Every module transforms the input to increase the selectivity and representation's invariance [56]. A system consisting of multiple non-linear layers can implement intricate functions of its input that are insensitive to irrelevant features (e.g., the pose or the background) while maintaining a high sensitivity to minute details [56].

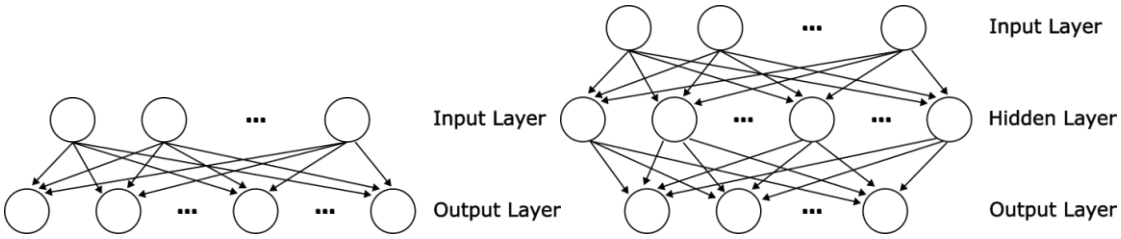
Deep learning is an improvement over conventional artificial neural networks, as networks with more than two layers can be constructed [86]. The advances in processing units, the availability of large data sets, and the improvements in learning algorithms cause the success of deep learning approaches [86]. Deep learning methods need large data sets during the training stage to be highly effective [86]. However, most medical applications have only smaller data sets available [86]. Applying deep learning to such small data sets to build models without suffering from overfitting is considered a primary challenge [86]. Artificially expanding the data set by applying affine transformations (i.e., data augmentation) is among the strategies used to overcome the challenge of small data sets [86].

Another method takes only image patches (instead of the full-sized images) as input to reduce the input dimensionality (thus the number of model parameters) [86].

5.1 Artificial Neural Networks

Artificial neural networks can be characterized as *computation models* with certain properties like the ability to learn, generalize, or cluster data while using parallel processingl [88]. Despite descriptions of parallels with the biological systems, artificial neural networks seem to be an oversimplification of biological systems, as knowledge about the biological systems is limited [88]. After the McCulloch and Pitts [63] introduced simplified neurons, the first wave of interest in neural networks emerged [88]. Shen et al. state that the perceptron [79] is the first trainable neural network, which has an input and output layer [86].

Figure 5.1 displays examples of a typical neural network architecture. The network shown in Figure 5.1a consists only of input and output layer, whereas the perceptron of Rosenblatt [79] is an example of such a network architecture [86]. Figure 5.1b displays a neural network that contains a hidden layer between the input and output layer as well.



(a) (b)

Figure 5.1: Examples for architectures of different neural networks. The neural network in (a) shows a network consisting only of input and output layer. The network displayed in (b) contains a hidden layer between the input and output layer as well. (Adapted from [86])

If data flows within a neural network strictly from input to output neurons, the neural networks are called *feed-forward* neural networks [88]. While the data processing can extend across multiple layers in feed-forward networks, there are no feedback connections present [88]. These feedback connections could extend from the output of neurons to the input of neurons in the same layer or the previous layer [88]. Neural networks that contain such feedback connections are called *recurrent* networks [88]. The perceptron of Rosenblatt [79] is a typical example of a feed-forward neural network [88]. The commercial interest in deep learning intensified when Krizhevsky et al. [51] won the ImageNet object recognition challenge by utilizing a deep learning network [29].

5.1.1 Basic Building Blocks of an Artificial Neural Network

Each neural network's architecture uses similar modules which perform the processing [88]. These modules communicate with each other by sending signals over many weighted connections [88]. Among the modules/building blocks of every neural network are [88]:

- a set of processing units (i.e., neurons),
- an activation state of every neuron, which is the output of the neuron at the same time,
- weighted connections between units, whereas the connection's weights determine the impact of the signal sent from one neuron to the other,
- an activation function that uses the current activation state and the effective input to determine the new activation state,
- an environment in which the system operates while it provides input signals and handles possible error signals.

Neurons & Weighted Connections

A neuron receives input from neurons connected or external sources, uses that to compute an output signal, and forwards the output signal to other neurons connected [88]. Apart from this processing task, a neuron adjusts its weights as well [88]. As many neurons can compute simultaneously, the system is inherently parallel [88]. There are three types of layers holding the neurons in a neural network: a) input layer, which receives the data from external sources (i.e., outside the neural network), and b) output layer, which sends the data out of the network, and c) hidden layer(s), which input signals origin from the layer before it, while it sends its output to the layer following it [88]. In Figure 5.1, the neurons are illustrated as circles, while the figure displays the layered structure of neural networks.

Mostly, each neuron contributes additively to the input provided by another neuron [88]. The total input of a neuron is the weighted sum of the outputs provided by the neurons connected, plus a possible bias term [88]. The term excitation refers to a positive contribution to the weight, while inhibition refers to a negative contribution [88]. In neural networks, every neuron of a layer has connections to neurons in the neighboring layers, but there are no connections between the neurons within the same layer [86]. Neural networks consisting of only input and output layers are regarded as linear models, prohibiting their application in tasks that involve complex data patterns [86]. Neural networks need hidden layers to overcome that limitation [86]. The role of hidden layers is to find informative features for a certain task [86]. Figure 5.1 illustrates the weighted connections within a neural network with arrows from neurons from one layer to neurons at the next layer, but there are no connections between neurons of the same layer.

Activation Function

The activation function of a neuron takes its current activation and the total input to produce a new activation value [88]. In many cases, the activation function is a non-decreasing function of the total input of a neuron, although other functions can be activation functions as well [88]. In general, some kind of threshold function is used as activation function [88]. Such a threshold function can be a hard limiting threshold function (e.g., a sgn function), a semi-linear function, or a smoothly limiting threshold function (e.g., a sigmoid function) [88]. Figure 5.2 displays three examples of possible activation functions: the left panel displays a sgn-function, the middle panel shows a semi-linear function, while the right figure depicts a sigmoid activation function.

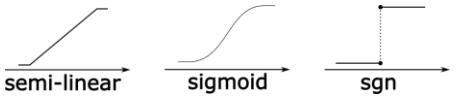


Figure 5.2: Examples of different activation functions used in neural networks. The activation function on the left shows a semi-linear function, while the function displayed in the middle illustrates a sigmoid activation function (i.e., smoothly limiting function). A sgn-function (i.e., a hard limiting function) is illustrated on the right. (Adapted from [88])

5.1.2 Training an Artificial Neural Network

The configuration of artificial neural networks demands that the application of an input set leads to the desired set of outputs [88]. This configuration is handled by setting the weights of the connections [88]. One option is to set the weights explicitly (using a priori knowledge) [88]. Another option feds the neural network teaching patterns and lets it change its weights according to a cost function [88]. The neural network aims to approximate some function, which takes the input and the weights to produce a certain output, making the training of an artificial neural network about finding the values for the weights leading to the best approximation for the function [29]. The training of a neural network requires a cost function, the optimization of the weights to minimize the costs of the cost function (by utilizing gradient descent), and back-propagation to compute the gradients [29].

Cost Function

The training process of an artificial neural network needs a cost function, aiming to find weights that minimize the cost function [29]. The choice of the cost function is a crucial aspect of the neural network's design [29]. Often the cost function uses the cross-entropy between the predictions and the training data (the maximum likelihood principle) [29]. The total cost function used to train an artificial neural network is often a combination of a primary cost function and a regularization term [29].

The specific form of the cost function varies slightly, depending on the model distribution [29]. Deriving the cost function from maximum likelihood provides the advantage that models can share the design of the cost function [29]. The gradient of the cost function must be predictable and large enough to impact the learning algorithm [29]. A saturating cost function undermines that goal when the gradient becomes too small [29]. The negative log-likelihood or cross-entropy helps to avoid this issue [29].

Goodfellow et al. show that the mean squared error can be used as a maximum likelihood estimation procedure since both criteria have the same optimal location despite having different values [29]. Willmott et al. [104] propose, that the Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|, \tag{5.1}$$

with n as the number of samples, y_j the ground truth label of a sample j, and $\hat{y}_j$ the value predicted by the model for j should be the preferred over the (root) mean squared error [104]. Measures of average errors based on the sum of squared errors (e.g., (root) mean squared error) do not describe the average error alone but rather the distribution of squared errors as well [104]. The analysis of Willmott et al. show that MAE is the most natural and unambiguous measure of average error magnitude, hence it should be used preferably compared to (root) mean squared error [104].

Gradient Descent

Gradient descent is commonly used as an algorithm to minimize the cost function $\vec{\mathbf{y}} = g$ by changing the parameters of the model d [29]. Most optimization problems are minimization problems since maximization problems can be turned into minimization problems by minimizing $-g$ [29]. Utilizing the first derivative of a function g' to decide how to change the parameters to minimize the cost is the basic idea behind gradient descent [29]. The first derivative of $g(b)$ gives the function's slope at a certain point b, thus determining how to scale a small change in the function's input to obtain a corresponding change in the output [29]. A function's derivative is useful for the function's minimization, as it determines how to make a small change in the network parameters to make a small improvement in the output [29].

The *learning rate* acts as a constant of proportionality for these infinitesimal steps [88]. A function's derivative (i.e., the gradient) is positive if the gradient points uphill and negative when it points downhill [29]. Moving in the direction of the negative gradient decreases the function f, hence the name gradient descent [29].

The derivative provides no information about in which direction to move, if $g'(b) = 0$ [29]. Points where $g'(b) = 0$ are called critical points, among which are local minima, local maxima, saddle points, and a global minimum [29]. If $g(b)$ is lower than at all neighboring points, the current point is a local minimum, as decreasing $g(b)$ is no longer possible by making infinitesimal steps [29]. If a point not only satisfies the criteria for a local minimum but is the absolute lowest value of $g(b)$, it is a global minimum [29]. The

counterpart to a local minimum is a local maximum that is a point where $g(b)$ is higher than at all neighboring points and thus can no longer be increased by making infinitesimal steps [29]. Critical points that are neither minima nor maxima are saddle points [29]. Figure 5.3 illustrates the three types of critical points, i.e., minimum, maximum, and saddle points.

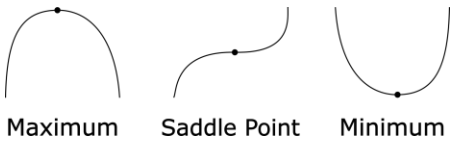


Figure 5.3: Illustrates how the functions of three types of critical points look like. Each of these critical points has a slope of zero [29]. (Adapted from [29])

Back-Propagation

Back-propagation describes a process where the errors of the output layer are propagated back to the hidden layers to determine their errors, thus providing a way to adjust the weights of the input connections [88]. Learning a pattern causes the activation values to propagate to the output layer, which compares the network's output to the output desired [88]. A disparity between the two values results in error values (usually, there is an error value for each neuron in the output layer) [88]. These error values should reach zero, whereas the simplest method to achieve this is the greedy method [88]. The greedy method changes the weights so that the next time around, the error will be zero for that pattern [88]. The output or another hidden layer distributes the error to all connected hidden neurons weighted by their connections to propagate the error to the hidden units [88]. Back-propagation repeats this process until the neurons' weights in all layers (including the input layer) are adjusted [56].

5.2 Convolutional Neural Networks

A particular type of deep, feed-forward network generalizes better than networks with full connectivity between neighboring layers and is easier to train; this is a CNN [56]. CNNs are a special kind of neural networks for processing data, which are known to have a grid-like structure [29]. As the name implies, a CNN uses an operation called *convolution* that is a special mathematical operation [29]. Another operation that CNNs usually employ is called *pooling*.

5.2.1 Typical Convolutional Neural Network Architecture

The typical architecture of a CNN is a series of stages [56]. Convolutional layers and pooling layers make up the first few stages of a CNN, while the rest are usually fully-connected layers [56]. The units in a convolutional layer are organized in feature maps, whereas each unit within is connected to local patches of the previous layer's feature

map through weights called a filter bank [56]. The algorithm passes the resulting local weighted sum through a non-linearity, such as the sgn-function [56]. A filter bank is shared by all units of a feature map, while different feature maps use different filter banks [56]. The idea behind sharing the filter bank is that a pattern can appear in multiple parts of an image [56]. As a result, units at different locations share their weights to possibly detect the same pattern in various parts of the image [56]. More theoretically speaking, local statistics of images are invariant to location, and local value groups are often highly correlated in image data [56].

The pooling layer merges semantically similar features into one [56]. A pooling layer typically computes the local patch's maximum of the units in one feature map or a few feature maps [56]. "*Neighboring pooling units take input from patches that are shifted by more than one row or column, thereby reducing the dimension of the representation and creating an invariance to small shifts and distortions*" ([56], p. 439). A CNN stacks multiple stages of convolution, non-linearity, and pooling, followed by more convolutional and fully-connected layers [56].

Back-propagation of gradients is not more difficult in a CNN than in a regular deep neural network, which allows the training of all weights in all filter banks [56]. In general, deep neural networks exploit the characteristic that natural signals are typically compositional hierarchies [56]. The composition of lower-level features allows the acquisition of higher-level features [56]. In images, such a compositional hierarchy can start with a combination of local edges that form patterns, where patterns assemble into parts, and parts make up objects [56].

Figure 5.4 displays the architecture of a typical CNN for a regression task. The first convolutional layer takes the input images. With a series of convolutional and pooling layers followed by some fully connected layers, the CNN calculates the regression and returns the output at the end [56]. The number of convolutional and pooling layers, and the number of fully connected layers, vary from network to network.

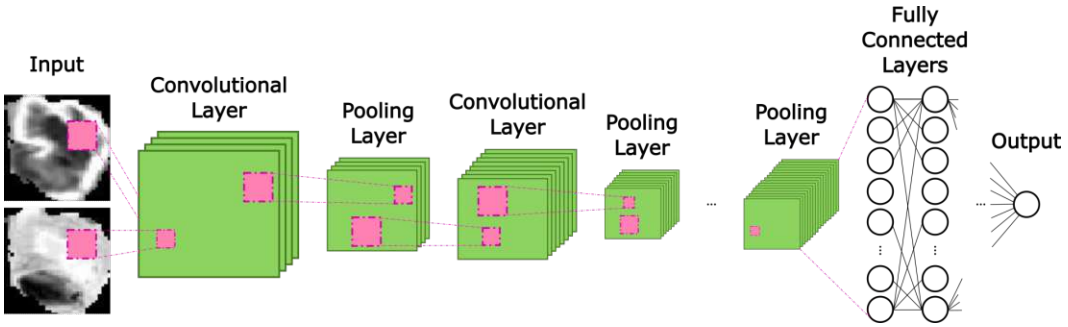


Figure 5.4: Illustrates the typical architecture of a CNN for a regression task. A series of convolutional and pooling layers extracts the information derived from the input. After these layers, some fully connected layers calculate the output of the regression and network. (Adapted from [55])

5.2.2 Convolution

"*In its most general form, convolution is an operation on two functions of a real-valued argument*" ([29], p. 327). Typically, an asterisk denotes the convolution operation, such as in equation 5.2 [29]

$$s = (\mathbf{I}_j^{NN} * w), \tag{5.2}$$

where $\mathbf{I}_j^{NN}$ is denoted as input, while the second function w is referred to as kernel [29]. The result of equation 5.2 s is typically called feature map [29]. Usually, the kernel is much smaller than the input image [29]. Mathematically speaking, the convolution operation is a weighted averaging operation [29]. In most applications, the input $\mathbf{I}_j^{NN}$ is a multidimensional data array, while the kernel w is a multidimensional array of parameters optimized by the learning algorithm [29].

To improve the machine learning system, convolution leverage three ideas: sparse interactions, parameter sharing, and equivariant representations [29]. Sparse interactions (also called sparse connectivity or weights) are accomplished by making kernels smaller than the input, as not every output needs every input [29]. Parameter sharing describes learning only one parameter set during a convolution operation instead of a separate set of parameters for each location [29]. Equivariant representation or equivariance means that the output changes the same way the input changes [29].

Figure 5.5 displays an example of a convolution in a CNN. From left to right, Figure 5.5 illustrates an input image, the convolution kernel, the multiplication result of the kernel and the input image, and the output image as a result of the convolution operation.

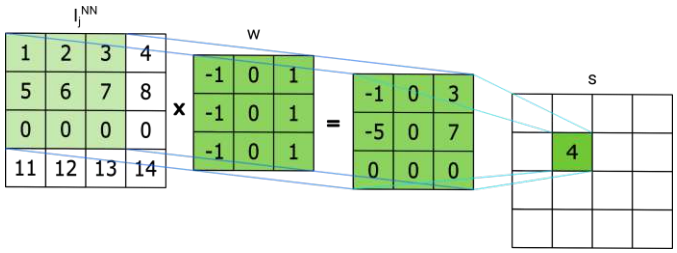


Figure 5.5: Illustrates the convolution operation of a CNN, with (l.t.r.) the input image, the convolution kernel, the multiplication result, and the output image of the convolution. (Adapted from [78])

5.2.3 Pooling

Pooling uses a kernel to generate an output at a specific location within the net with summary statistics of the previous output at that location and of previously neighboring outputs [29]. An example of a pooling function is *max pooling*, which calculates and reports the maximum output from a rectangular neighborhood [29]. The average of a rectangular neighborhood or the L^2 norm of a rectangular neighborhood is another example of pooling functions [29].

Pooling supports making the representation (approximately) invariant to small input translations, independently of the pooling function used [29]. With invariance to translation, a small translation of the input should not change most pooled outputs [29]. If it is more important to know whether a feature is present than its exact location, invariance to local translation is a convenient characteristic [29]. As pooling summarizes the neighborhood's outputs, the computational efficiency of the CNN improves since the next layer has fewer inputs to process [29].

Figure 5.6 shows an example of an averaging pooling operation, where the average of the input image becomes the value in the output image. Figure 5.6 illustrates the input image, the pooling operation, and the resulting output, from left to right.

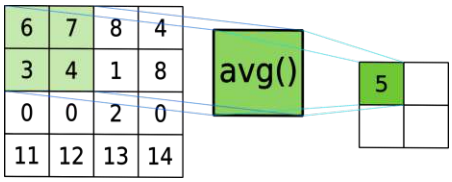


Figure 5.6: Illustrates the average pooling operation of a CNN layer, with (l.t.r.) the input image, the pooling operation (i.e., the averaging operation), and the output image of the pooling. (Adapted from [78])

5.2.4 Residual Learning

Improving the learning power of neural networks is not as simple as stacking more layers in the network due to the vanishing/exploding gradient problem [32]. When a deep neural network starts to converge, a degradation problem has been exposed: with increasing network depth, accuracy gets saturated and then degrades rapidly [32]. He et al. [32] propose a solution for this problem by introducing *deep residual learning* [32]. He et al. [32] approximate the residual function

$$F(\mathbf{I}_j^{NN}) := H(\mathbf{I}_j^{NN}) - \mathbf{I}_j^{NN} \tag{5.3}$$

instead of the expected mapping $H(\mathbf{I}_j^{NN})$, where $\mathbf{I}_j^{NN}$ denotes the input to the first layer of the mapping [32]. The solution proposed is based on the idea that when multiple nonlinear layers can approximate complex functions, then it should be possible that these layers approximate the residual function [32]. When the layers added can be constructed as identity mappings, the training error of the deeper model should not be greater than from a corresponding shallower model [32].

He et al. [32] adopt residual learning to every few stacked layers, with a building block formally defined as [32]:

$$\hat{y}_j = F(\mathbf{I}_j^{NN}, W_i) + \mathbf{I}_j^{NN} \tag{5.4}$$

where $\mathbf{I}_j^{NN}$ denotes the input, $\hat{y}$ the output, and the function $F(\mathbf{I}_j^{NN}, W_i)$ the residual mapping to be learned [32]. Figure 5.7 illustrates a building block of residual learning,

where it shows how neural networks use the residual learning idea presented in Equation 5.4 practically.

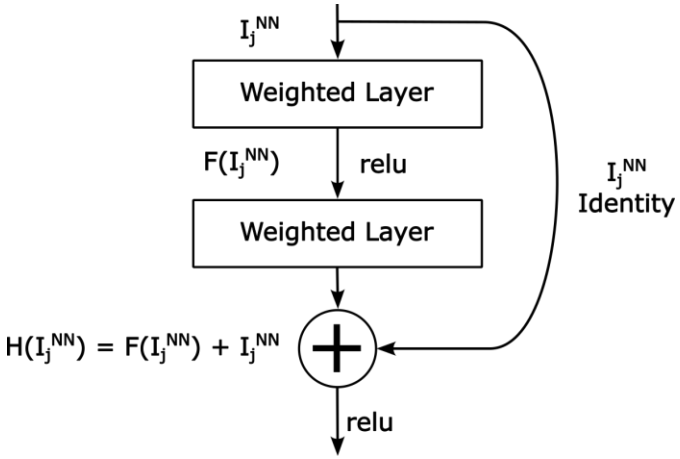


Figure 5.7: Illustrates an example of a building block for residual learning. (Following [32])

5.3 Summary

First, a general overview of deep learning is provided, followed by a description of neural nets and CNNs. This chapter presents the basic building blocks and inner workings of neural nets. In addition, the key components for training a neural net are described. CNNs are a subtype of neural nets used in this thesis. This chapter presents key components and processes of CNNs, and residual learning since this thesis uses a ResNet50 for the deep learning approach.

CHAPTER 6

State-of-the-Art: Image Analysis in Brain Tumors

This chapter presents the related work of this thesis. In detail, the related work focuses on studies that use radiomics in combination with GBMs, and studies that utilize deep learning to tackle problems related to GBMs. In addition, this chapter covers studies that predict TILs as well.

6.1 Radiomics & Glioblastoma

In the research of GBMs, radiomics is gaining momentum as studies (e.g., [6], [21], [27], [45], [46], [72]) make use of it in their research. Kickingereder et al. [45] evaluate if signatures of radiomics features extracted from MRI data allow the prediction of the patient's survival and the stratification of patients with newly diagnosed GBMs. Kickingereder et al. [45] predict progression-free and overall survival with Cox proportional hazards models based on supervised principal component analysis. The findings of Kickingereder et al. show that it is possible to predict survival and that radiomics signatures can stratify patients with newly diagnosed GBMs [45]. The findings further show that accuracy improved compared to established clinical and radiological risk models [45]. Bae et al. [6] find that improving survival prediction is possible while indicating that radiomic MRI phenotyping integrated with genetic and clinical profiles can create a potentially practical imaging biomarker.

While Bae et al. [6] and Kickingereder et al. [45] analyze GBMs directly, Kim et al. [46] utilize radiomics to feasibly distinguish GBMs from primary central nervous system lymphoma, which is another type of brain tumor. As the treatment of the tumor types investigated differ substantially, a feasible differentiation before surgery can be useful [46]. Kim et al. use the minimum redundancy maximum likelihood and the LASSO

algorithm with 10-FCV based on radiomics features extracted from multi-parametric MRI data [46]. Kim et al.'s findings suggest that radiomics features derived from MRI data can differentiate the two tumor types with high accuracy [46].

The research question of Priya et al. [72] is similar to the one of Kim et al. [46], but Priya et al. differentiate GBMs from intracranial metastatic disease. Radiomics features extracted from MRI data build the basis of the research presented by Priya et al. [72]. Compared to Kim et al. [46], Priya et al. compare findings of single-parametric and multi-parametric MRI, as well as combinations of MRI sequences, and how the different tumor segmentations affect the results [72]. To determine the optimized configurations, Priya et al. [72] cross-compared multiple machine learning models based on radiomics using multi-parametric MRI, where the LASSO model performs the best for the multi- and single-parametric MRI. Further findings of Priya et al. suggest that the radiomics features used are more important than the sequence(s) used, as the results show no significant difference between the top-performing models [72].

The research of Gao et al. [27] aims to predict tumor grades and pathological biomarkers using machine learning algorithms on radiomics features extracted from MRI data. With this work, Gao et al. tackle the issue that grading and pathological biomarkers of GBMs have important guiding significance for the individual treatment [27]. The machine learning algorithms investigated are logistic regression, support vector machines, and random forests [27]. Compared with the other algorithms, the results achieved by the random forests are consistently better [27]. The findings of Gao et al. suggest that predicting GBM grades and pathological biomarkers non-invasively, pre-surgery, and with good predictive accuracy and stability is possible with machine learning algorithms based on radiomics data [27].

Choi et al. [21] investigate the potential of radiomics to serve as an imaging biomarker for GBM patients while using a radio-genomics approach to explore the molecular rationale behind radiomics. Choi et al. extract the radiomics features used in the research from multiple habitats of the GBM and multi-parametric MRI data [21]. Choi et al. use the Cox-LASSO algorithm to build a survival prediction model that is the basis for their findings [21]. Based on their results, Choi et al. [21] conclude that radiomics has the potential to act as an imaging biomarker regarding the clinical and genomic significance, which they confirmed by the integrated radio-genomics approach [21].

6.2 Deep Learning & Glioblastoma

With the increasing popularity of deep learning, several studies (e.g., [5], [26], [55], [105]) incorporate deep learning in research regarding GBMs. Training deep learning models can improve results compared to (traditional) machine learning algorithms but require data sets that are large [55]. In medical imaging analysis, data sets are often not large enough for deep learning algorithms to reach their full potential [55], which is why Lao et al. [55] use transfer learning and fine-tuning. Apart from the features extracted via deep learning, Lao et al. extract (handcrafted) radiomics features from MRI data [55].

Combining radiomics features and features extracted via deep learning, Lao et al. generate radiomics signatures with the LASSO Cox regression model to predict the overall survival and patient stratification, which indicates the potential of incorporating deep learning regarding pre-surgery care of GBM patients [55].

Bae et al. [5] evaluate the generalizability and diagnostic performance of deep learning and traditional machine learning models for differentiating single brain metastasis from GBMs using radiomics. This differentiation is significant, as the diagnostic workup and treatment following differs between the two diseases since the treatment depends on identifying the primary tumor and the current tumor-spreading status in case of a brain metastasis [5]. Deep learning using radiomics features can help differentiate metastasis from GBMs without neglecting generalizability [5].

Fu et al. propose a fully automatic workflow for survival prediction of GBM patients using deep learning [26]. Fu et al. use a 3D CNN for automatic segmentation of the GBMs, which they use to generate tumor contours for several GBM patients [26]. Handcrafted radiomics features are extracted from auto-contours of explicitly designed algorithms, while a pre-trained CNN extracts the deep learning features [26]. To create handcrafted and deep learning signatures, Fu et al. [26] train Cox regression models with regularization techniques. The 3D CNN proposed generates accurate GBM contours, and the deep learning-based signature outperforms the signature created from handcrafted radiomics features [26]. Fu et al.'s results demonstrate the potential of improving survival prediction and patient stratification with the automatic workflow proposed [26].

Wong et al. [105] use deep learning to discover prognostic genes for GBM patients' survival. To assess the predictive value of the deep learning features in addition to clinical, methylation, and mutation factors, Wong et al. use univariate and multivariate Cox survival models [105]. Compared to traditional machine learning methods, including the ridge, adaptive LASSO, and elastic net Cox regression models, deep learning provides non-redundant prognostic covariates for patient survival [105]. The findings of Wong et al. show that the deep learning model learns genes related to GBM stem cells and treatment-resistant genes [105]. Using the approach proposed, Wong et al. identify many specific genes which can be potential biomarkers or targets for treatment [105].

In their study, Ma et al. [60] propose two CNN models to predict the glioma's grade based on pathological and radiological data. One of the models presented by Ma et al. is based on a ResNet for classifying 2D slices of pathological data [60]. The results achieved with the CNN-based models proposed suggest that these models can improve the accuracy of glioma grading since the models performed well at the CPM-RadPath-2019 challenge [60]. The models presented by Ma et al. have the potential to support the diagnosis and treatment planning of glioma for radiologists and pathologists [60].

6.3 Prediction of Tumor-Infiltrating Lymphocytes

The prediction of TILs based on analysis of MRI data is subject to some studies, e.g., [9], [37], [52], [57], [106], [110]. Ku et al. [52] predict TILs in triple-negative breast cancer patients based on MRI data. Ku et al. divide the patients into two groups for the analysis, one with high TIL levels and one with low TIL levels [52]. The results of Ku et al. demonstrate that the prediction model proposed can help to identify TIL levels in triple-negative breast cancer patients and has the potential to be used as an imaging biomarker [52].

The prediction model proposed by Bian et al. [9] could predict the TILs for patients with pancreatic ductal adenocarcinoma. Bian et al. split the patients into score-high TILs and score-low TILs groups, where the Cox regression model acquires the TILs score [9]. They use the LASSO and the extreme gradient boosting to select features and construct the prediction model [9]. Bian et al.'s findings demonstrate that the model based on the extreme gradient boosting could predict the TILs and support clinical decision-making for immune therapies [9].

Li et al. [57] present another study investigating pancreatic ductal adenocarcinoma, where they use a multilayer perceptron network to predict the CD20+ expression in that tumor [57]. Novel therapeutic targets are necessary for treating pancreatic ductal adenocarcinoma as conventional chemotherapy has limited benefit [57]. The network proposed is based on radiomics features extracted from MRI data and selected by the minimum absolute contraction and selective operator logistic regression algorithms [57]. The findings of Li et al. demonstrate that predicting the CD20+ expression for patients suffering from pancreatic ductal adenocarcinoma is possible by utilizing a multilayer perceptron network [57].

Jeon et al. [37] propose an MRI-derived radiomics signature to predict CD8+ TIL density changes in chemoradiotherapy patients with rectal cancer. They utilize the LASSO method on MRI-derived radiomics data to establish a radiomics signature [37]. The findings of Jeon et al. suggest the utilization of radiomics-immunophenotype modeling in clinics for evaluating the tumor immune status following neoadjuvant chemoradiation in rectal cancer [37].

As TILs establish themselves as a prognostic indicator of immunotherapy, Çelebi [110] investigate the effectiveness of imaging features regarding the prediction of histologic stromal TIL levels in invasive breast cancer patients [110]. Çelebi et al. use logistic regression analysis to find the statistically significant parameters in predicting histologic stromal TIL levels [110]. The findings of Çelebi et al. show that imaging features can play a role as an adjunct tool in uncertain situations and could improve the biopsy results' accuracies [110]. As Çelebi et al. conclude, their approach could give imaging features an opportunity for the prognosis prediction of invasive breast cancer patients [110].

Wu et al. provide an overview of studies about radiogenomics in the era of immunotherapy [106]. They inform about the high potential of studies confirming the link between

radiological imaging and tumor immune microenvironment [106]. However, most studies have issues regarding the small sample size and external validation [106]. As future studies will advance and investigate radiogenomic relations on promising evidence, these studies can propose predictive biomarkers for selecting patients who will benefit from immunotherapy [106]. Wu et al. state that it may be possible to non-invasively monitor and assess the molecular and tumor microenvironment's evolutions during the treatment with reliable radiogenomic surrogates [106].

6.4 Summary

This chapter presents state-of-the-art image analyses of brain tumors. First, this chapter focuses on image analysis using radiomics as a methodology. Afterward, state-of-the-art using deep learning as a method is presented. Then, various studies are described that predict TILs using image analysis.

CHAPTER 7

Methodology

This chapter describes the methodology of the thesis. The state-of-the-art and background information provided in the previous chapters are the basis for this thesis' methodology. The beginning of the chapter introduces TIL markers as they serve as the actual prediction targets for the experiments. Following the TIL markers, a section provides the image preprocessing steps taken. Preprocessing of the medical images is used since the images contain the entire head, while only the brain is needed. In addition, the same section describes the ROI segmentation used for the experiments. Afterward, the chapter provides the methodology of the radiomics approach, consisting of descriptions regarding the feature extraction, the machine learning models, and the methods used for voxel-based experiments. The beginning of the methodology regarding the deep learning approach provides the preprocessing and data augmentation used, while the section presents the actual deep learning model afterward. In the end, this chapter describes the evaluation.

7.1 Prediction Targets

This study does not use the TILs directly as prediction targets, but specific characteristics of the TILs, the so-called *TIL markers*. This section describes the individual TIL markers used as prediction targets. Afterward, it provides the management applied to handle outliers of TIL markers.

TIL marker

TIL markers are characteristics of a TIL used in the experiments.

Positive describes the total number of TILs in the tissue analyzed. The term y_j^{pos} denotes the ground truth value of this marker for a sample j.

Negative provides the total amount of cells not regarded as *Positive*. The term y_j^{neg} denotes the ground truth value of this marker for a sample j.

Density defines the density of the TILs per mm². The term y_j^{den} denotes the ground truth value of this marker for a sample j.

Percentage labels the relative share of the TILs in the tissue analyzed. Equation 7.1 provides the formula for calculating this marker. The term y_j^{per} denotes the ground truth value of this marker for a sample j.

$$y_j^{per} = \frac{y_j^{pos}}{y_j^{pos} + y_j^{neg}} \tag{7.1}$$

Each TIL marker is present for every TIL, i.e., *CD3+ Density*, *CD3+ Percentage*, *CD3+ Positive*, *CD3+ Negative*, *CD8+ Density*, and so on. As a result, there are 12 separate prediction targets for the experiments. Clinicians acquire the values of each TIL marker for the patients in the data set with the Definiens software (Definiens AG, Munich, Germany), but not every TIL marker is present for each patient. In most cases, clinicians do not capture the TIL (and thus the TIL markers) during the data acquisition. Additionally, this study regards a few individual TIL markers as outliers and excludes those from the experiments.

Outlier Management

This study uses a criterion for possible outliers of TIL markers as outliers struck out while studying the data set. If a value y_j fulfills the criterion denoted in Equation 7.2, this study uses it in the experiments, otherwise, this value is regarded as an outlier:

$$\mu_{tm} - 2 * \sigma_{tm} < y_j < \mu_{tm} + 2 * \sigma_{tm}, \tag{7.2}$$

where μ_{tm} is the mean of the values of TIL marker tm and σ_{tm} is the corresponding standard deviation. As this criterion is applied for each TIL marker individually, the experiments may use *PD1+ Density* from a patient, but not *PD1+ Negative*. The TIL markers cleared of outliers are used for all experiments of both radiomics and deep learning approaches.

7.2 Image Preprocessing

This study preprocesses the MRI sequences, FLAIR and T1c, before using them in the radiomics or deep learning approach. This section provides the methodology used for image preprocessing. One part of the preprocessing is the ROI segmentation, while a step referred to as *skull stripping* preprocesses the MRI images. Equation 7.3 denotes the image preprocessing mathematically for a sample j:

$$\mathbf{I}_j^{preproc} = f_{pp}(\mathbf{I}_j^{acquired}), \tag{7.3}$$

where $\mathbf{I}_j^{preproc}$ denotes the resulting preprocessed image, $f_{pp}()$ describes the preprocessing steps, and $\mathbf{I}_j^{acquired}$ is the imaging data acquired.

7.2.1 Region-of-Interest Segmentation

The ROIs used in this thesis are segmented by two independent clinicians, whereas they agreed on a usual segmentation if their segmentations differ. One segmentation builds upon the FLAIR sequence, where the segmentation covers the high contrast area, which refers to the GBM and the peritumoral edema - this study calls that segmentation *edema*. The T1c sequence is the basis for the other segmentation that covers the GBM, which is the contrast-enhanced region displayed in the MRI image. This thesis refers to that segmentation as *tumor*.

7.2.2 Skull Stripping

The preprocessing of the MRI sequences aims to remove the unnecessary parts of the head displayed in the MRI, which is everything except the brain and to homogenize the MRI images. The process of *skull stripping* consists of a few steps described in the following and uses FMRIB Software Library (FSL; `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL`), ImageMagick (`https://imagemagick.org/`), FreeSurfer (`https://surfer.nmr.mgh.harvard.edu/`) and Advanced Normalization Tools (ANTs; `https://stnava.github.io/ANTs/`).

Resampling: This step ensures that every voxel covers the same volume across all samples [95], e.g., a voxel covers 1mm³. Without resampling, the size represented by a voxel can differ between samples making the radiomics features extracted incomparable.

Reorient and Crop: Reorienting the MRI sequences is necessary, as some patients might have their heads slightly tilted during the MRI scan. Cropping the MRI sequences removes unwanted parts or artifacts outside the skull [35]. Reorienting and cropping the MRI sequences supports the following preprocessing steps as these steps remove unwanted data and simplify brain extraction and registration.

Bias Field correction: As Juntu et al. [40] explain, MRI machines can corrupt MRI images with the so-called bias field signal, which is a smooth and low-frequency signal. The results of algorithms using the pixel's graylevel values (e.g., texture analysis) will be unsatisfactory if the experiments use corrupted MRI images. Hence, correcting the bias field of the MRI images is a necessary step for the experiments of this thesis.

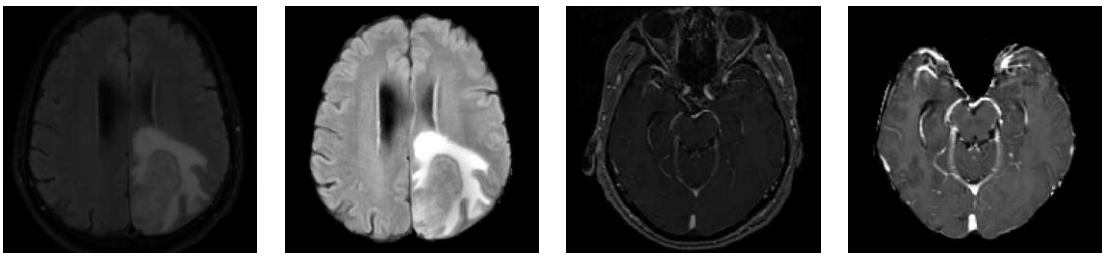
Brain extraction: For the experiments of this thesis extracting the brain from the head displayed in the MRI image is advantageous. Brain extraction causes the images to contain only the brain and reduce the file size, which speeds up computation since less information needs to be processed. Furthermore, the intensities of the bones, eyes, etc. do not affect the intensity rescaling step and thus do not affect the features extracted.

Registration: Registration ensures that both MRI sequences are within the same coordinate system, which is necessary for feature extraction [95]. Without registration, the ROI would only be available to the MRI sequences for which it is segmented. However, as this study uses both MRI sequences (FLAIR and T1c) for feature extraction, the ROI must be available for both MRI sequences. This thesis investigates all three possibilities.

The first option regards the FLAIR sequence as the registration target - this method is later called *flair*. Using the T1c sequence as the target is the second method, called *t1c*. The third option registers only the ROI while the sequences remain in their respective spaces - *orig* refers to this method.

Intensity Rescaling: The intensity of an MRI image impacts the features extracted significantly [95]. Without rescaling, the predictive features learned by a machine learning model can focus on the intensity difference between the images caused by the MRI scanner settings and may not originate from the tissue in the ROI [95]. Intensity rescaling reduces that impact and causes the features extracted to focus more on the actual content of the MRI image.

Figure 7.1 displays how the image preprocessing steps affect the MRI images. The images on the left display the original images captured by the MRI with the FLAIR sequenced shown in Figure 7.1a, and T1c displayed in Figure 7.1c. Figure 7.1b displays the FLAIR image after the preprocessing steps, Figure 7.1d the T1c image after preprocessing. In both examples, the intensity rescaling causes a higher contrast, while the removal of the skull can be seen as well, especially for the FLAIR sequence.



(a) (b) (c) (d)

Figure 7.1: Illustrates the effects of the image preprocessing steps on the MRI images acquired. Figure a displays the original FLAIR sequence while Figure b shows the image preprocessed. For the T1c sequences, Figure 7.1c shows the original MRI image acquired and Figure d displays the T1c image preprocessed. (Source: The GBM images displayed are from the data set used in this thesis.)

7.3 Radiomics Approach

This section provides the methodology of the radiomics approach and build upon the state-of-the-art provided in the chapter above (see Chapter 3 and 4). Firstly, this section explains the methods used for feature extraction. The machine learning models use radiomics features extracted to find patterns and predict new, unseen data samples. This part presents the methods used for elastic net and random forest. Voxel-based experiments can visualize the origin of the (predictive) features extracted, providing an opportunity to see where the most predictive parts of the ROI are. The end of this section provides the methods for the voxel-based experiments.

7.3.1 Feature Extraction

The different image registration methods and ROIs provided during the preprocessing allow for six different settings (2 ROIs * 3 image registration methods). Moreover, an additional setting is created by applying wavelet filters for all six original settings, as studies report their positive impact [1], [42]. As a result, there are twelve settings investigated in total by the radiomics approach - six settings that use the radiomics features only and another six settings that utilize the radiomics features and the wavelet features. This thesis uses the open-source Python-based software PyRadiomics [30] to extract the radiomics features. The features extracted can be categorized as either

- Shape-based,
- First-order statistics,
- Gray Level Co-occurrence Matrix (GLCM),
- Gray Level Size Zone Matrix (GLSZM),
- Gray Level Run Length Matrix (GLRLM),
- Neighboring Gray Tone Difference Matrix (NGTDM),
- Gray Level Dependence Matrix (GLDM), or
- Wavelet-based (only present if the setting uses wavelet-filters)

features [73]. The features are extracted for each setting individually. The feature values of different settings are not the same since the settings used for their extraction are not the same, e.g., the settings differ regarding the ROI used (i.e., one setting uses the *edema* ROI, while the other uses the *tumor* ROI), or regarding the use of wavelet filter.

Equation 7.4 describes the extraction of radiomics features for sample j mathematically:

$$\vec{\mathbf{x}}_j^{radiomics} = f_{fe}(\mathbf{I}_j^{preproc}) \qquad with\ \vec{\mathbf{x}}_j^{radiomics} \in \mathbb{R}, \tag{7.4}$$

where $\vec{\mathbf{x}}_j^{radiomics}$ is the radiomics feature vector, $f_{fe}()$ the feature extraction process, and $\mathbf{I}_j^{preproc}$ the preprocessed image.

7.3.2 Machine Learning Model

The radiomics features extracted build the basis for the machine learning models, which try to find patterns in the training data and predict new, unseen data samples. The thesis uses the two machine learning methods *Elastic Net* $\mathbf{m}_{EN}$ and *Random Forest* $\mathbf{m}_{RF}$. Afterward, this part describes the methodology used for the voxel-based experiments.

Elastic Net

As described in the state-of-the-art in Section 4.3, elastic nets are a generalization of LASSO that use additional parameters. For the experiments, this thesis uses the Python-based glmnet [25] for the elastic net. Before the radiomics features extracted are handed to the elastic nets as the input, they are z-scored. The parameter α is a value in the range $[0, 0.1, 0.2, \ldots, 1]$, resulting in 11 elastic nets trained for each TIL marker and setting. The other parameter λ is chosen by the elastic net itself, whereas out of 100 values, the one that results in a minimal MSE. Equation 7.5 describes how a elastic net model $\mathbf{m}_{EN}$ predicts a value $\hat{y}_j$ mathematically based on the feature vector $\vec{\mathbf{x}}_j^{radiomics}$.

$$\hat{y}_j = \mathbf{m}_{EN}(\vec{\mathbf{x}}_j^{radiomics}) \tag{7.5}$$

Random Forest

The methodology of random forests follows the state-of-the-art described in Section 4.4. Before random forests use the radiomics features extracted, the algorithm z-scores them. This study uses the Python-based software scikit-learn [71] as random forest regressor. Each random forest consists of 100.000 trees. Compared to elastic nets, only one random forest is calculated per combination of TIL marker and setting since the random forest does not have a parameter like the elastic net's α. Equation 7.6 describes how a random forest model $\mathbf{m}_{RF}$ predicts a value $\hat{y}_j$ mathematically based on the feature vector $\vec{\mathbf{x}}_j^{radiomics}$.

$$\hat{y}_j = \mathbf{m}_{RF}(\vec{\mathbf{x}}_j^{radiomics}) \tag{7.6}$$

7.3.3 Visualizing Predictive Features

The voxel-based experiments use the results obtained from either elastic net or random forest and are only exemplary. They visualize the most predictive features of a result and illustrate their origin within the ROI. The steps taken to accomplish this are presented in the following.

1. Three results in total are chosen - one result per TIL - e.g., the results of elastic nets with $\alpha = 0.1$ for *PD1+ Density* in the setting *tumor/orig*. The following steps are done for each of the three results chosen.

2. The top 5 features of the result are identified. This is accomplished by focusing on the feature coefficients (if it is an elastic net model $\mathbf{m}_{EN}$) or the Gini-impurities (in the case of a random forest model $\mathbf{m}_{RF}$). Both indicate the importance of a feature, the higher the value, the more important the feature. As a result, these coefficients are ranked, and the top 5 entries are chosen.

3. Each of the 5 features chosen is re-calculated for each voxel of ROI used in the setting by PyRadiomics [30] (e.g., for each voxel of the *tumor* ROI in the example mentioned above). This results in one file per feature, which holds the information where that feature is how strong.

4. These files are in the `.nrrd` format and need to be converted to the NIFTI format used by the original files for visualizing them. This thesis uses a special Python-based tool (see [11]) for this step.

5. Furthermore, this work uses ANTs and FSL for rescaling the intensity of the voxel features and for aligning the images with the voxel features with the original imaging data.

Overlaying the original imaging data with the voxel images highlights the areas of the ROI where that feature is particularly strong. This illustrates the regions that give these top features their predictive power and show which parts of the ROI have a significant impact on predicting the TIL marker.

7.4 Deep Learning Approach

This section presents the methodology of the deep learning approach based on the state-of-the-art presented in Chapter 5. A task before using the deep learning model is applying image preprocessing and data augmentation. The deep learning model uses these further preprocessed and augmented images. This section provides the methods used for preprocessing, data augmentation, and the deep learning model. The experiments with the deep learning approach are not as far-reaching as the radiomics approach's experiments since this study considers the deep learning approach a proof of concept. Regarding the TIL markers available, the deep learning approach investigates only the density markers, i.e., *CD3+ Density*, *CD8+ Density*, and *PD1+ Density*.

7.4.1 Preprocessing & Data Augmentation

The preprocessed images (see Section 7.2) need further preprocessing for the CNN to use them. As the CNN uses the images directly as input, compared to the radiomics approach that uses the radiomics features extracted, the images need to be further reduced in size to achieve a feasible computational speed. First, the algorithm uses only a 2D slice instead of the entire 3D MRI sequence (similar to Shboul et al. [85]), whereas the algorithm chooses the 2D slice with the largest ROI in the horizontal plane.

After slicing all images, the algorithm rescales them to a consistent size. Afterward, augmentation of the images increases the image number for training the CNN. The algorithm applies the following data augmentations to each image:

1. Flipping horizontally,
2. Flipping vertically,
3. Flipping horizontally and vertically,
4. Rotating the original image 90° counter-clockwise,

5. Rotating the image flipped horizontally 90° counter-clockwise,
6. Rotating the image flipped vertically 90° counter-clockwise,
7. Rotating the image flipped horizontally & vertically 90° counter-clockwise,
8. Performing a random affine transformation to all of the above and the original image,
9. Performing random noise to the original image and all augmentations from (1-7), and
10. Performing a random elastic deformation to the original image and all augmentations from (1-7).

With this, the data augmentation creates 31 images for each original slice, resulting in 32 images per sample in the original data set. The algorithm uses PyTorch (`https://pytorch.org/`) for the augmentations 1-7 mentioned above, and TorchIO [74] for the augmentations described in 8-10.

7.4.2 Deep Learning Model

This study uses a modified PyTorch ResNet50 as a deep learning model $\mathbf{m}_{NN}$ for the experiments as Ma et al [60] achieved favorable results with a ResNet50 in a related study. The modifications concern the convolutional first and the fully-connected layer of the ResNet. The first convolutional layer's modification is necessary to accommodate the two images (one for each MRI sequence) as a single input with two channels to the ResNet50. The fully-connected layer's adaptation causes the output layer to consist of only one neuron, as the output is a single number. The CNN uses MAE as a cost function with the Adam optimizer. The ResNet50 used is not pretrained, causing the results to only depend on the input images. The algorithm tries to find applicable values for the parameters *learning rate* and *number of epochs* with a grid search. The possible values for the learning rate are $5 * 10^{-4}$, $5 * 10^{-5}$, and $5 * 10^{-6}$. The values for the number of epochs are 90, 100, 110, and 120.

Compared to the radiomics approach, the deep learning approach investigates fewer combinations of ROI and registration methods due to the characteristics of a CNN. The deep learning approach does not investigate the combinations using the wavelet filters, as the CNN uses the images directly as input and not features extracted. On the contrary, a CNN extracts the features from the images provided, which means investigating these combinations of ROI and registration method is unintended. Furthermore, this approach does not investigate all settings with the registration method *orig* since the CNN receives the two images as input channels. The images registered with the *orig* method are (most likely) not congruent as the ROI is registered and therefore altered. As a result, the images used as input channels would not be exactly of the same size or show the same part of the GBM. This concludes that the deep learning approach only investigates the remaining

four combinations of ROI and registration method: *tumor/flair*, *tumor/t1c*, *edema/flair*, and *edema/t1c*.

Equation 7.7 describes how a CNN $\mathbf{m}_{NN}$ predicts a value $\hat{y}_j$ mathematically based on the imaging data $\mathbf{I}_j^{NN}$.

$$\hat{y}_j = \mathbf{m}_{NN}(\mathbf{I}_j^{NN}) \tag{7.7}$$

7.5 Evaluation

This section presents the evaluation methods used in this thesis. Firstly, this section provides the evaluation methods used for the radiomics approach. Afterward, this section describes the evaluation methodology of the deep learning approach. This work uses the Python-based software library seaborn [99] for visualizing evaluation results.

7.5.1 Evaluation of Radiomics Approach

We evaluate the accuracy of predicting TIL markers from imaging data with $\mathbf{m}_{EN}$ and $\mathbf{m}_{RF}$. The evaluation methodology used by elastic nets and random forests is the same and builds upon cross-validation, which follows the state-of-the-art described in Section 4.5. With cross-validation, all samples are part of a test set at some point, which is advantageous for this study since a dedicated test set could contain only high or low values of a TIL marker if chosen at random. In addition, cross-validation allows testing the generalizability of the results with all samples. The experiments of elastic nets and random forests use LOOCV and 10-FCV. LOOCV uses all but one sample for training, which provides the machine learning method with more data for learning. On the contrary, 10-FCV allocates more samples for testing purposes, making it more representative.

The evaluation process correlates the TIL marker values for the test samples predicted with their corresponding ground truth values. This thesis uses the Spearman method to calculate the correlation, because it focuses on the monotonic relationship instead of the linear relationship [81], which is an advantageous property since the Python-based glmnet implements the naïve elastic net [25]. In addition, the Spearman coefficient is relatively robust against outliers [81]. A positive correlation indicates that the model learned (at least part of) the TIL marker values' distribution based on the radiomics features extracted. Due to that, a negative correlation would make no sense, as that would mean that the values predicted would decrease while the ground truth increases. As a result, a negative correlation can be regarded as no correlation since the model learned is not predictive.

For summarizing evaluations of the correlations achieved, results are regarded as predictive if the Spearman correlation $r > 0.2$ and $p < 0.05$. The elastic net algorithm trains multiple elastic nets for each TIL marker and setting due to the α parameter. The predictive results are counted for each TIL marker and setting combination, evaluating how robust the prediction is.

The stability of the predictive features chosen is subject to further evaluation. This evaluation builds on the feature coefficients of the elastic nets and the Gini-impurity of the random forests. In both cases, the values are averaged across all samples and normalized. Visualizing these averaged results allows a qualitative evaluation. The evaluation of the voxel-based experiments is qualitative and visual.

7.5.2 Evaluation of Deep Learning Approach

We evaluate the accuracy of predicting TIL markers from imaging data with $\mathbf{m}_{NN}$. The evaluation of the deep learning approach does not use cross-validation since that would be infeasible with a CNN. Instead, it builds upon splitting the available data into two subsets, the training set, and the test set. The test set contains 20% of the original data set, while the remaining 80% are part of the training data set. While the size of the test set is higher in related work, e.g., Fu et al. [26] uses about 25%, the total size of the data set in related work is higher as well. However, since data set available to this work is only half to about a third (depending on the TIL marker) of Fu et al.'s data set size, the training set size is increased to 80% to provide more information to the CNN during the training. The training process of the CNN uses only data from the training set. A randomized stratification algorithm chooses the samples for the test set. The stratification allows a more equalized representation of the data distribution by the test set since the target TIL marker values distribution is unequal. A non-stratified randomized choice of test samples bears the risk that the test set can contain only samples with high TIL marker values. In that case, the training set could consist of samples with lower TIL marker values only, which could cause the model learned to miss important information for samples in the test set. While the training uses all augmented images, the testing process only uses the original slices. Algorithm 7.1 illustrates the stratification process of the test set. The TIL marker values investigated are z-scored based on the samples in the training set.

Algorithm 7.1: ResNet Test set Stratification Algorithm

Input: A vector $\vec{\mathbf{x}}$ with the samples; a vector $\vec{\mathbf{y}}$ with the targets
Output: $\vec{\mathbf{x}}_{train}$ with the training samples; $\vec{\mathbf{x}}_{test}$ with the test samples

$n_{test} \leftarrow \lfloor n * 0.2 \rfloor$
$n_{bins} \leftarrow n \ / \ n_{test}$
$\vec{\mathbf{y}}_{sorted} \leftarrow sort(\vec{\mathbf{y}})$
$\vec{\mathbf{x}}_{test} \leftarrow \emptyset$
for $q \leftarrow 0$ **to** $n_{test} - 1$ **do**
    $z \leftarrow$ random number from $\lfloor q * n_{bins} \rfloor$ to $min(\lfloor (q+1) * n_{bins} \rfloor, n_{bins} - 1)$
    add element of $\vec{\mathbf{x}}$ corresponding with $\vec{\mathbf{y}}_{sorted_z}$ to $\vec{\mathbf{x}}_{test}$
end
$\vec{\mathbf{x}}_{train} \leftarrow \vec{\mathbf{x}} \ \setminus \ \vec{\mathbf{x}}_{test}$
return $\vec{x}_{train}$ *and* $\vec{x}_{test}$

At first, the evaluation process evaluates the CNNs trained with the samples for the training set. The process correlates the values predicted with their respective ground truth with the Spearman correlation method. The next step is choosing the parameter values for the learning rate and the number of epochs of the CNN with the highest correlation for each grid search. In addition, the significance of the CNN has to be lower than 0.05, i.e., $p < 0.05$. Afterward, the process evaluates the CNN chosen with the test set to assess the generalizability of the deep learning model. The evaluation of the test set correlates the values predicted with their corresponding ground truth with the Spearman method. A positive correlation indicates that the ResNet model learned generalizes well, while a correlation near 0 and a negative correlation suggests the model does not generalize well. The evaluation process also visualizes the gradients of the model's layers for a training and test sample. The visualizations build upon the PyTorch-based software tool (`https://github.com/vickyliin/gradcam_plus_plus-pytorch`) and provide Gradient-weighted Class Activation Mapping (GradCAM) [84] and GradCAM++ [20]. These visualizations allow a qualitative evaluation of areas where the predictive features originate.

7.6 Summary

This chapter introduces the methodology of this work. First, the targets of the predictions are provided. Following that, the image preprocessing steps taken are provided, including the ROI segmentation used in this study and the *skull stripping* process. The methods used by the radiomics approach are described, consisting of feature extraction methods and machine learning models (elastic nets and random forests) to analyze the features extracted. In addition, methods to visualize predictive features are presented. Then the deep learning approach used in this work is presented, which consists of further image preprocessing and data augmentation, and the deep learning model itself - a modified ResNet50. In the end, the methodology used for evaluating the prediction accuracy of the different approaches is introduced.

CHAPTER 8

Experiments & Results

This chapter describes and presents the results of the thesis. First, this chapter provides the results acquired through the experiments with the radiomics approach. The part about the radiomics results presents the experiments & results with elastic nets and random forests. Furthermore, that part shows the results of exemplary voxel-based experiments. Afterward, the results achieved with the deep learning approach are presented. This chapter only provides the presentation and description of the experiments and results, while the following chapter provides the discussion. We compare results for different ROI definitions, different features, and different MRI sequences.

The different ROIs provided and image registration methods used during the preprocessing allow various combinations. The labels used have a dedicated structure, which is

`<ROI><wavelet filters>/<registration method>`.

The first part provides the ROI used for that setting, followed by information about the use of wavelet filter features. The second part shows the registration method used during the image preprocessing. The three components can have the following values:

ROI:

- *edema* indicates that this setting uses the segmentation of the FLAIR sequence
- *tumor* indicates that this setting uses the segmentation of the T1c sequence

wavelet filters: The suffix *_wv* of the ROI used indicates the use of wavelet filters for that setting. If the ROI does not have such a suffix, only the radiomics features without the wavelet filters are used in that setting.

registration method:

- *flair* indicates that the method registers the T1c sequence of the patient onto the FLAIR sequence
- *t1c* indicates that the method registers the FLAIR sequence of that patient onto the T1c sequence
- *orig* indicates that neither sequence is registered, which means that this setting used both sequences with their respective original space

Due to these values, possible labels for settings are, e.g., *edema/orig*, *tumor_wv/t1c*, or *tumor/flair*.

8.1 Data

Clinicians use MRI scanners to capture the imaging data and the Definiens software (Definiens AG, Munich, Germany) to calculate the TIL markers from histology. Table 8.1 summarizes the number of samples used in the experiments. The number of samples available for a TIL marker counts how many samples have the *FLAIR* and *T1c* sequences and the corresponding TIL marker. The outlier column displays how many samples are outliers since these samples do not meet the rule mentioned in Section 7.1. The last column provides the number of samples for each TIL marker available for the experiments.

TIL marker	Samples available	Outliers	Total data samples
CD3+ Density	91	3	88
CD3+ Positive	91	3	88
CD3+ Negative	91	3	88
CD3+ Percentage	91	4	87
CD8+ Density	77	1	76
CD8+ Positive	77	0	77
CD8+ Negative	77	2	75
CD8+ Percentage	77	1	76
PD1+ Density	59	2	57
PD1+ Positive	59	3	56
PD1+ Negative	59	2	57
PD1+ Percentage	59	3	56

Table 8.1: Summarizes the data available for each TIL marker

8.2 Results of the Radiomics Approach

This section provides the results achieved with the radiomics approach. First, this section presents the results of the elastic nets. The second part shows the results acquired with random forests. Finally, the results of the visualization of predictive features in the imaging data are shown.

8.2.1 Results: Elastic Net

Figure 8.1 displays the results achieved by the experiments with elastic nets and 10-FCV. Figure 8.1a summarizes the correlations between the ground truth of the TIL values and their predicted counterparts for different combinations of features, ROI, and registration. The x-axis holds the different TIL markers, namely *Density*, *Positive*, *Negative*, and *Percentage* for *CD3+*, *CD8+*, and *PD1+*. The labels for the y-axis (rows of the heatmap) show the combination (features, ROI, registration) used for the experiment.

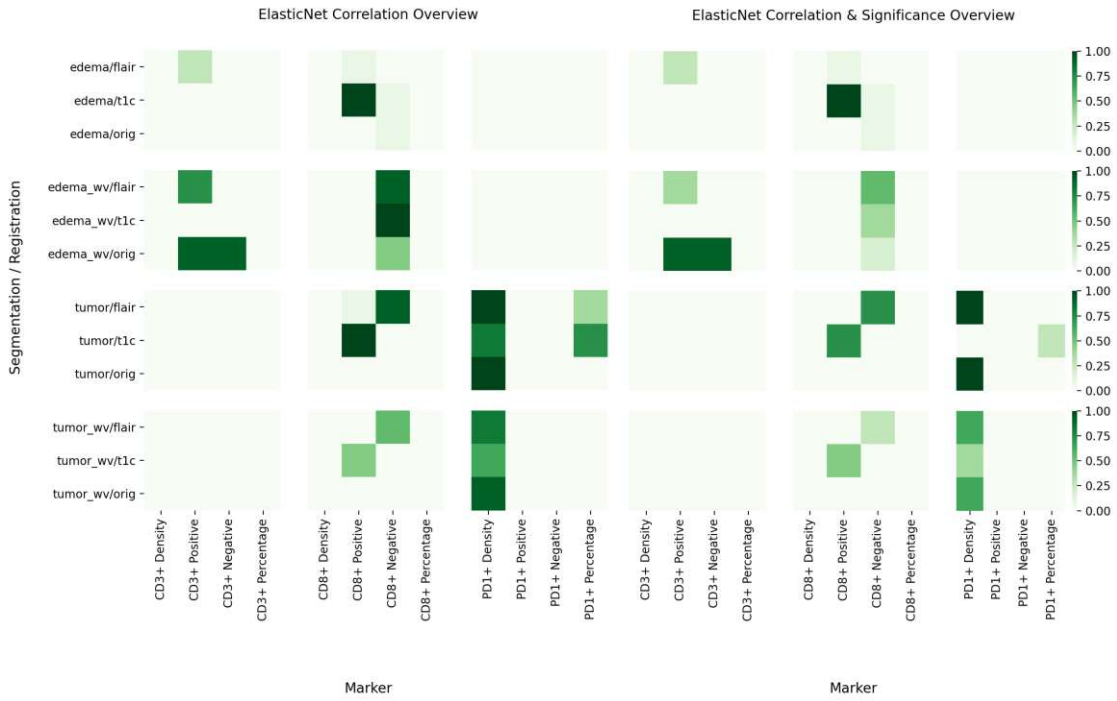
For each entry in the heatmap, e.g., for *PD1+ Density* in the setting *tumor/orig*, 11 elastic nets with $\alpha = \{0, 0.1, \ldots, 1\}$ are fitted. The elastic net with $\alpha = 0$ corresponds to a ridge regression, while an elastic net with $\alpha = 1$ corresponds to a LASSO [109]. Due to the absence of a dedicated test set, the evaluation method uses 10-FCV. The evaluation method calculates the Spearman correlation coefficient r of TIL values predicted with the ground truth provided. If $r > 0.2$, the elastic net with the α used is considered as predictive. Figure 8.1b is structured in the same way as Figure 8.1a. The resulting elastic nets for each entry in Figure 8.1b are considered predictive, if not only $r > 0.2$, but the corresponding p-value p meets $p < 0.05$ as well.

The results displayed indicate that *PD1+ Density* can be predicted as long as the *tumor* ROI is used. On the contrary, all markers of *PD1+* remain unpredictable if the *edema* ROI is part of the setting, likewise the markers of *CD3+* are unpredictable when using the *tumor* ROI. Moreover, *PD1+* is the only TIL where the markers *Density* and *Percentage* can be predicted, while *Positive* and *Negative* are unpredictable. Contrary to that, only the markers *Positive* and *Negative* can be predicted for the TILs *CD3+* and *CD8+*. The results displayed in both figures are similar, which shows that most elastic nets which have $r > 0.2$ also meet $p < 0.05$.

Experiments for the setting *tumor/orig*

The experiments provided here evaluate the predictability of the elastic nets for various TIL markers for the setting *tumor/orig*. Figure 8.2 displays the correlation results. The x-axis holds the different α values for the elastic nets, while the y-axis shows the TIL markers. A red cell indicates a positive correlation, and a blue cell a negative correlation. The darker the color, the stronger the correlation, i.e., a dark red cell indicates a stronger positive correlation, and a dark blue cell indicates a strong negative correlation.

The results displayed in Figure 8.2 show that elastic nets can predict *PD1+ Density* the best. Additionally, the results displayed reveal that the α parameter influences the



(a) (b)

Figure 8.1: Overview of the results achieved with the elastic nets and 10-FCV. The more elastic nets are predictive for a combination of TIL marker and setting, the darker the green tone.

correlation's strength, but not if a TIL marker is predictable (has a positive r) or not. This indicates that this is related to the TIL markers themselves and not to the choice of the α parameter.

Prediction accuracy of *PD1+ Density*

This part presents experiments and detailed results for the TIL marker *PD1+ Density* in the setting *tumor/orig*. The experiments provided here evaluate the predictability of the elastic nets for *PD1+ Density* in the setting *tumor/orig*. Figure 8.3 illustrates the detailed correlation results for the different α values of *PD1+ Density* in the *tumor/orig* setting. Each plot displays the ground truth, the prediction results, and the correlation resulting from them for the elastic nets with a certain α value. The header of each plot shows the α value used, the Spearman correlation coefficient r, and the significance of the correlation p. As an example, for the plot in the top right corner of Figure 8.3, a α value of 0.1 has been used and the resulting r is 0.423 with a p of 0.001. The x-axes display the ground truth labels, while the y-axes show the predicted values.

The correlations displayed in Figure 8.3 are similar, which suggests that *PD1+ Density*

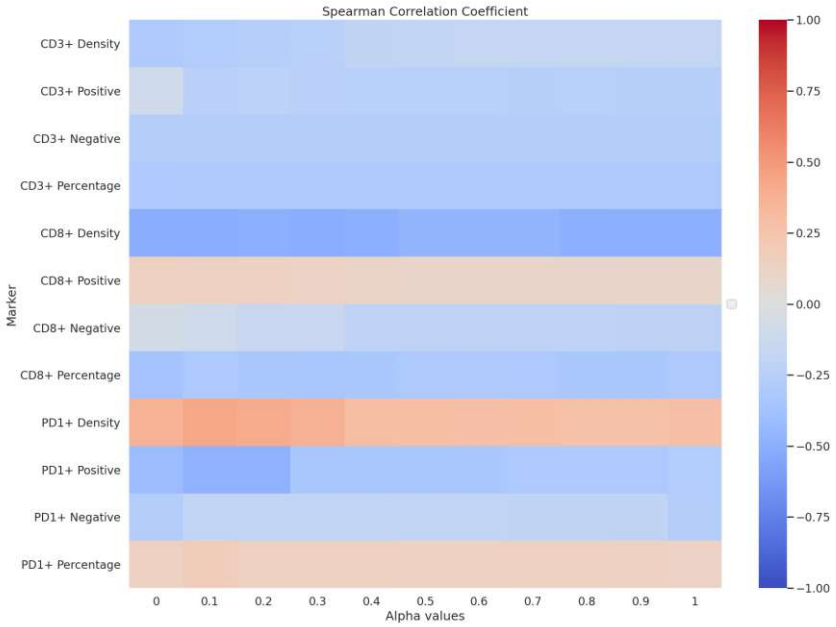


Figure 8.2: Correlation results of the elastic net with 10-FCV in the setting *tumor/orig*. The darker the color tone, the stronger the correlation, whereas red represents a positive correlation and blue a negative correlation.

can be predicted in the *tumor/orig* setting with elastic nets. However, the results also indicate that the α parameter influences the prediction accuracy achieved. A lower value for α, e.g., 0.1 or 0.2 causes the elastic nets to predict the TIL markers more accurately, compared to elastic nets using a higher value for α, e.g., 0.8 or 0.9.

Stability of features predicting for *PD1+ Density*

The experiments provided in this part evaluate the stability of the elastic nets for *PD1+ Density* in the setting *tumor/orig* in choosing predictive features. Figure 8.4 shows the predictive features chosen by the elastic nets for the setting *tumor/orig*. A feature is predictive and chosen by the elastic net algorithm if its coefficient is not zero. The x-axis of Figure 8.4 holds all the features, while the y-axis holds the different α values of the elastic nets. Counting how many times a feature coefficient is non-zero in the elastic nets results for each feature and α value creates the heatmap displayed in Figure 8.4. The darker the green tone of a cell is, the more often the corresponding feature coefficient is non-zero. As a result, the heatmap shows the features chosen and the choice's stability.

The results displayed in Figure 8.4 suggest that elastic nets choose the most predictive features independent of the α value used. This indicates the stability of the predictive features as elastic nets keep repeating their choice of features driving the prediction. Appendix A provides further results of the elastic net experiments. These outcomes

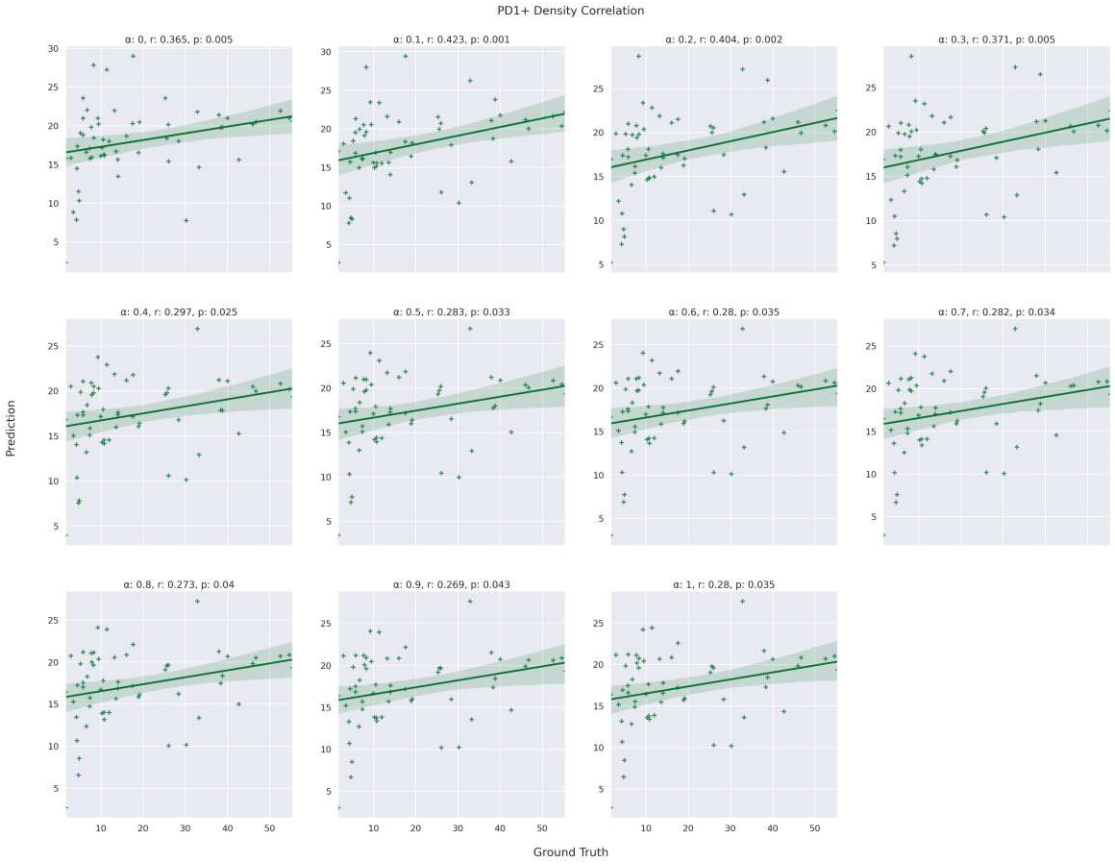


Figure 8.3: Correlation plots of the elastic nets with 10-FCV for the *PD1+ density* marker in the *tumor/orig* setting. A different α value is the basis for each result displayed in a plot. The header of each plot displays the α value, the Spearman correlation coefficient r, and the corresponding significance p. The x-axes display the ground truth labels, while the y-axes show the predicted values.

demonstrate that the radiomics approach using elastic nets can predict other TIL markers in different settings.

8.2.2 Results: Random Forest

Figure 8.5 displays the results achieved by the experiments with random forests and 10-FCV. Figure 8.5a displays the correlations between the TIL values' ground-truth and their predicted counterparts for multiple settings. The x-axis holds the different TIL markers, namely density, positive, negative, and percentage for *CD3+*, *CD8+*, and *PD1+*, e.g., the last column of the heatmap contains *PD1+ percentage*. The labels for the y-axis show the setting used for the experiments. Compared to the results of the elastic net and Figure 8.1a, Figure 8.5a displays only two green tones, since the random forests do not have a parameter α like an elastic net. Due to that, the correlation either meets the

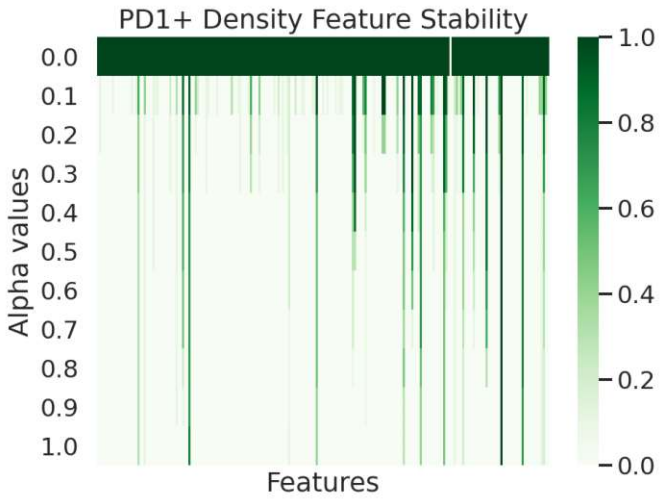


Figure 8.4: Stability of the features chosen by the elastic net for the *tumor/orig* setting and the *PD1+ density* marker for various α values. The darker the green tone, the more often the corresponding feature is considered predictive by the elastic nets. The x-axis holds the radiomics features extracted, while the y-axis holds the different α values.

threshold of $r > 0.2$ or does not.

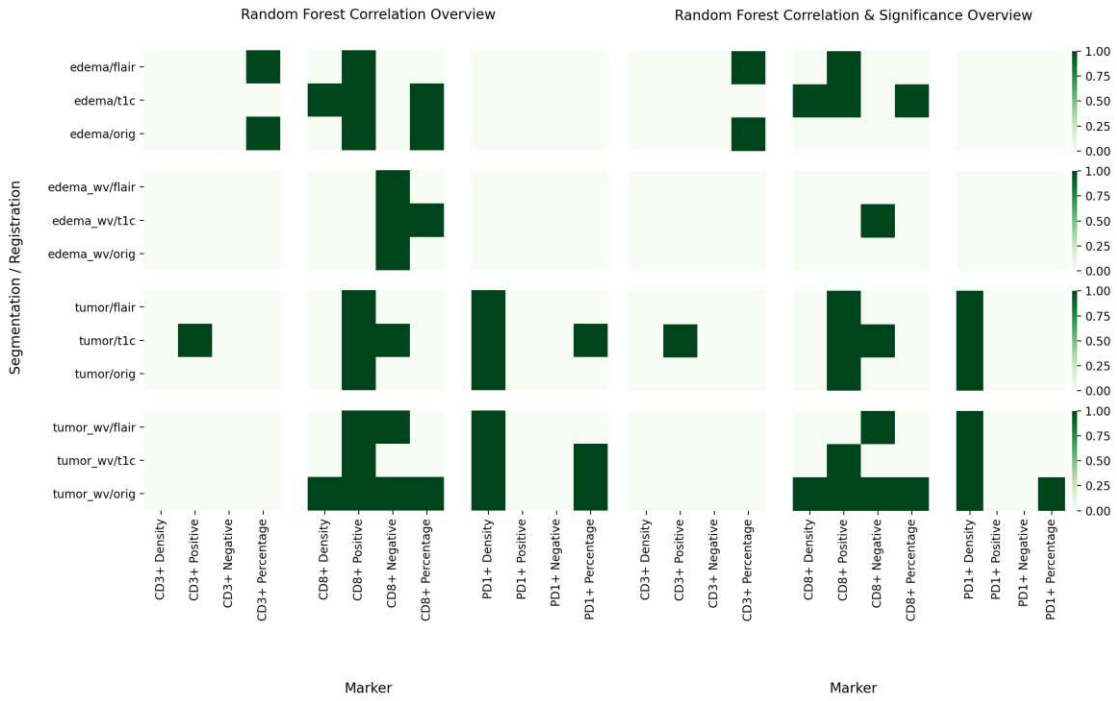
The results displayed in Figure 8.5b take the correlation and the significance into account. A result is predictive, if $r > 0.2$ and $p < 0.05$.

The results displayed indicate that *PD1+ Density* can be predicted as long as the *tumor* ROI is used. On the contrary, all markers of *PD1+* remain unpredictable if the *edema* ROI is part of the setting. Moreover, *CD8+* is the TIL whose markers can be predicted most with *CD8+ Positive* sticking out especially. Contrary to that, *CD3+* is almost unpredictable. Furthermore, the results are worse when using the ROI *edema* with the wavelet filters compared to the other settings. The results displayed in both figures are similar, which shows that the majority of random forests meet both criteria - $r > 0.2$ and $p < 0.05$.

Prediction accuracy for the *tumor/orig* setting

The experiments evaluate the prediction accuracy of random forests for the *tumor/orig* setting. Figure 8.6 shows the correlation between the TIL values predicted and the ground truth obtained with a random forest and 10-FCV. The predictions are along the y-axes of the plots, the ground truth along the x-axes. The title of each figure states the TIL, the Spearman correlation coefficient r, and the p value.

The results displayed in Figure 8.6 show that random forests can predict *PD1+ Density* and *CD8+ Positive*. However, random forests are not able to predict the other TIL markers in the *tumor/orig* setting, as the results displayed show that the values predicted



(a) (b)

Figure 8.5: Illustrates an overview of the results achieved with random forests and 10-FCV.

and the ground truth are uncorrelated or their correlation is not strong enough to be considered predictive.

Stability of predictive features for the *tumor/orig* setting

The experiments presented in this part evaluate the stability of the features chosen by random forests for all TIL markers in the setting *tumor/orig* in choosing predictive features. Figure 8.7 displays the mean values of the radiomics feature's Gini-impurities. The x-axis holds the Gini-impurities of the radiomics features, while the y-axis holds the TIL markers. The darker a green cell is, the higher the Gini-impurity of that feature for the TIL marker. With this, Figure 8.7 shows the importance of each feature (for every TIL marker) for the prediction of the TIL values.

The results displayed in Figure 8.7 suggest that random forests choose the most predictive features independent of the cross-validation fold, as there are only a few features with a high averaged Gini-impurity. This indicates the stability of the predictive features as random forests keep repeating their choice of features driving the prediction.

Appendix A provides further results of the random forest experiments. These outcomes demonstrate that the radiomics approach using random forests can predict other TIL

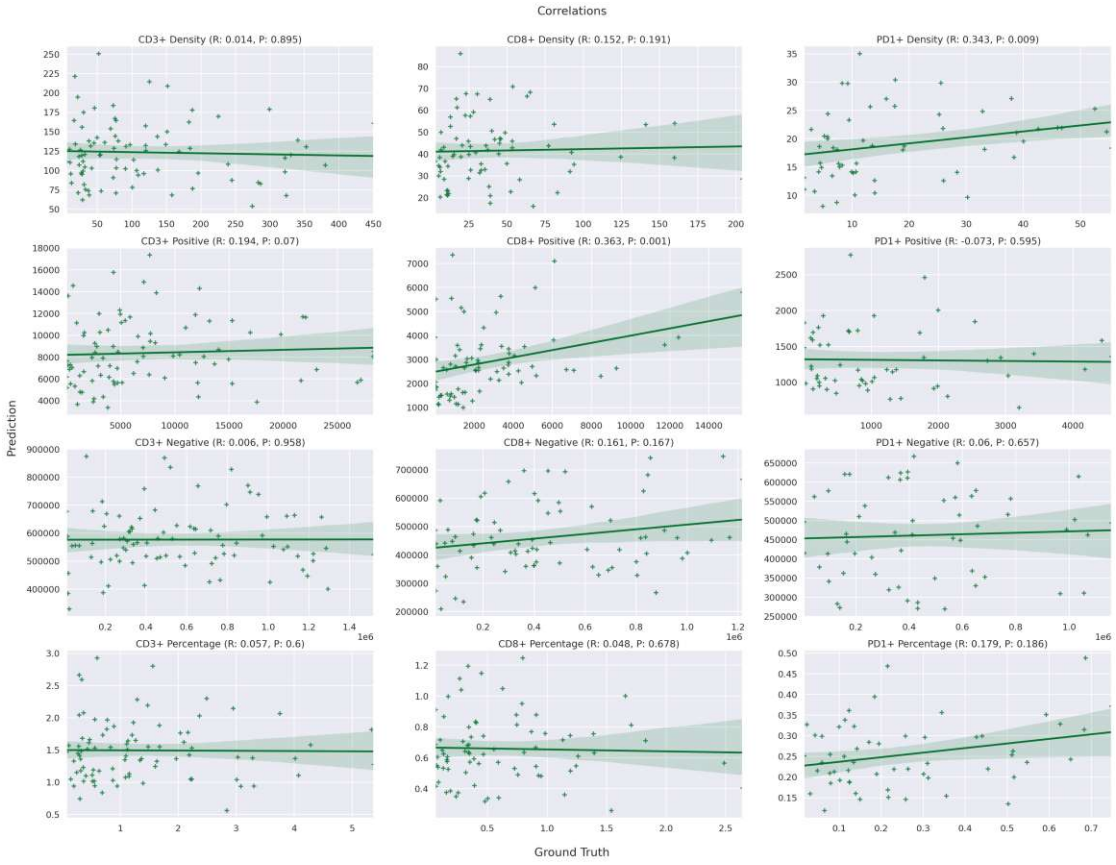


Figure 8.6: Correlations of the predicted TIL values and the ground truths for the *tumor/orig* setting. The experiments use a random forest and 10-FCV to obtain the predictions.

markers in different settings.

8.2.3 Visualizing predictive Features

This part presents exemplary results for voxel-based maps of features. These results illustrate the top image features predicting a TIL marker. The illustrations can provide novel medical insights into the relations between TILs and GBMs. The following parts present figures of the top 5 features found for the TIL markers *PD1+ Density*, *CD8+ Density*, and *CD3+ Percentage*. In addition, the parts provide a brief explanation of the top features investigated and displayed in the visualizations.

Features Predicting PD1+ Density

The voxel-based experiments illustrate the area of origin of the predictive features for *PD1+ Density* in the *tumor/orig* setting. The basis for these visualizations are results

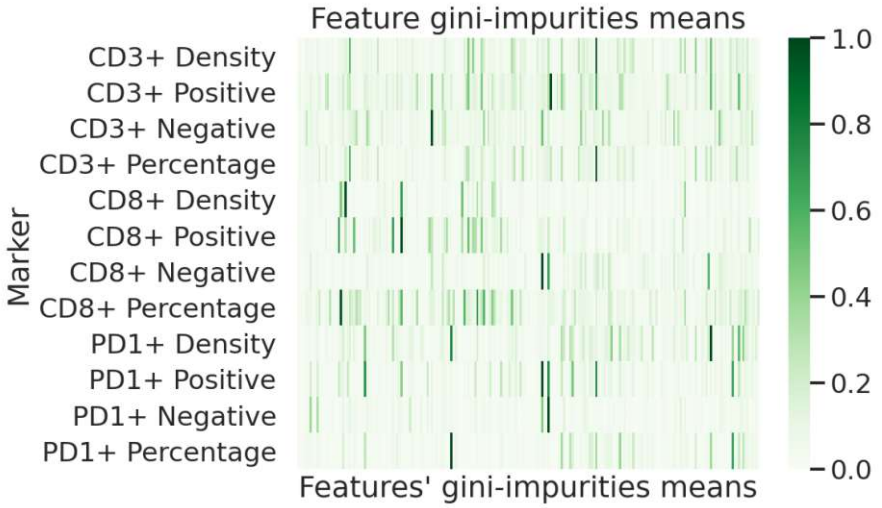


Figure 8.7: Averages of Gini-impurities of all radiomics features for every TIL marker of the *tumor/orig* setting. The darker a cell, the higher the average Gini-impurity for that feature. A higher Gini-impurity average indicates higher importance of the feature for predicting the TIL values.

obtained with the elastic net ($\alpha = 0.1$) with 10-FCV. Figure 8.8 displays the areas investigated with their respective ROIs and the results of these voxel-based experiments. Figure 8.8a shows the FLAIR sequence, Figure 8.8b depicts the segmentation with the FLAIR sequence. The T1c sequence is shown in Figure 8.8e with the Figure 8.8f displaying the sequence with ROI.

The Figures 8.8c and 8.8d display the top features found that originate from the FLAIR sequence. The Figures 8.8g, 8.8h and 8.8i display the top features found that originate from the T1c sequence. The illustrations display the features as colored overlays, whereas the more opaque the color is, the stronger the feature expressed in that voxel. Table 8.2 summarizes displays which plot shows which top feature, the sequence from which the feature originates, the feature's name, and the category to which the feature belongs. The top 5 features used express roughly the following information[1]:

Figure	Sequence	Feature	Feature Category
8.8g	T1c	*Strength*	NGTDM
8.8h	T1c	*ShortRunHighGrayLevelEmphasis*	GLRLM
8.8i	T1c	*HighGrayLevelRunEmphasis*	GLRLM
8.8c	FLAIR	*Strength*	NGTDM
8.8d	FLAIR	*MCC*	GLCM

Table 8.2: This table summarizes which top feature is depicted in which figure and to what radiomics feature category it belongs.

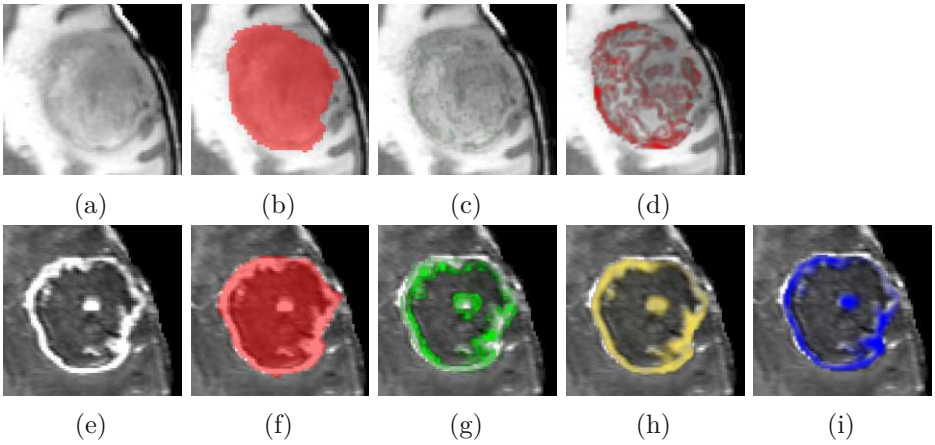


Figure 8.8: Illustrates parts of the FLAIR (a) and T1c (e) sequences showing a GBM case, with the respective ROIs (b and f). These images show how the sequences' parts look without overlay, while the ROIs indicate what area is part of the GBM. The Figures (c, d) illustrate the origin of top features from the FLAIR sequence. The Figures (g, h, i) illustrate the origin of top features from the T1c sequence. The more opaque the color overlay of a voxel is, the stronger the feature expressed at that voxel.

- *Strength* measures the image primitives, i.e., the value is high when the intensity changes slowly but with larger coarse differences in gray level intensities [73]. This means *Strength* measures if the brightness changes slowly but the brightness differences are high.
- *ShortRunHighGrayLevelEmphasis* measures the joint distribution of short lengths of pixels with the same high gray level value [73], which means it measures if there is a small number of consecutive pixels with the same high brightness value.
- *HighGrayLevelRunEmphasis* measures high gray level values' distribution [73], which means it measures if there is a high concentration of bright pixels.
- *MCC* measures how complex the texture of the image is [73], which means a more complex/inhomogeneous structure results in a higher value.

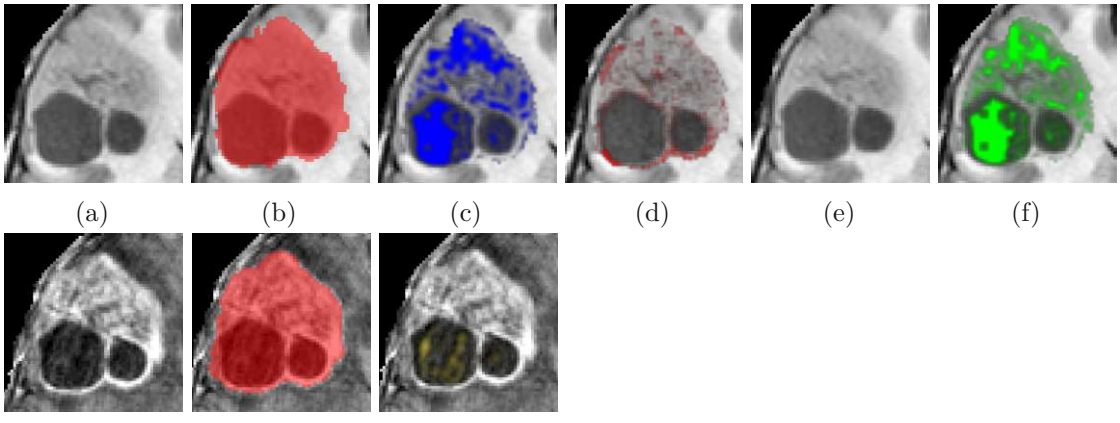
Features Predicting CD8+ Density

The voxel-based experiments illustrate the area of origin of the predictive features for *CD8+ Density* in the *tumor/orig* setting. The basis for these visualizations are results obtained with the random forest and LOOCV. Figure 8.9 displays the areas investigated with their respective ROIs and the results of these voxel-based experiments. Figure 8.9a shows the FLAIR sequence, Figure 8.9b depicts the segmentation with the FLAIR

[1] detailed information about the features can be found at [73]

sequence. The T1c sequence is shown in Figure 8.9g with the Figure 8.9h displaying the sequence with ROI.

Figure 8.9i displays the top features found originating from the T1c sequence. The Figures 8.9c, 8.9d, 8.9e and 8.9f display the top features found that originate from the FLAIR sequence. The illustrations display the features as colored overlays, whereas the more opaque the color is, the stronger the feature expressed in that voxel. Table 8.3



(a) (b) (c) (d) (e) (f)

(g) (h) (i)

Figure 8.9: Illustrates parts of the FLAIR (a) and T1c (g) sequences showing a GBM case, with the respective ROIs (b and h). These images show how the sequences' parts look without overlay, while the ROIs indicate what area is part of the GBM. The Figures (c, d, e, and f) illustrate the origin of top features from the FLAIR sequence. The Figure (i illustrates the origin of top features from the T1c sequence. The more opaque the color overlay of a voxel is, the stronger the feature expressed at that voxel.

summarizes displays which plot shows which top feature, the sequence from which the feature originates, the feature's name, and the category to which the feature belongs.

Figure	Sequence	Feature	Feature Category
8.9c	FLAIR	*LargeDependenceEmphasis*	GLDM
8.9d	FLAIR	*SmallDependenceHighGrayLevelEmphasis*	GLDM
8.9e	FLAIR	*Imc1*	GLCM
8.9f	FLAIR	*LongRunEmphasis*	GLRLM
8.9i	T1c	*LowGrayLevelRunEmphasis*	GLRLM

Table 8.3: This table summarizes which top feature is depicted in which figure and to what radiomics feature category it belongs.

The top 5 features used express roughly the following information[2]:

[2]detailed information about the features can be found at [73]

- *LargeDependenceEmphasis* measures the distribution of larger dependencies, whereas a dependency is the number of connected pixels within a certain distance that depend on the center voxel [73]. This means a larger value of *LargeDependenceEmphasis* indicates a more homogeneous texture.
- *SmallDependenceHighGrayLevelEmphasis* measures the joint distribution of small dependencies with high gray level values [73], which means it measures small areas with high brightness.
- *Imc1* measures the correlation of the texture's complexity [73].
- *LongRunEmphasis* measures the distribution of the length of consecutive pixels with the same gray level value [73], which means it measures if there are more coarse/homogeneous structural textures.
- *LowGrayLevelRunEmphasis* measures the distribution of low gray level values [73], which means it measures if there are darker areas in the image.

Features Predicting CD3+ Percentage

The voxel-based experiments illustrate the area of origin of the predictive features for *CD3+ Percentage* in the *edema/flair* setting. The basis for these visualizations are results obtained with the random forest and 10-FCV. Figure 8.9 displays the areas investigated with their respective ROIs and the results of these voxel-based experiments. Figure 8.10a shows the FLAIR sequence, Figure 8.10b depicts the segmentation with the FLAIR sequence. The T1c sequence is shown in Figure 8.10e with the Figure 8.10f displaying the sequence with ROI.

The Figures 8.10c and 8.10d display the top features found that originate from the FLAIR sequence. The Figures 8.10g and 8.10h display the top features found that originate from the T1c sequence. The illustrations display the features as colored overlays, whereas the more opaque the color is, the stronger the feature expressed in that voxel. The fifth top feature (*Flatness*) is not displayed here since it is based on the shape of the segmentation, and information about the shape is not available at the voxel level. Table 8.4 summarizes displays which plot shows which top feature, the sequence from which the feature originates, the feature's name, and the category to which the feature belongs.

Figure	Sequence	Feature	Feature Category
8.10c	FLAIR	*InverseVariance*	GLCM
8.10d	FLAIR	*DependenceVariance*	GLDM
8.10g	T1c	*InverseVariance*	GLCM
8.10h	T1c	*LongRunEmphasis*	GLRLM

Table 8.4: This table summarizes which top feature is depicted in which figure and to what radiomics feature category it belongs.

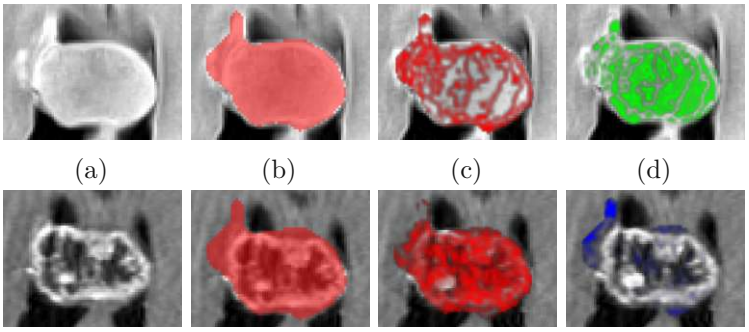


Figure 8.10: Illustrates parts of the FLAIR (a) and T1c (e) sequences showing a GBM case, with the respective ROIs (b and f). These images show how the sequences' parts look without overlay, while the ROIs indicate what area is part of the GBM. The Figures (c, d) illustrate the origin of top features from the FLAIR sequence. The Figures (g, h) illustrate the origin of top features from the T1c sequence. The more opaque the color overlay of a voxel is, the stronger the feature expressed at that voxel.

The top 4 features used express roughly the following information[3]:

- *InverseVariance* measures the inverse variance of the image [73].
- *DependenceVariance* measures the variance of the dependencies' sizes in the image [73]. This means *DependenceVariance* measures the variance of the different numbers of connected voxels within a certain distance that depend on the center voxel.
- *LongRunEmphasis* measures the distribution of the length of consecutive pixels with the same gray level value [73], which means it measures if there are more coarse/homogeneous structural textures.

8.3 Results of the Deep Learning Approach

This section provides the results achieved with the deep learning approach. The section contains three parts, one for each TIL marker investigated, i.e., *CD3+ Density*, *CD8+ Density*, and *CD8+ Density*. Each of these parts presents the experiments and the results for that TIL marker in a setting.

8.3.1 Experiments for *CD3+ Density* in the *edema/t1c* setting

This part presents the experiments and results for the TIL marker *CD3+ Density* in the *edema/t1c* setting. This segment provides the experiments and results achieved with the

[3] detailed information about the features can be found at [73]

training data at first, while it gives the results with the test data afterward. Note that the results on training data are provided as a reference and are not informative regarding how well the model would perform on unseen data.

Results with training data

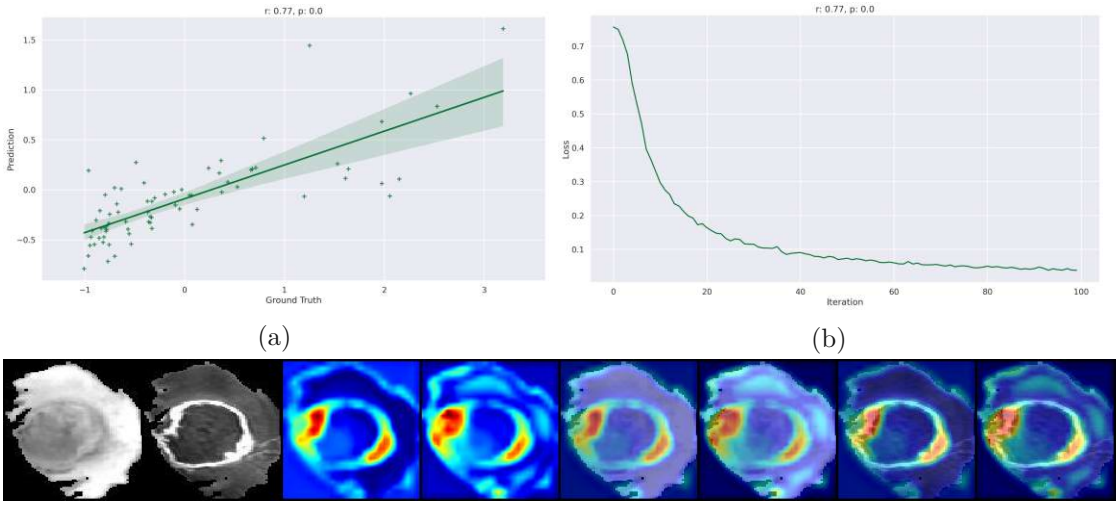
This segment provides the experiments and results achieved with the training data. The illustrations use the results obtained from the CNN with a parameter configuration leading to the highest Spearman correlation r. Figure 8.11a presents the correlation of the data used for the training of the CNN with a learning rate of $5 * 10^{-5}$ and 100 epochs. Figure 8.11b shows the loss' course throughout the epochs for the CNN. Figure 8.11c illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the training data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved for the training data in this setting is $r = 0.77$ with a significance of $p < 0.01$. The MAE of the last iteration is 0.0384.

Results with test data

This segment provides the experiments and results achieved with the test data. The illustrations use the same CNN used for the training data. Figure 8.12a presents the correlation of the test data. Figure 8.12b illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the test data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved is $r = 0.224$ with a significance of $p = 0.405$. The MAE of the test data is 0.6983.

8.3.2 Experiments for *CD8+ Density* in the *tumor/flair* setting

This part presents the experiments and results for the TIL marker *CD8+ Density* in the *tumor/flair* setting. This segment provides the experiments and results achieved with the training data at first, while it gives the results with the test data afterward. Appendix A holds the result figures of this part. Note that the results on training data are provided as a reference and are not informative regarding how well the model would perform on unseen data.



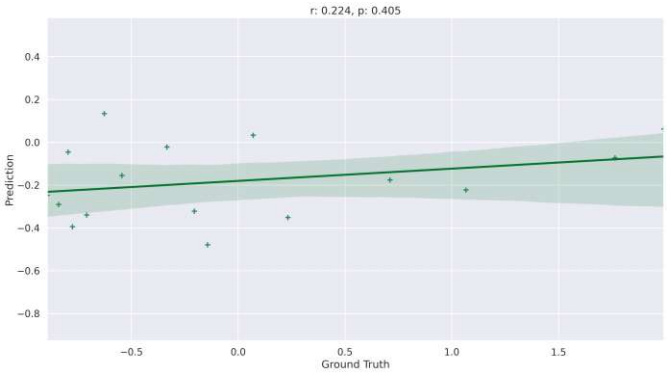
(a) (b)

(c)

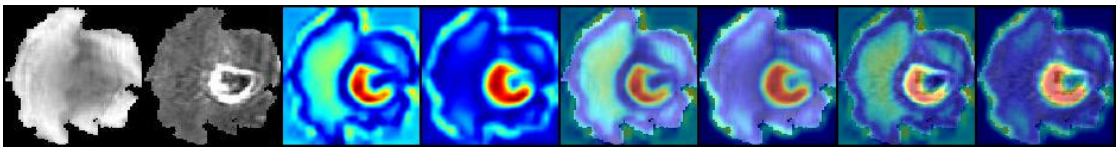
Figure 8.11: Illustrates the results with the training data for the *CD3+ Density* and *edema/t1c* setting. Figure (a) displays the correlation of the CNN with the height Spearman r, while Figure (b) shows the loss throughout the epochs. Figure (c) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (c) has a high *CD3+ Density* value.

Results with training data

This segment provides the experiments and results achieved with the training data. The illustrations use the results obtained from the CNN with a parameter configuration leading to the highest Spearman correlation r. Figure A.16a presents the correlation of the data used for the training of the CNN with a learning rate of $5 * 10^{-5}$ and 100 epochs. Figure A.16b shows the loss' course throughout the epochs for the CNN. Figure A.16c illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the training data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved is $r = 0.715$ with a significance of $p < 0.01$. The MAE of the last iteration is 0.0294.



(a)



(b)

Figure 8.12: Illustrates the outcomes with the test data for the *CD3+ Density* and *edema/t1c* setting. Figure (a) displays the correlation of the CNN with the height Spearman r. Figure (b) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (b) has a high *CD3+ Density* value.

Results with test data

This segment provides the experiments and results achieved with the test data. The illustrations use the same CNN used for the training data. Figure A.17a presents the correlation of the test data. Figure A.17b illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the test data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved is $r = 0.108$ with a significance of $p = 0.714$. The MAE of the test data is 0.6388.

8.3.3 Experiments for *PD1+ Density* in the *edema/flair* setting

This part presents the experiments and results for the TIL marker *PD1+ Density* in the *edema/flair* setting. This segment provides the experiments and results achieved with the training data at first, while it gives the results with the test data afterward. Appendix A holds the result figures of this part. Note that the results on training data are provided as a reference and are not informative regarding how well the model would perform on unseen data.

Results with training data

This segment provides the experiments and results achieved with the training data. The illustrations use the results obtained from the CNN with a parameter configuration leading to the highest Spearman correlation r. Figure A.18a presents the correlation of the data used for the training of the CNN with a learning rate of $5 * 10^{-5}$ and 120 epochs. Figure A.18b shows the loss' course throughout the epochs for the CNN. Figure A.18c illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the training data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved is $r = 0.695$ with a significance of $p < 0.01$. The MAE of the last iteration is 0.0435.

Results with test data

This segment provides the experiments and results achieved with the test data. The illustrations use the same CNN used for the training data. Figure A.19a presents the correlation of the test data. Figure A.19b illustrates the GradCAM [84] and GradCAM++ [20] of the first layer as heatmaps for a sample of the test data. From left to right, the first image shows the FLAIR sequence slice, the second the T1c slice, the third the GradCAM, and the fourth the GradCAM++. The fifth image illustrates the FLAIR slice with the GradCAM image as an overlay, while the sixth depicts the FLAIR slice with the GradCAM++ as an overlay. The seventh illustration presents the T1c sequence with the GradCAM images as an overlay, and the eighth shows the T1c sequence with a GradCAM++ overlay. The Spearman correlation achieved is $r = 0.212$ with a significance of $p = 0.556$. The MAE of the test data is 0.6887.

8.4 Summary

In this chapter, the evaluation results of the different approaches are shown. A quantitative evaluation shows that the radiomics approach - both elastic nets and random forests

- can predict TIL markers in various settings. The results further suggest an impact the settings (ROI chosen, registration method) have on the results. The quantitative evaluation shows that the predictive features chosen are stable, while the qualitative evaluation displays that predictive features tend to originate from specific parts of the ROIs or focus on specific textures. The results of the deep learning approach show that this method cannot accurately predict the TIL markers of the test set.

CHAPTER 9

Discussion

This chapter summarizes, compares, and discusses the results obtained through the experiments. In addition, this chapter provides further insights and results aside from the predictability of TILs. In the end, this chapter describes the limitations of this study.

9.1 Radiomics Approach

This part discusses the elastic net's and random forest's results before comparing them. Additionally, this chapter discusses the results of the exemplary voxel-based experiments.

9.1.1 Elastic Net Results

Figure 8.1b provides a useful overview of the results obtained with elastic nets, as it displays which TIL markers are predictable with a relevant significance. The figure demonstrates that some TIL markers (e.g., *PD1+ Density* in the setting *tumor/orig*) can be predicted constantly across various settings, while others do not seem to be predictable at all (e.g., *PD1+ Negative* - independent of the setting). The stability varies substantially among the predictable combinations of setting and TIL markers. Some results such as *CD8+ Negative* with *edema/t1c* are almost unpredictable, while others such as *PD1+ Density* with *tumor/orig* are completely stable and predictable.

Since the populations are rather small (56 to 88 patients, depending on the TIL marker), it might be useful to look at the results without taking the significance into account, as the size of the test set is rather small with 10-FCV (around 5 to 9 samples). Figure 8.1a summarizes these results. Comparing the results in the two figures shows their similarities, indicating that most of the predictable results can achieve statistical significance. Despite that, the differences between the figures suggest that some combinations may predict TIL marker values, but more data is needed to verify this trend.

The results displayed in both figures indicate that the segmentation chosen affects the predictability of the TIL markers. On one hand, all markers of *CD3+* remain unpredictable when using the *tumor* segmentation, on the other hand, that *tumor* segmentation appears to be vital for the prediction of any *PD1+* marker. This indicates, that the *edema* segmentation contains important information regarding *CD3+*, information that is missing in the *tumor* segmentation. Apparently, the features of *PD1+* lose their predictive power when the larger *edema* segmentation is chosen over the smaller *tumor* segmentation. On the contrary, *CD8+* appears unaffected by the segmentation used.

The choice of registration method during the preprocessing steps affects the predictability of the TIL markers. Especially the results shown in Figure 8.1b indicate, that *CD3+* and *PD1+* have their lowest predictability with an elastic net, when the *t1c* registration method is chosen. These results suggest that the *FLAIR* sequence transformed at the *t1c* registration method contains important information needed for a feasible prediction.

Figures like Figure 8.4 present the predictive features chosen by the elastic net. Apparently, the elastic nets choose the same features over and over again as predictive features, as the results show in Figure 8.4. The outcomes demonstrate the stability of the feature choice regarding the cross-validation and the different α. Despite the rising α values causing stricter rules for the elastic net, the same features are chosen again (although they are thinned out) instead of different ones. Especially with higher α values, the number of features chosen is low. The low number indicates that the same features are chosen independently of the cross-validation fold, suggesting that they do not originate from certain samples or a subpopulation. These results are not limited to one setting and TIL marker, since Figure A.6 illustrates the same behavior.

9.1.2 Random Forest Results

Figure 8.1b presents a result overview of the predictability of TIL markers with a feasible significance using random forests. The figure shows in which settings what TIL marker is predictable, e.g., *CD8+ Density* in the *tumor_wv/orig* setting can be predicted feasibly, but *CD8+ Density* can not be predicted with feasible significance in the *edema/flair* setting. Figure 8.5a summarizes TIL markers that are predictable regardless of the significance achieved. These results can be worth a look because of the small population sizes of the different TIL markers in the various settings.

The choice of segmentation seems to affect not only the results of the elastic nets but also the results achieved with the random forests. While *PD1+* appears to be unpredictable with a certain significance when the *edema* segmentation is used the results indicate that *CD8+* is unaffected of the segmentation chosen. Even though *CD3+* is predictable with feasible significance by using either segmentation, the amount of successful predictions for *CD3+* remains sparse, making the TIL almost unpredictable. The results presented in Figure 8.5a show that the almost unpredictability of *CD3+* is independent of the significance achieved by the results of the random forests.

While the choice of registration method seems to affect the results of the elastic nets, the results achieved by the random forests appear to be affected by the segmentation chosen. This can be seen in Figure 8.5b as TIL markers remain the barely unpredictable when the *edema_wv* segmentation is used and the results' significance is taken into account. However, more TIL markers are predictable with feasible significance when the *edema* segmentation is used, which indicates that the wavelet filters used for the segmentation *edema_wv* appear to worsen some results, e.g., *CD8+ Density* for the *edema/t1c* setting.

As the elastic net chooses the more predictive features, some features have a higher impact on the prediction result of a random forest than other features. Figure 8.7 illustrates which features are regarded as more predictive than others by the random forests for the setting *tumor/orig*. As the figure displays, a few features are driving the prediction of each TIL marker, while most of the other features have a negligible impact. The same features are among the more predictive features for most of the cross-validations' folds since the averages displayed in Figure 8.7 appear to be high for mostly those features. These results indicate the stability of the features' choice, as they appear to be independent of the samples in a cross-validation fold.

9.1.3 Comparison of the Radiomics Approaches

A comparison of the results achieved by the elastic nets and random forest shows their similarities.

- Some TIL markers can be predicted with feasible significance by both methods, but not all.
- The choice of segmentation appears to have a significant impact on prediction accuracy for the TILs *PD1+* and *CD3+*
- The methods choose most of the features regarded as predictive stably, which indicates that the information described by these features is crucial for the prediction, and not the result of few outlier examples.
- Both methods can almost constantly predict the TIL marker *PD1+ Density*, as long as they use the *tumor* segmentation.
- TIL *CD3+* appears to be the most unpredictable TIL among the three investigated in this work, as results indicating predictability (even ignoring significance) are rare.

There are some differences between the results achieved by both methods. For example, the random forests appear to achieve predictability for more TIL markers than the elastic nets. However, the common characteristics outweigh the differences, demonstrating that the radiomics approach can predict TIL markers, even though not all.

9.1.4 Voxel-based maps of relevant features

The voxel-based experiments investigate where the predictive features' origin in the GBM is. As these experiments support visualization of the results achieved, they are only exemplary. The results of the voxel-based results presented in Figure 8.8 illustrate the top 5 predictive features for the TIL marker *PD1+ Density*. The colored overlays illustrate the top features, with the intensity of the colored overlay at a voxel indicating the feature's strength. The features extracted from the T1c sequence focus on the (border of the) contrast-enhanced part of the GBM. Compared to that, the features extracted from the FLAIR sequence focus on the inhomogeneity of the ROI. These results indicate that a contract-enhanced area visible in the T1c sequence and an inhomogeneous texture displayed in the FLAIR sequence are important for predicting *PD1+ Density*.

Contrary to that, most of the top features of the TIL marker *CD8+ Density* displayed in Figure 8.9 originate from the FLAIR sequences. 2 of the top 5 features are almost congruent while coming from the same sequence and focusing on homogeneous textures, highlighting the importance of that characteristic. In addition, the feature extracted from the T1c sequence focuses on darker areas. As a result, homogeneous textures displayed in the FLAIR sequence and dark areas in the T1c sequence are significant for predicting *CD8+ Density*.

The top features for *CD3+ Percentage* displayed in Figure 8.10 seem to be extracted evenly from the FLAIR and T1c sequences. Compared to the top features of the other TIL markers investigated, the top features for *CD3+ Percentage* appear to be scattered over the complete *edema* segmentation used. Especially, a possible combination of the features presented in Figure 8.10d and Figure 8.10c appear to cover the entire segmentation, as both features seems to be weak in parts where the other is strong.

All in all, the voxel-based experiments allow a spatial interpretation of the predictive features extracted concerning the GBM. These visualizations provide insights into which regions of a GBM drive the different TIL markers. Especially the results found for *CD8+ Density* and *PD1+ Density* demonstrate this as their results are almost contradictory. The results of *CD8+ Density* focus on the FLAIR sequence, the results of *PD1+ Density* favor the T1c sequence, and the features extracted from the T1c sequence focus either on darker areas or on brighter ones. These findings suggest that different areas of a GBM are important for different TIL markers, indicating that the tissue displayed in MRI images is indeed related to the TIL markers.

9.2 Deep Learning Results

This section summarizes and discusses the results of the deep learning approach. The results of all three TIL markers presented show that the CNN can predict the values of the training data (see, e.g., Figure 8.11a). In addition, the course of the loss curve throughout the training epochs (see, e.g., Figure 8.11b) implies that the CNNs are learning to predict the TIL marker values based on the training images. Despite that, non of the ResNets

generalizes well enough to predict the test set with feasible significance ($p < 0.05$; see, e.g., Figure 8.12a). Even without considering the significance achieved, the correlation of the test results are substantially lower than the correlations achieved with the training data (compare, e.g, Figure 8.12a and Figure 8.11a).

Using GradCAM [84] and GradCAM++ [20], this study investigates the origin of the CNNs predictive power. The heatmaps created depict where the most important areas of the image are - for an image of the training set, e.g., Figure 8.11c and an image of the test set , e.g., Figure 8.12b. These visualizations suggest that the CNNs focus on the more prominent areas of the input images, e.g., the heatmaps displayed in Figure 8.11c resemble the contrast-enhanced region of T1c slices and slightly highlight the brighter area of the FLAIR image. The test sample illustrated in Figure 8.12b shows a similar distribution indicating that CNN potentially learned to focus on these features. The heatmaps based on results for another TIL marker and depicted in Figure A.18c and Figure A.19b show a different distribution. While the heatmaps of the training sample (see Figure A.18c) still focus on some characteristics of the GBMs displayed, the GradCAM heatmap of the test sample displayed in Figure A.19b does not show a focus on any part of the image. Even the GradCAM++ heatmap appears to focus just outside the GBM displayed in the images. These findings do not only demonstrate the poor generalizability of the CNN model learned but raise the question of what exactly the CNN learned here.

The results achieved with the deep learning approach suggest that the methodology used cannot predict TIL marker values for unseen MRI images. There are various possible reasons for these results, e.g., the potential overfitting of the CNNs on the training data or a loss of predictive information in the images due to image preprocessing. Moreover, the data sets available are comparatively small for CNNs, e.g., Bae et al. [5] use 166 samples which almost doubles the size of the largest data set available for this thesis (*CD3+ Density* has 88 samples available). Changing the methodology could improve the results with the test data, possibly achieving a model with a feasible generalizability.

9.3 Radiomics vs. Deep Learning

A comparison of the experiments and the results achieved with the radiomics and deep learning approach is non-trivial since the methodologies used for the two approaches differ in various aspects. While the radiomics approach utilizes cross-validation, the deep learning approach uses a dedicated test set due to time constraints regrading repeated retraining. The radiomics approach makes use of the entire (3D) ROI, while the deep learning approach has only access to a (2D) slice of the ROI. The deep learning approach only investigates the *Density* TIL markers, while the radiomics approach studies all of them.

Nevertheless, the methodology of the radiomics approach allows a prediction of the TIL marker values based on the MRI sequences. At the same time, the deep learning approach cannot achieve this goal. In addition, the voxel-based experiments of the radiomics approach reveal that the most predictive features mostly focus on prominent

aspects of the ROI (see, e.g., Figure 8.8h). However, a similar investigation of the results achieved with the deep learning approach shows that such features do not necessarily focus on prominent aspects (see, e.g., Figure A.19b).

9.4 Limitations

The methodology underlies the limitations discussed in this section. The size of the data set is a limitation, as up to 88 samples are problematic for the deep learning approach. Additionally, manual segmentations are a limitation for all experiments due to the segmentation's imprecision. Stratification can partly rebalance the unbalanced TIL marker values, but this does not fix the issue entirely. Machine learning models trained with unbalanced data sets focus on the more prominent part, which reduces the model's generalizability.

The deep learning approach does not use the entire ROI defined as an input but only the slice where the ROI is the largest, which is a limitation of the deep learning approach. Although using only one slice speeds up computation by a significant margin (and maybe using the complete ROI or even MRI is infeasible), this step drops a big amount of information. Removing so much information influences the results since the information outside the slice chosen is unavailable to the CNN. Thus it might be a reason for the particularly poor performance of CNN in our experiment.

9.5 Summary

This chapter discussed the evaluation results and insights gained from them. The results obtained with the radiomics approach are discussed, first the ones with elastic nets, and afterward the results obtained with random forests. Additionally, insights regarding the impact of the ROIs chosen or the stability of the predictive features are discussed. The methods of elastic nets and random forests and their results are compared, focusing on their similarities and differences. The visualizations of predictive features are discussed, and insights gained from them are presented. The evaluation results of the deep learning approach are discussed, including possible reasons for the prediction inability. A comparison of the two approaches is given, discussing their differences and comparing the results achieved. If one should be recommended it would be the radiomics approach since it can accurately predict the target values of new, unseen data samples.

CHAPTER 10

Conclusion & Future Work

This thesis introduces methods for predicting TIL marker values of GBMs based on MRI images. This study describes and evaluates a radiomics approach using either an elastic net or random forest and a deep learning approach. The methodology utilizes three different registration methods to register both sequences with one ROI into the same space during the preprocessing, as two ROIs are available for two MRI sequences. The radiomics approach utilizes handcrafted radiomics features extracted from an ROI in the MRI images preprocessed with cross-validation to avoid overfitting. The evaluation of the results predicted by the machine learning models focuses on their ability to predict TIL marker values based on features extracted and their robustness. The results demonstrate that some TIL markers are predictable based on the MRI images by the radiomics approach, while other TIL markers remain unpredictable. Especially the TIL marker *PD1+ Density* appears to be predictable as long as the experiments use the *tumor* ROI. As a density marker is the only marker investigated independent of the GBMs size/volume, the results achieved for the density marker are of interest from a medical point of view.

Investigating the results achieved by the radiomics approach further with voxel-based experiments yielded insights into the origin of the predictive features in the ROI. These voxel-based results show a particular characteristic of the top features for the prediction of *PD1+ Density* and *CD8+ Density*, suggesting that different areas of a GBM are predictive for varying TIL markers. From a medical point of view, the voxel-based results are of interest as the significant areas visualized can be identified by clinicians.

The deep learning approach utilizes a modified ResNet50 to learn the relations between 2D slices of the MRI images and the TIL markers. The deep learning model is independent of the handcrafted radiomics features since it uses the images preprocessed directly. At the same time, the deep learning approach does not utilize cross-validation to prevent overfitting but a splitting of the data set in train and a stratified test set. Even though the 2D slices are the only feasible option, the model misses possible predictive features

since this step removes most of the ROI. The results achieved with the deep learning approach indicate that a CNN cannot predict TIL marker values based on slices of MRI sequences in our experiment setting. Although the deep learning experiments show that the CNNs can learn an association between the MRI slices and the TIL marker values, they cannot predict new, unseen data samples. Moreover, the visualizations of the predictive areas reveal that the CNNs do not appear to focus on the more prominent areas as the top predictive features of the radiomics approach do. These findings suggest that the methodology chosen for the deep learning approach cannot produce results similar to the radiomics approach.

While this thesis demonstrates that TIL marker values are predictable with information derived from MRI images, future work can deepen the understanding and possibly answer questions that arose during this thesis. Possible future work can include research questions such as the following:

- Can the findings of this thesis be reproduced with a different data set?
- Why does the predictability of some TIL markers depend on the ROI used? Why is *PD1+* with the radiomics approach only predictable when using the smaller ROI?
- Can changes in the methodology for the deep learning approach cause a CNN to predict the TIL markers based on the MRI images?
- How can a deep learning approach use more information of the ROI feasibly?
- Is it possible to visualize the difference between higher and lower TIL marker values like the top predictive features by the voxel-based experiments?

APPENDIX A

Result Appendix

This appendix provides further results of this thesis. First, this appendix presents additional outcomes of the elastic net experiments. The appendix gives more results of the random forest experiments afterward. The end of this chapter gives further results of the CNN experiments.

A.1 ElasticNet Results

This section provides further results of the elastic net experiments. Figures are structured as described in the results chapter. The results presented originate from different ROI and registration method combinations and TIL markers, e.g., Figure A.1 displays the correlation overview for the elastic nets of the *edema_wv/orig* setting, while Figure A.5 shows the correlations for *CD8+ Positive* with the *edema/t1c* setting.

Figure A.7 displays the results for the *tumor/orig* setting where the elastic nets use LOOCV instead of 10-FCV. The results presented show, that some TIL markers can be predicted (e.g., *PD1+ Density*) with LOOCV while others cannot (e.g., *CD3+ Density*).

A.2 Random Forest Results

This section presents additional results of the random forest experiments. Figures are structured as described in the results chapter. The results presented originate from different ROI and registration method combinations, e.g., Figure A.10 displays the correlation overview for the random forests of the *tumor/flair* setting, while Figure A.12 shows the correlations of the *edema/flair* setting.

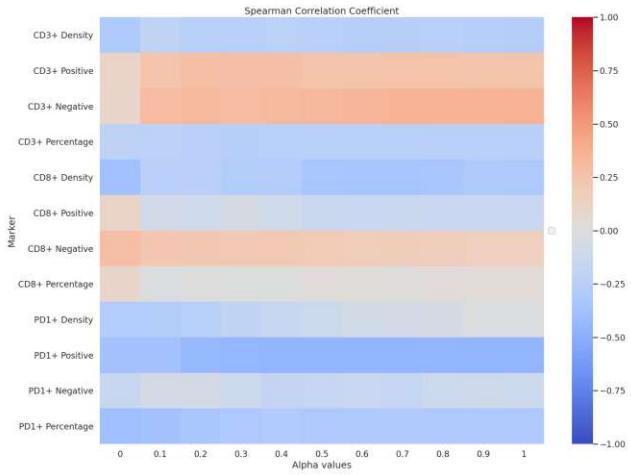


Figure A.1: Correlation results of the elastic net with 10-FCV in the setting *edema_wv/orig*. The darker the color tone, the stronger the correlation, whereas red represents a positive correlation and blue a negative correlation.

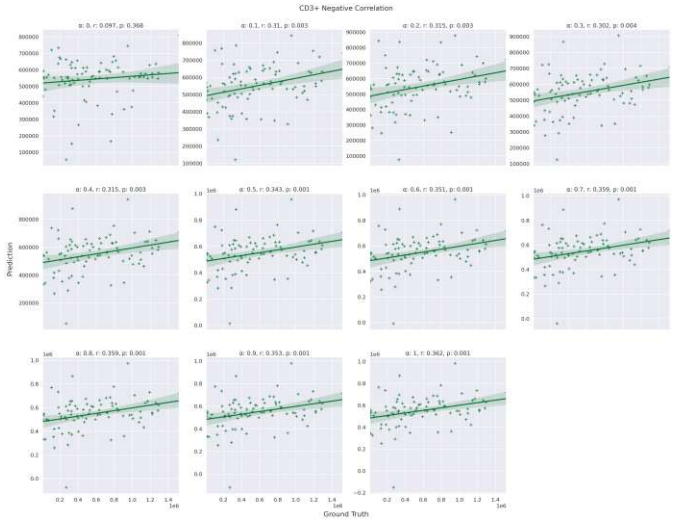


Figure A.2: Correlation plots of the elastic nets with 10-FCV for the *CD-3 negative* marker in the *edema_wv/orig* setting. Each plot is obtained with a different α value, whereas that α, the Spearman correlation coefficient r, and the corresponding significance p are stated in the header of each plot. The ground truth is along the x-axes, while along the y-axes the results predicted are displayed.

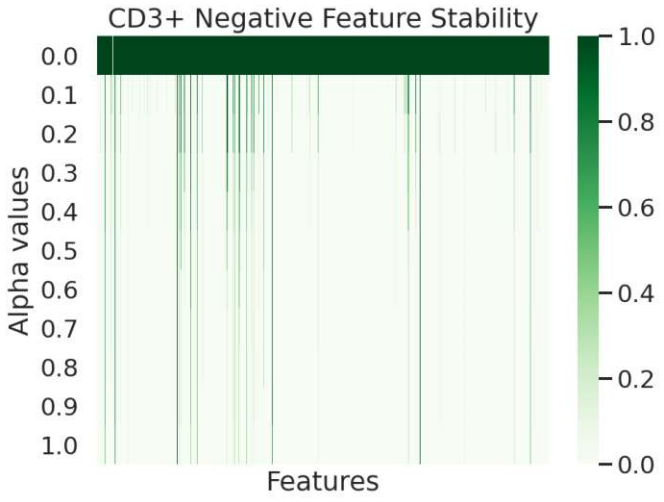


Figure A.3: Stability of the features chosen by the elastic net for the *edema_wv/orig* setting and the *CD-3 negative* marker for various α values. The darker the green tone, the more often the corresponding feature is considered predictive by the elastic nets. The x-axis holds the radiomics features extracted, while the y-axis holds the different α values.

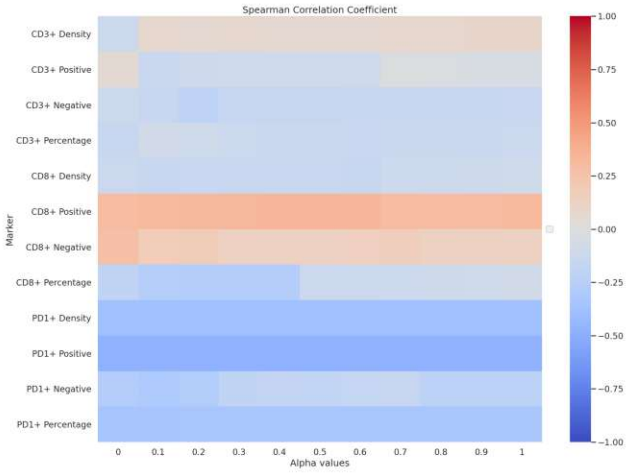


Figure A.4: Correlation results of the elastic net with 10-FCV in the setting *edema/t1c*. The darker the color tone, the stronger the correlation, whereas red represents a positive correlation and blue a negative correlation.

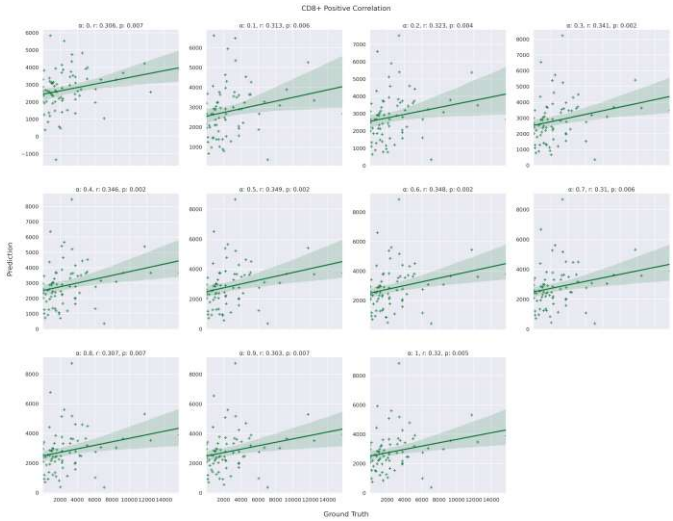


Figure A.5: Correlation plots of the elastic nets with 10-FCV for the *CD-8 positive* marker in the *edema/t1c* setting. Each plot is obtained with a different α value, whereas that α, the Spearman correlation coefficient r, and the corresponding significance p are stated in the header of each plot. The ground truth is along the x-axes, while along the y-axes the results predicted are displayed.

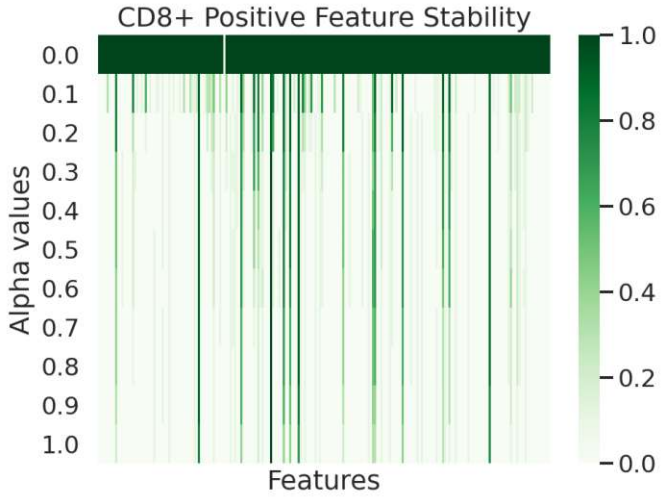


Figure A.6: Stability of the features chosen by the elastic net for the *edema/t1c* setting for various α values. The darker the green tone, the more often the corresponding feature is considered predictive by the elastic nets.

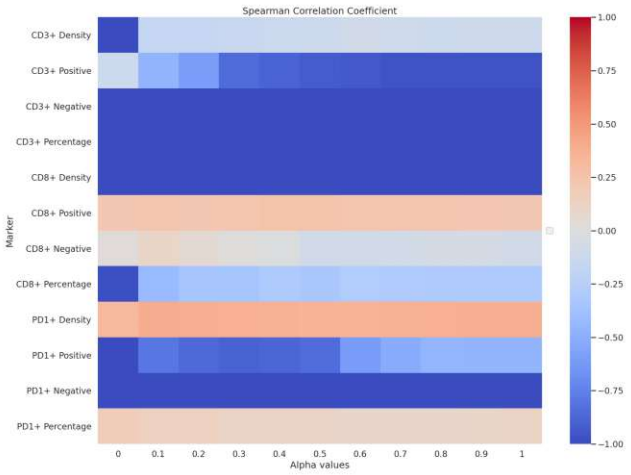


Figure A.7: Correlation results of the elastic net with LOOCV in the setting *tumor/orig*. The darker the color tone, the stronger the correlation, whereas red represents a positive correlation and blue a negative correlation.

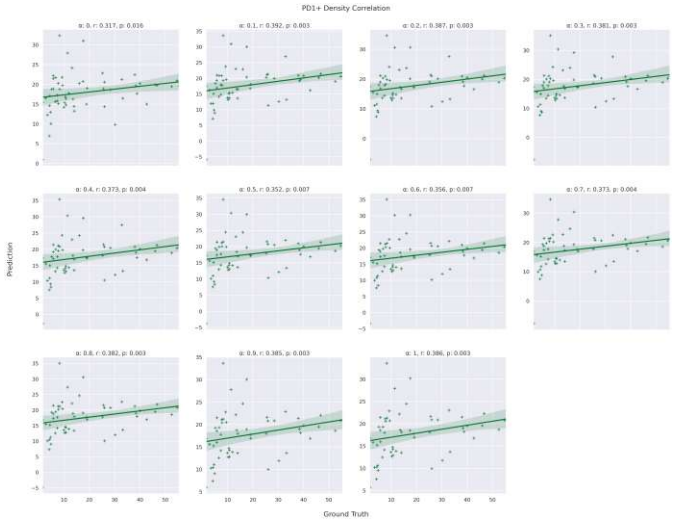


Figure A.8: Correlation plots of the elastic nets with LOOCV for the *PD-1 density* marker in the *tumor/orig* setting. Each plot is obtained with a different α value, whereas that α, the Spearman correlation coefficient r, and the corresponding significance p are stated in the header of each plot. The ground truth is along the x-axes, while along the y-axes the results predicted are displayed.

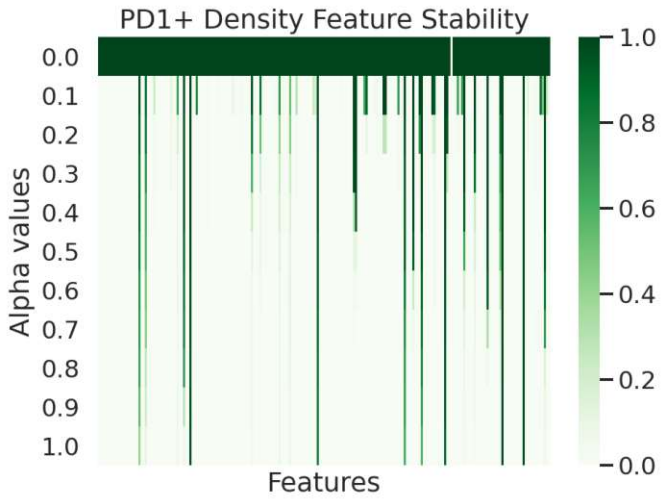


Figure A.9: Stability of the features chosen by the elastic net for the *tumor/orig* setting and the *PD-1 density* marker for various α values. The darker the green tone, the more often the corresponding feature is considered predictive by the elastic nets. The x-axis holds the radiomics features extracted, while the y-axis holds the different α values.

A.3 Deep Learning Results

This section presents additional results of the CNN experiments. Figures are structured as described in the results chapter.

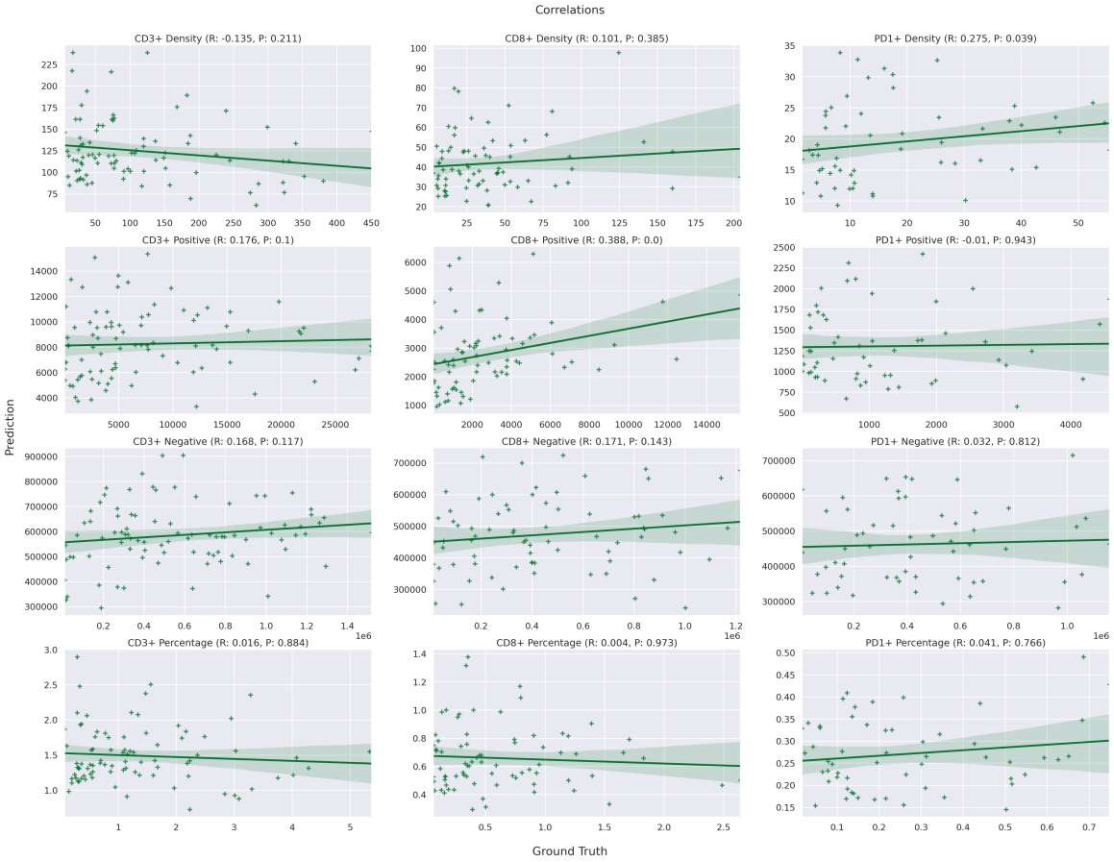


Figure A.10: Correlations of the predicted TIL values and the ground truths for the *tumor/flair* setting. The predictions are obtained with a random forest and 10-fold cross validation.

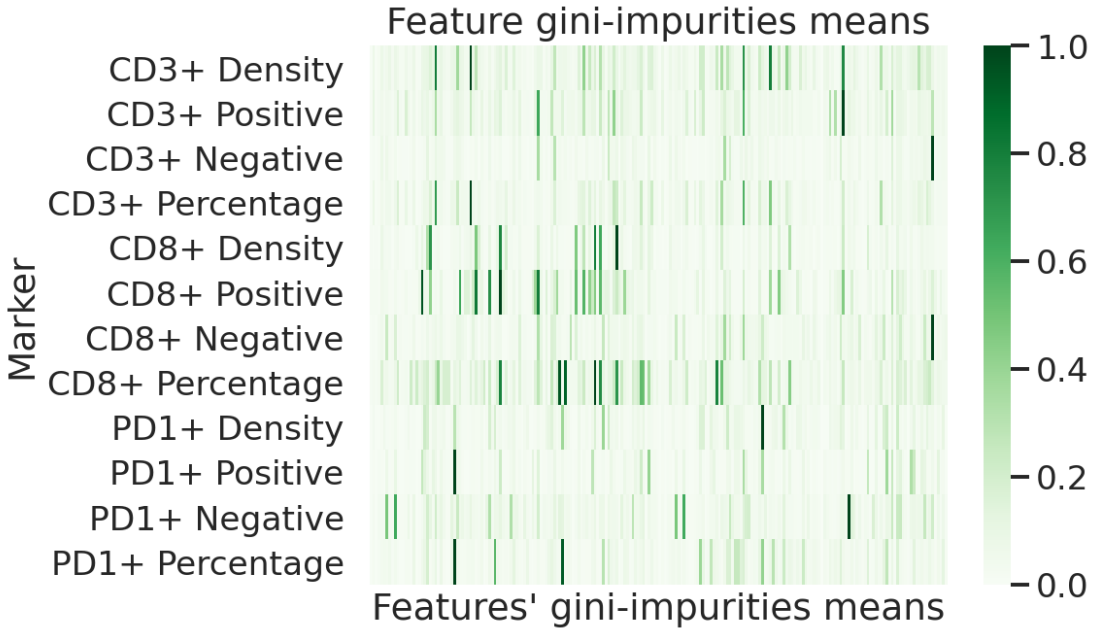


Figure A.11: Gini-impurity means of all radiomics features for every TIL marker of the *tumor/flair* setting. The darker a cell, the higher the average gini-impurity for that feature. A higher gini-impurity mean indicates a higher importance of the feature for the prediction of the TIL values.

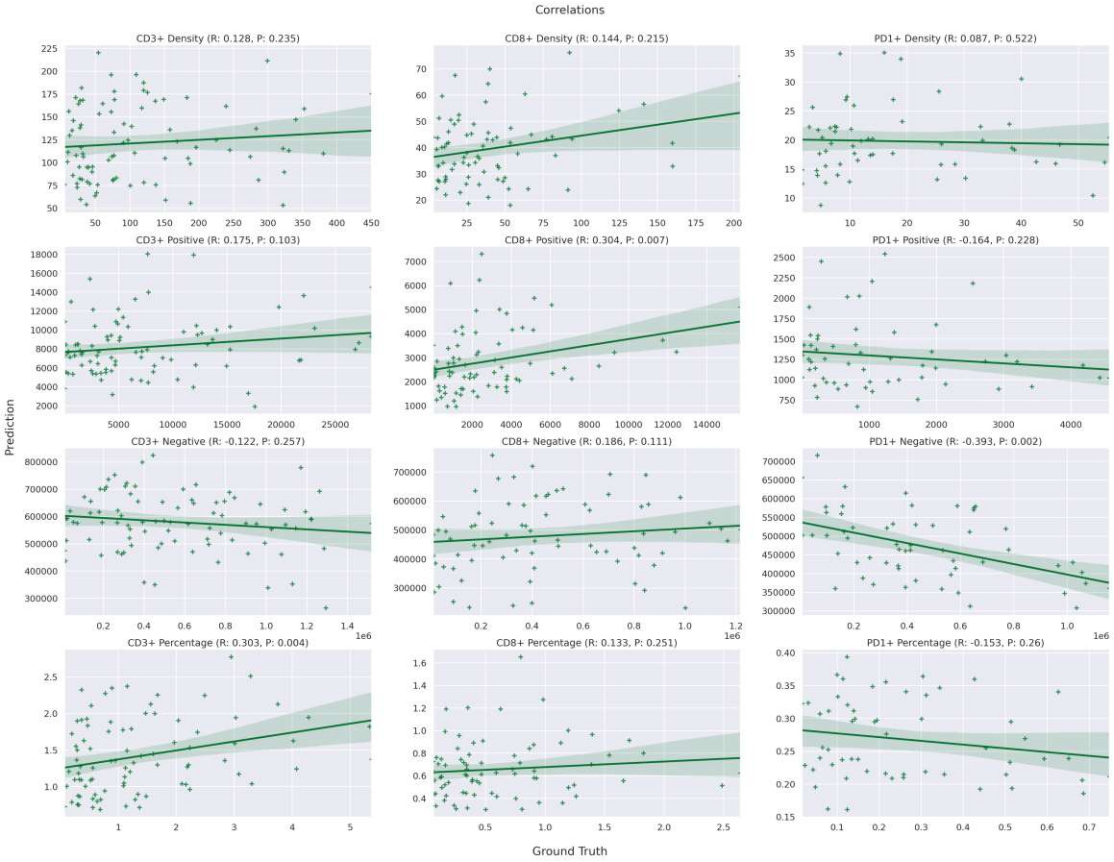


Figure A.12: Correlations of the predicted TIL values and the ground truths for the *edema/flair* setting. The predictions are obtained with a random forest and 10-fold cross validation.

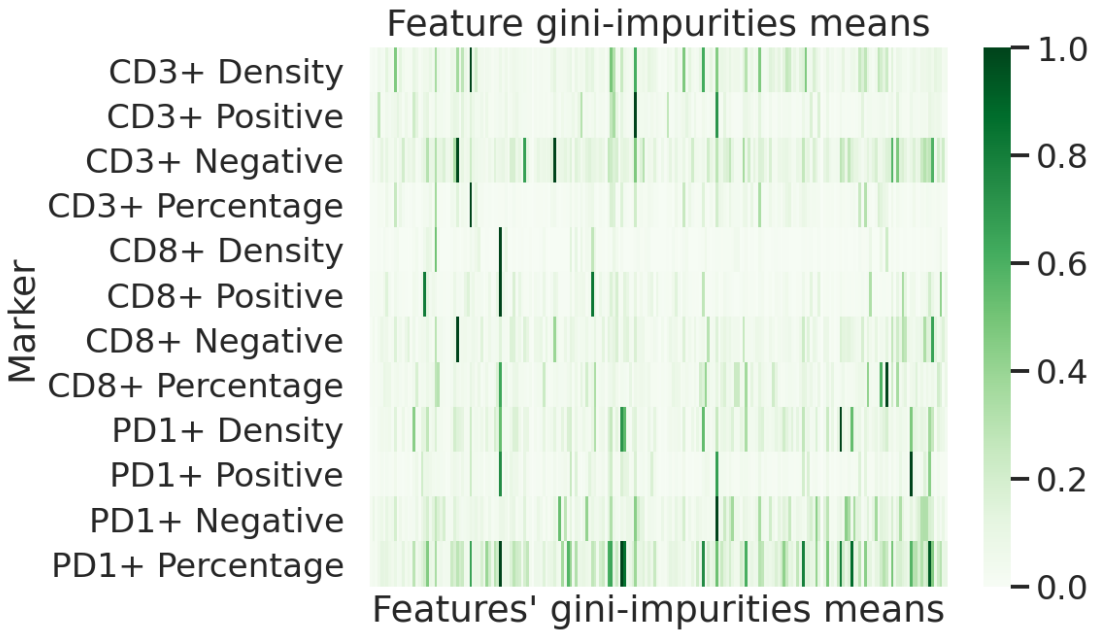


Figure A.13: Gini-impurity means of all radiomics features for every TIL marker of the *edema/flair* setting. The darker a cell, the higher the average gini-impurity for that feature. A higher gini-impurity mean indicates a higher importance of the feature for the prediction of the TIL values.

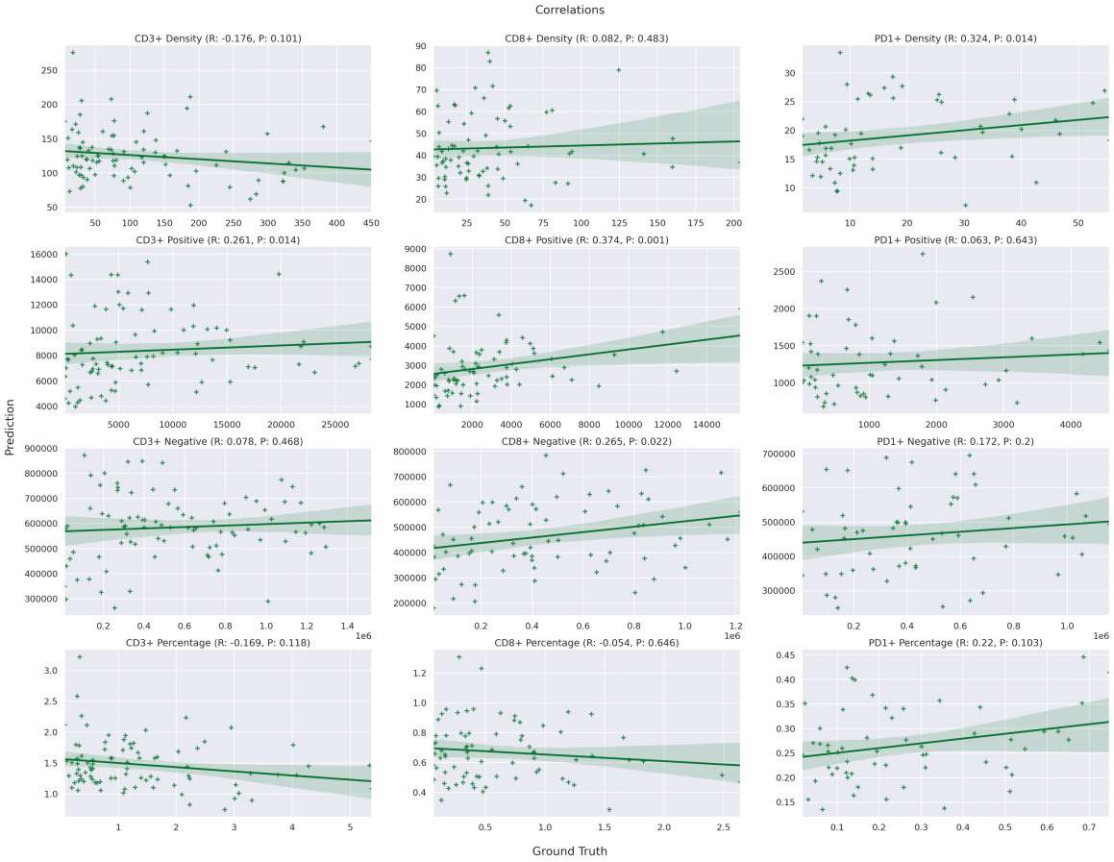


Figure A.14: Correlations of the predicted TIL values and the ground truths for the *edema/t1c* setting. The predictions are obtained with a random forest and 10-fold cross validation.

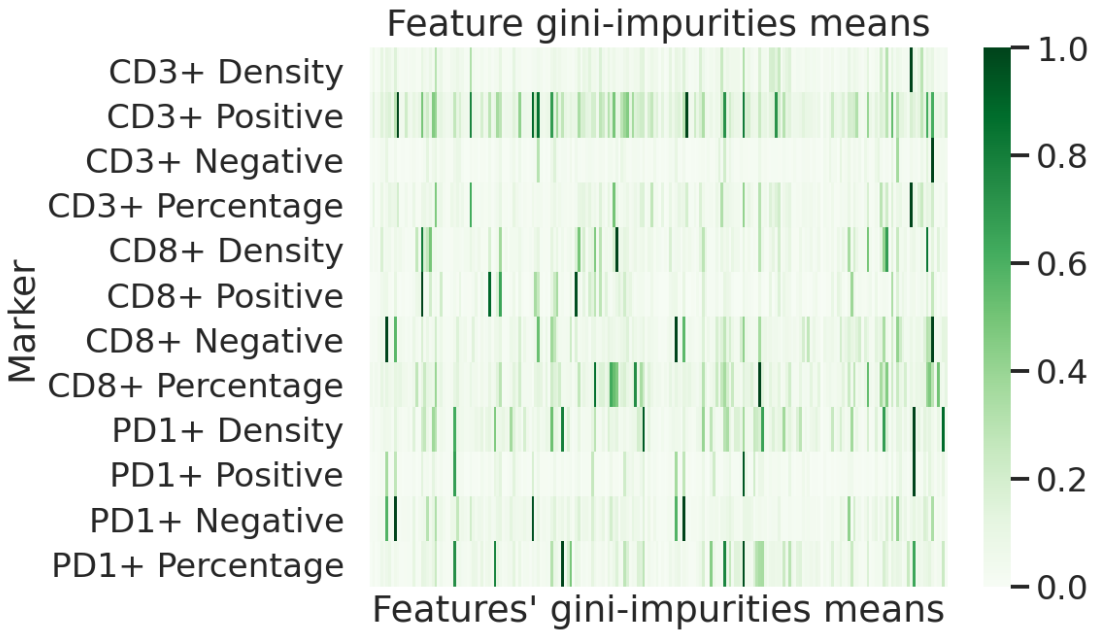
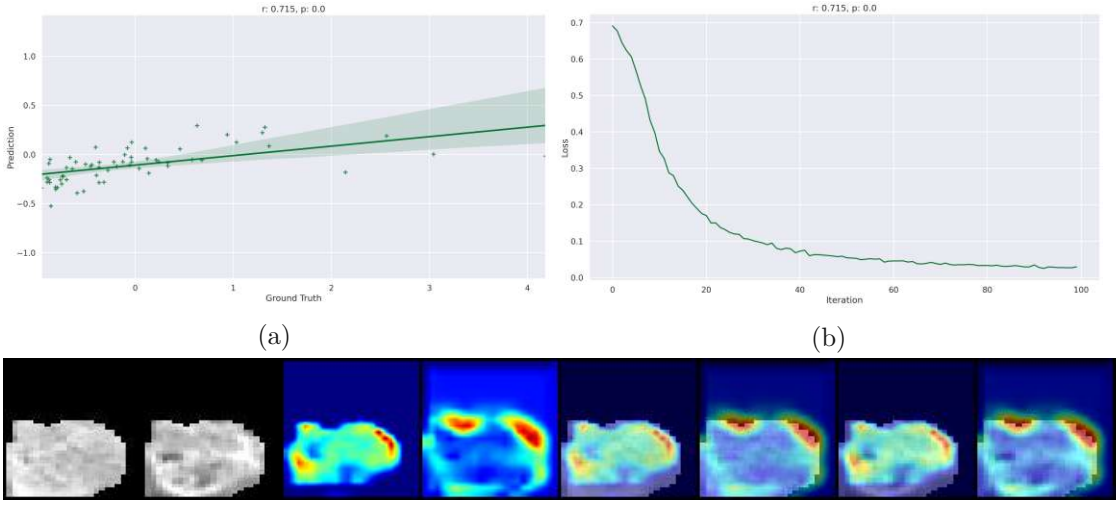


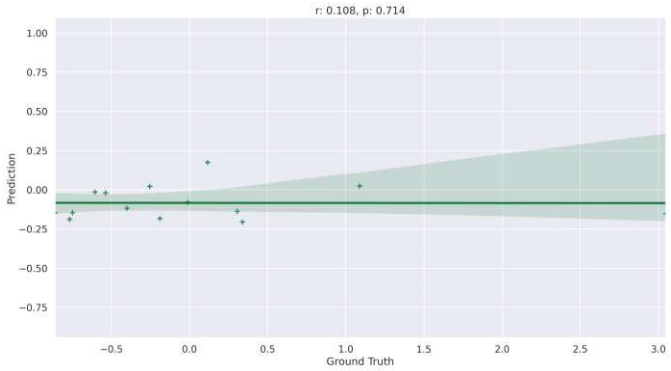
Figure A.15: Gini-impurity means of all radiomics features for every TIL marker of the *edema/t1c* setting. The darker a cell, the higher the average gini-impurity for that feature. A higher gini-impurity mean indicates a higher importance of the feature for the prediction of the TIL values.



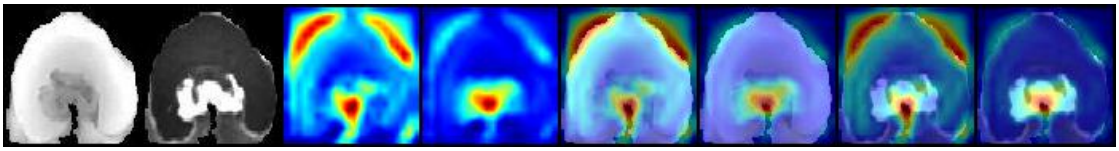
(a) (b)

(c)

Figure A.16: Illustrates the results with the training data for the *CD8+ Density* and *tumor/flair* setting. Figure (a) displays the correlation of the CNN with the height Spearman r, while Figure (b) shows the loss throughout the epochs. Figure (c) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (c) has a high *CD8+ Density* value.

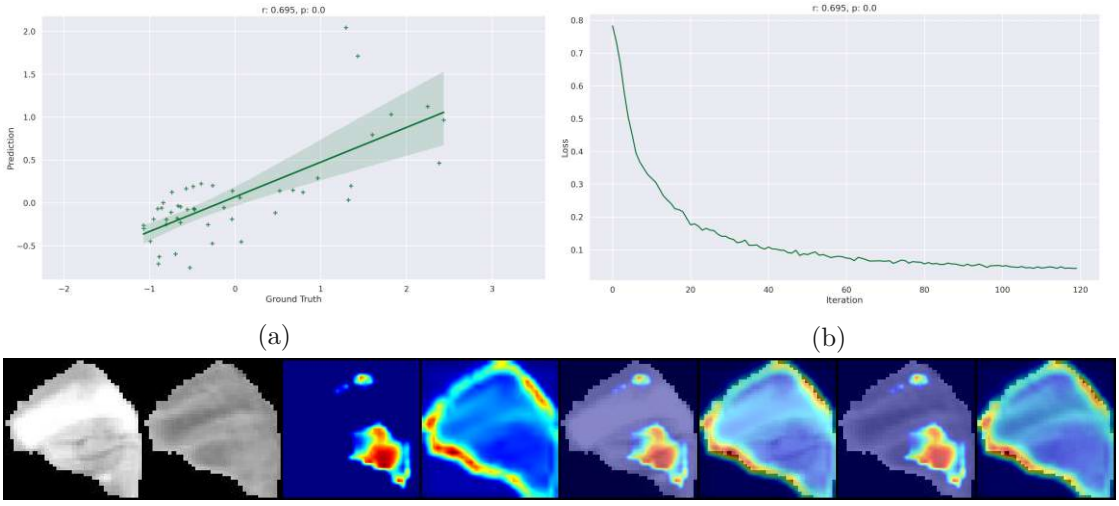


(a)



(b)

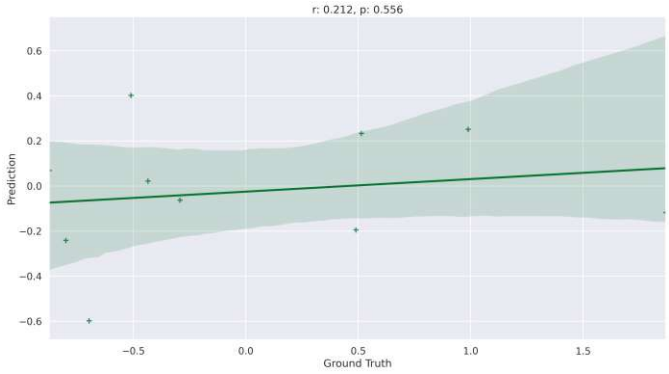
Figure A.17: Illustrates the outcomes with the test data for the *CD8+ Density* and *tumor/flair* setting. Figure (a) displays the correlation of the CNN with the height Spearman r. Figure (b) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (b) has a high *CD8+ Density* value.



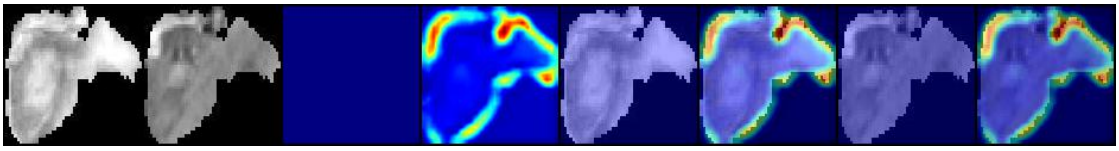
(a) (b)

(c)

Figure A.18: Illustrates the results with the training data for the *PD1+ Density* and *edema/flair* setting. Figure (a) displays the correlation of the CNN with the heighest Spearman r, while Figure (b) shows the loss throughout the epochs. Figure (c) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (c) has a low *PD1+ Density* value.



(a)



(b)

Figure A.19: Illustrates the outcomes with the test data for the *PD1+ Density* and *edema/flair* setting. Figure (a) displays the correlation of the CNN with the heighest Spearman r. Figure (b) presents the GradCAM and GradCAM++ of the first layer, where the first two image (from left to right) are the FLAIR and T1c slices, followed by the GradCAM and the GradCAM++ illustrations as heatmaps. The other images depict the slices with GradCAM/GradCAM++ as overlay: 5. FLAIR & GradCAM, 6. FLAIR & GradCAM++, 7. T1c & GradCAM, and 8. T1c & GradCAM++. The sample presented in Figure (b) has a low *PD1+ Density* value.

Bibliography

[1] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach", *Nature Communications*, vol. 5, no. 1, p. 4006, Jun. 3, 2014, Number: 1 Publisher: Nature Publishing Group, DOI: 10.1038/ncomms5006.

[2] A. C. Al-Rikabi, M. O. Al-Sohaibani, A. Jamjoom, and M. M. Al-Rayess, "Metastatic deposits of a high-grade malignant glioma in cervical lymph nodes diagnosed by fine needle aspiration (FNA) cytology—case report and literature review", *Cytopathology*, vol. 8, no. 6, pp. 421–427, 1997, DOI: 10.1111/j.1365-2303.1997.tb00573.x.

[3] G. Alatrash, H. Jakher, P. D. Stafford, and E. A. Mittendorf, "Cancer immunotherapies, their safety and toxicity", *Expert Opinion on Drug Safety*, vol. 12, no. 5, pp. 631–645, Sep. 1, 2013, DOI: 10.1517/14740338.2013.795944.

[4] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. Kam-Thong, J. Siebourg-Polster, B. Steiert, and J. D. Zhang, "An introduction to machine learning", *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 871–885, 2020, DOI: 10.1002/cpt.1796.

[5] S. Bae, C. An, S. S. Ahn, H. Kim, K. Han, S. W. Kim, J. E. Park, H. S. Kim, and S.-K. Lee, "Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: Model development and validation", *Scientific Reports*, vol. 10, no. 1, p. 12 110, Jul. 21, 2020, DOI: 10.1038/s41598-020-68980-6.

[6] S. Bae, Y. S. Choi, S. S. Ahn, J. H. Chang, S.-G. Kang, E. H. Kim, S. H. Kim, and S.-K. Lee, "Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction", *Radiology*, vol. 289, no. 3, pp. 797–806, Oct. 2, 2018, DOI: 10.1148/radiol.2018180200.

[7] A. Berger, "Magnetic resonance imaging", *BMJ : British Medical Journal*, vol. 324, no. 7328, p. 35, Jan. 5, 2002, DOI: 10.1136/bmj.324.7328.35.

[8] D. Berrar, "Cross-validation", in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 542–545, DOI: 10.1016/B978-0-12-809633-8.20349-X.

[9] Y. Bian, Y. F. Liu, H. Jiang, Y. Meng, F. Liu, K. Cao, H. Zhang, X. Fang, J. Li, J. Yu, X. Feng, Q. Li, L. Wang, J. Lu, and C. Shao, "Machine learning for MRI radiomics: A study predicting tumor-infiltrating lymphocytes in patients with pancreatic ductal adenocarcinoma", *Abdominal Radiology*, vol. 46, no. 10, pp. 4800–4816, Oct. 1, 2021, DOI: 10.1007/s00261-021-03159-9.

[10] G. Biau and E. Scornet, "A random forest guided tour", *Test*, vol. 25, no. 2, pp. 197–227, 2016, DOI: 10.1007/s11749-016-0481-7.

[11] T. Billah, S. Bouix, and Y. Rathi, "Various MRI Conversion Tools", 2019, DOI: 10.5281/zenodo.2584003.

[12] T. A. Birbilis, G. K. Matis, S. G. Eleftheriadis, E. N. Theodoropoulou, and E. Sivridis, "Spinal metastasis of glioblastoma multiforme: An uncommon suspect?", *Spine*, vol. 35, no. 7, E264, Apr. 1, 2010, DOI: 10.1097/BRS.0b013e3181c11748.

[13] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1, 1996, DOI: 10.1023/A:1018054314350.

[14] ——, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 1, 2001, DOI: 10.1023/A:1010933404324.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Boca Raton: Routledge, Oct. 25, 2017, 368 pp., DOI: 10.1201/9781315139470.

[16] M. L. D. Broekman, R. Risselada, J. Engelen-Lee, W. G. M. Spliet, and B. H. Verweij, "Glioblastoma multiforme in the posterior cranial fossa in a patient with neurofibromatosis type i", *Case Reports in Medicine*, vol. 2009, e757898, Dec. 16, 2009, Publisher: Hindawi, DOI: 10.1155/2009/757898.

[17] C. M. d. Castro-Costa, R. W. B. d. Araújo, M. A. d. Arruda, P. M. d. Araújo, and E. G. d. Figueiredo, "Increased intracranial pressure in a case of spinal cervical glioblastoma multiforme: Analysis of these two rare conditions", *Arquivos de Neuro-Psiquiatria*, vol. 52, pp. 64–68, Mar. 1994, Publisher: Academia Brasileira de Neurologia - ABNEURO, DOI: 10.1590/S0004-282X1994000100011.

[18] A. Chaddad, M. J. Kucharczyk, P. Daniel, S. Sabri, B. J. Jean-Claude, T. Niazi, and B. Abdulkarim, "Radiomics in Glioblastoma: Current Status and Challenges Facing Clinical Implementation", *Frontiers in Oncology*, vol. 9, 2019, Publisher: Frontiers, DOI: 10.3389/fonc.2019.00374.

[19] J. E. Chang, D. Khuntia, H. I. Robins, and M. P. Mehta, "Radiotherapy and radiosensitizers in the treatment of glioblastoma multiforme", *Clinical Advances in Hematology & Oncology*, vol. 5, no. 11, pp. 894–902, 2007.

[20] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks", in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2018, pp. 839–847, DOI: 10.1109/WACV.2018.00097.

[21] S. W. Choi, H.-H. Cho, H. Koo, K. R. Cho, K.-H. Nenning, G. Langs, J. Furtner, B. Baumann, A. Woehrer, H. J. Cho, J. K. Sa, D.-S. Kong, H. J. Seol, J.-I. Lee, D.-H. Nam, and H. Park, "Multi-Habitat Radiomics Unravels Distinct Phenotypic Subtypes of Glioblastoma with Clinical and Genomic Significance", *Cancers*, vol. 12, no. 7, Jun. 27, 2020, DOI: 10.3390/cancers12071707.

[22] B. D. Coene, J. V. Hajnal, P. Gatehouse, D. B. Longmore, S. J. White, A. Oatridge, J. M. Pennock, I. R. Young, and G. M. Bydder, "MR of the brain using fluid-attenuated inversion recovery (FLAIR) pulse sequences.", *American Journal of Neuroradiology*, vol. 13, no. 6, pp. 1555–1564, Nov. 1, 1992, [Online]. Available: http://www.ajnr.org/content/13/6/1555.

[23] L. B. Daniels, M. Shaya, M. L. Nordberg, C. D. Shorter, M. Fowler, and A. Nanda, "Glioblastoma multiforme in two non-nuclear family members", *The Journal of the Louisiana State Medical Society*, vol. 159, no. 4, pp. 215–222, Jul. 1, 2007.

[24] M. Diehn, C. Nardini, D. Wang, S. McGovern, M. Jayaraman, Y. Liang, K. Aldape, S. Cha, and M. Kuo, "Identification of noninvasive imaging surrogates for brain tumor gene-expression modules", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 13, pp. 5213–5218, 2008, DOI: 10.1073/pnas.0801279105.

[25] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent", *Journal of Statistical Software*, vol. 33, pp. 1–22, Feb. 2, 2010, DOI: 10.18637/jss.v033.i01.

[26] J. Fu, K. Singhrao, X. Zhong, Y. Gao, S. Qi, Y. Yang, D. Ruan, and J. H. Lewis, "An automatic deep learning-based workflow for glioblastoma survival prediction using pre-operative multimodal MR images", *arXiv:2001.11155 [physics]*, Jan. 29, 2020, arXiv: 2001.11155, [Online]. Available: http://arxiv.org/abs/2001.11155.

[27] M. Gao, S. Huang, X. Pan, X. Liao, R. Yang, and J. Liu, "Machine Learning-Based Radiomics Predicting Tumor Grades and Expression of Multiple Pathologic Biomarkers in Gliomas", *Frontiers in Oncology*, vol. 10, Sep. 11, 2020, DOI: 10.3389/fonc.2020.01676.

[28] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data", *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 1, 2016, Publisher: Radiological Society of North America, DOI: 10.1148/radiol.2015151169.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, [Online]. Available: https://www.deeplearning.org.

[30] J. J. M. v. Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype", *Cancer Research*, vol. 77, no. 21, e104–e107, Nov. 1, 2017, Publisher: American Association for Cancer Research Section: Focus on Computer Resources, DOI: 10.1158/0008-5472.CAN-17-0339.

[31] E. Grips, N. Wentzensen, C. Sutter, O. Sedlaczek, J. Gebert, R. Weigel, A. Schwartz, M. von Knebel-Doeberitz, and M. Hennerici, "Glioblastoma multiforme als Manifestation des Turcot-Syndroms", *Der Nervenarzt*, vol. 73, no. 2, pp. 177–182, Feb. 1, 2002, DOI: 10.1007/s00115-001-1233-8.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, ISSN: 1063-6919, vol. 2016-December, 2016, pp. 770–778, DOI: 10.1109/CVPR.2016.90.

[33] C. Helissey, C. Vicier, and S. Champiat, "The development of immunotherapy in older adults: New treatments, new toxicities?", *Journal of Geriatric Oncology*, vol. 7, no. 5, pp. 325–333, Sep. 2016, DOI: 10.1016/j.jgo.2016.05.007.

[34] M. P. Houben, D. C. van, J. W. Coebergh, and C. C. Tijssen, "Gliomas: the role of environmental risk factors and genetic predisposition", *Nederlands tijdschrift voor geneeskunde*, vol. 149, no. 41, pp. 2268–2272, Oct. 1, 2005.

[35] C.-T. Hsu and J.-L. Wu, "Hidden digital watermarks in images", *IEEE Transactions on Image Processing*, vol. 8, no. 1, pp. 58–68, Jan. 1999, Conference Name: IEEE Transactions on Image Processing, DOI: 10.1109/83.736686.

[36] A. B. Jamjoom, Z. A. B. Jamjoom, N.-U. Rahman, and R. A. C. Al, "Cervical Lymph Node Metastasis from A Glioblastoma Multiforme in A Child: Report of A Case and A Review of The Literature", *Annals of Saudi Medicine*, vol. 17, no. 3, pp. 340–343, May 1, 1997, Publisher: King Faisal Specialist Hospital & Research Centre, DOI: 10.5144/0256-4947.1997.340.

[37] S. H. Jeon, Y. J. Lim, J. Koh, W. I. Chang, S. Kim, K. Kim, and E. K. Chie, "A radiomic signature model to predict the chemoradiation-induced alteration in tumor-infiltrating CD8+ cells in locally advanced rectal cancer", *Radiotherapy and Oncology*, vol. 162, pp. 124–131, Sep. 1, 2021, DOI: 10.1016/j.radonc.2021.07.004.

[38] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science*, Jul. 17, 2015, Publisher: American Association for the Advancement of Science, DOI: 10.1126/science.aaa8415.

[39] W. H. Jung, S. Choi, K. K. Oh, and J. G. Chi, "Congenital Glioblastoma Multiforme: report of an autopsy case", *Journal of Korean Medical Science*, vol. 5, no. 4, pp. 225–231, May 26, 2009, Publisher: The Korean Academy of Medical Sciences, DOI: 10.3346/jkms.1990.5.4.225.

[40] J. Juntu, J. Sijbers, D. Van Dyck, and J. Gielen, "Bias field correction for MRI images", in *Computer Recognition Systems*, M. Kurzyński, E. Puchała, M. Woźniak, and A. żołnierek, Eds., ser. Advances in Soft Computing, Berlin, Heidelberg: Springer, 2005, pp. 543–551, DOI: 10.1007/3-540-32390-2_64.

[41] G. C. Kabat, A. M. Etgen, and T. E. Rohan, "Do steroid hormones play a role in the etiology of glioma?", *Cancer Epidemiology and Prevention Biomarkers*, vol. 19, no. 10, pp. 2421–2427, Oct. 1, 2010, Publisher: American Association for Cancer Research Section: Minireview, DOI: 10.1158/1055-9965.EPI-10-0658.

[42] D. Kang, J. E. Park, Y.-H. Kim, J. H. Kim, J. Y. Oh, J. Kim, Y. Kim, S. T. Kim, and H. S. Kim, "Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: Development and multicenter external validation", *Neuro-Oncology*, vol. 20, no. 9, pp. 1251–1261, Aug. 2, 2018, DOI: 10.1093/neuonc/noy021.

[43] S. Karcher, H.-H. Steiner, R. Ahmadi, S. Zoubaa, G. Vasvari, H. Bauer, A. Unterberg, and C. Herold-Mende, "Different angiogenic phenotypes in primary and secondary glioblastomas", *International Journal of Cancer*, vol. 118, no. 9, pp. 2182–2189, 2006, DOI: 10.1002/ijc.21648.

[44] C. D. Katsetos, E. Dráberová, A. Legido, C. Dumontet, and P. Dráber, "Tubulin targets in the pathobiology and therapy of glioblastoma multiforme. i. class III β-tubulin", *Journal of Cellular Physiology*, vol. 221, no. 3, pp. 505–513, 2009, DOI: 10.1002/jcp.21870.

[45] P. Kickingereder, S. Burth, A. Wick, M. Götz, O. Eidel, H.-P. Schlemmer, K. H. Maier-Hein, W. Wick, M. Bendszus, A. Radbruch, and D. Bonekamp, "Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models", *Radiology*, vol. 280, no. 3, pp. 880–889, Sep. 1, 2016, Publisher: Radiological Society of North America, DOI: 10.1148/radiol.2016160845.

[46] Y. Kim, H.-h. Cho, S. T. Kim, H. Park, D. Nam, and D.-S. Kong, "Radiomics features to distinguish glioblastoma from primary central nervous system lymphoma on multi-parametric MRI", *Neuroradiology*, vol. 60, no. 12, pp. 1297–1305, Dec. 1, 2018, DOI: 10.1007/s00234-018-2091-4.

[47] P. Kleihues and H. Ohgaki, "Primary and secondary glioblastomas: From concept to clinical diagnosis", *Neuro-Oncology*, vol. 1, no. 1, pp. 44–51, Jan. 1, 1999, DOI: 10.1093/neuonc/1.1.44.

[48] P. Kleihues, F. Soylemezoglu, B. Schäuble, B. W. Scheithauer, and P. C. Burger, "Histopathology, classification, and grading of gliomas", *Glia*, vol. 15, no. 3, pp. 211–221, 1995, DOI: 10.1002/glia.440150303.

[49] P. Klener Jr., P. Otáhal, L. Lateckova, and P. Klener, "Immunotherapy approaches in cancer treatment", *Current Pharmaceutical Biotechnology*, vol. 16, no. 9, pp. 771–781, 2015, DOI: 10.2174/1389201016666150619114554.

[50] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, ser. IJCAI'95, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Aug. 20, 1995, pp. 1137–1143, DOI: 10.1007/3-540-59286-5_57.

[51] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, DOI: 10.1145/3065386.

[52] Y. J. Ku, H. H. Kim, J. H. Cha, H. J. Shin, E. Y. Chae, W. J. Choi, H. J. Lee, and G. Gong, "Predicting the level of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: Usefulness of breast MRI computer-aided detection and diagnosis", *Journal of Magnetic Resonance Imaging*, vol. 47, no. 3, pp. 760–766, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25802, DOI: 10.1002/jmri.25802.

[53] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. W. L. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies, "Radiomics: The process and the challenges", *Magnetic Resonance Imaging*, Quantitative Imaging in Cancer, vol. 30, no. 9, pp. 1234–1248, Nov. 1, 2012, DOI: 10.1016/j.mri.2012.06.010.

[54] S. E. Lakhan and L. Harle, "Difficult diagnosis of brainstem glioblastoma multiforme in a woman: A case report and review of the literature", *Journal of Medical Case Reports*, vol. 3, no. 1, p. 87, Oct. 30, 2009, DOI: 10.1186/1752-1947-3-87.

[55] J. Lao, Y. Chen, Z.-C. Li, Q. Li, J. Zhang, J. Liu, and G. Zhai, "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme", *Scientific Reports*, vol. 7, no. 1, p. 10 353, Sep. 4, 2017, Number: 1 Publisher: Nature Publishing Group, DOI: 10.1038/s41598-017-10649-8.

[56] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7553 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computer science;Mathematics and computing Subject_term_id: computer-science;mathematics-and-computing, DOI: 10.1038/nature14539.

[57] Q. Li, J. Yu, H. Zhang, Y. Meng, Y. F. Liu, H. Jiang, M. Zhu, N. Li, J. Zhou, F. Liu, X. Fang, J. Li, X. Feng, J. Lu, C. Shao, and Y. Bian, "Prediction of tumor-infiltrating CD20+ b-cells in patients with pancreatic ductal adenocarcinoma using a multilayer perceptron network classifier based on non-contrast MRI", *Academic Radiology*, Dec. 16, 2021, DOI: 10.1016/j.acra.2021.11.013.

[58] E. J. Limkin, R. Sun, L. Dercle, E. I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, E. Deutsch, and C. Ferté, "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology", *Annals of Oncology*, vol. 28, no. 6, pp. 1191–1206, Jun. 1, 2017, DOI: 10.1093/annonc/mdx034.

[59] M. Lun, E. Lok, S. Gautam, E. Wu, and E. T. Wong, "The natural history of extracranial metastasis from glioblastoma multiforme", *Journal of Neuro-Oncology*, vol. 105, no. 2, pp. 261–273, Nov. 1, 2011, DOI: 10.1007/s11060-011-0575-8.

[60] X. Ma and F. Jia, "Brain tumor classification with multimodal MR and pathology images", in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 343–352, DOI: 10.1007/978-3-030-46643-5_34.

[61] E. Maher, F. Furnari, R. Bachoo, D. Rowitch, D. Louis, W. Cavenee, and R. DePinho, "Malignant glioma: Genetics and biology of a grave matter", *Genes and Development*, vol. 15, no. 11, pp. 1311–1333, 2001, DOI: 10.1101/gad.891601.

[62] M. J. Mair, B. Kiesel, K. Feldmann, G. Widhalm, K. Dieckmann, A. Wöhrer, L. Müllauer, M. Preusser, and A. S. Berghoff, "LAG-3 expression in the inflammatory microenvironment of glioma", *Journal of Neuro-Oncology*, vol. 152, no. 3, pp. 533–539, May 2021, DOI: 10.1007/s11060-021-03721-x.

[63] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943, DOI: 10.1007/BF02478259.

[64] R. Medikonda, G. Dunn, M. Rahman, P. Fecci, and M. Lim, "A review of glioblastoma immunotherapy", *Journal of Neuro-Oncology*, vol. 151, no. 1, pp. 41–53, Jan. 1, 2021, DOI: 10.1007/s11060-020-03448-1.

[65] M. P. Moore, S. B. Rodney, L. H. Michael, and P. R. Gavin, "Intracranial tumors", *Veterinary Clinics of North America: Small Animal Practice*, vol. 26, no. 4, pp. 759–777, Jul. 1, 1996, DOI: 10.1016/S0195-5616(96)50104-X.

[66] R. K. Moorthy and V. Rajshekhar, "Development of glioblastoma multiforme following traumatic cerebral contusion: Case report and review of literature", *Surgical Neurology*, vol. 61, no. 2, pp. 180–184, Feb. 1, 2004, DOI: 10.1016/S0090-3019(03)00423-3.

[67] A. Mujic, A. Hunn, A. B. Taylor, and R. M. Lowenthal, "Extracranial metastases of a glioblastoma multiforme to the pleura, small bowel and pancreas", *Journal of Clinical Neuroscience*, vol. 13, no. 6, pp. 677–681, Jul. 1, 2006, DOI: 10.1016/j.jocn.2005.08.016.

[68] W. H. Organization, "Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019", Geneva, 2020.

[69] C. Padmalatha, R. C. Harruff, D. Ganick, and G. B. Hafez, "Glioblastoma multiforme with tuberous sclerosis. Report of a case", *Archives of pathology & laboratory medicine*, vol. 104, no. 12, pp. 649–650, Dec. 1, 1980.

[70] D. Pardoll, "Cancer and Immune System: Basic Concepts and Targets for Intervention", *Seminars in Oncology*, vol. 42, no. 4, pp. 523–538, 2015, DOI: 10.1053/j.seminoncol.2015.05.003.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[72] S. Priya, Y. Liu, C. Ward, N. H. Le, N. Soni, R. Pillenahalli Maheshwarappa, V. Monga, H. Zhang, M. Sonka, and G. Bathla, "Machine learning based differentiation of glioblastoma from brain metastasis using MRI derived radiomics", *Scientific Reports*, vol. 11, no. 1, p. 10 478, May 18, 2021, Number: 1 Publisher: Nature Publishing Group, DOI: 10.1038/s41598-021-90032-w.

[73] PyRadiomics Community. (). "PyRadiomics Features", Radiomic Features, [Online]. Available: https://pyradiomics.readthedocs.io/en/v3.0.1/features.html.

[74] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning", *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106 236, Sep. 1, 2021, DOI: 10.1016/j.cmpb.2021.106236.

[75] A. S. Rathore, S. Kumar, R. Konwar, A. Makker, M. Negi, and M. M. Goel, "CD3+, CD4+ & CD8+ tumour infiltrating lymphocytes (TILs) are predictors of favourable survival outcome in infiltrating ductal carcinoma of breast", *The Indian Journal of Medical Research*, vol. 140, no. 3, pp. 361–369, Sep. 2014, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248382/.

[76] D. A. Reardon, A. A. Brandes, A. Omuro, P. Mulholland, M. Lim, A. Wick, J. Baehring, M. S. Ahluwalia, P. Roth, O. Bähr, S. Phuphanich, J. M. Sepulveda, P. De Souza, S. Sahebjam, M. Carleton, K. Tatsuoka, C. Taitt, R. Zwirtes, J. Sampson, and M. Weller, "Effect of Nivolumab vs Bevacizumab in Patients With Recurrent Glioblastoma: The CheckMate 143 Phase 3 Randomized Clinical Trial", *JAMA Oncology*, vol. 6, no. 7, pp. 1003–1010, Jul. 1, 2020, DOI: 10.1001/jamaoncol.2020.1024.

[77] M. Robert and M. Wastie, "Glioblastoma multiforme: a rare manifestation of extensive liver and bone metastases", *Biomedical Imaging and Intervention Journal*, vol. 4, no. 1, e3, Jan. 1, 2008, DOI: 10.2349/biij.4.1.e3.

[78] R. Robinson. (). "Machine Learning Notebook", Machine Learning Notebook, [Online]. Available: https://mlnotebook.github.io/.

[79] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, DOI: 10.1037/h0042519.

[80] A. M. Sanli, E. Turkoglu, H. Dolgun, and Z. Sekerci, "Unusual manifestations of primary Glioblastoma Multiforme: A report of three cases", *Surgical Neurology International*, vol. 1, p. 87, Dec. 22, 2010, DOI: 10.4103/2152-7806.74146.

[81] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation", *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018, DOI: 10.1213/ANE.0000000000002864.

[82] S. Schultz, G. S. Pinsky, N. C. Wu, M. C. Chamberlain, A. S. Rodrigo, and S. E. Martin, "Fine needle aspiration diagnosis of extracranial glioblastoma multiforme: Case report and review of the literature", *CytoJournal*, vol. 2, p. 19, Nov. 14, 2005, DOI: 10.1186/1742-6413-2-19.

[83] J. A. Schwartzbaum, J. L. Fisher, K. D. Aldape, and M. Wrensch, "Epidemiology and molecular pathology of glioma", *Nature Clinical Practice Neurology*, vol. 2, no. 9, pp. 494–503, Sep. 2006, Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Reviews Publisher: Nature Publishing Group, DOI: 10.1038/ncpneuro0289.

[84] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization", *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 1, 2020, DOI: 10.1007/s11263-019-01228-7.

[85] Z. A. Shboul, M. Alam, L. Vidyaratne, L. Pei, M. I. Elbakary, and K. M. Iftekharuddin, "Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction", *Frontiers in Neuroscience*, vol. 13, 2019, Publisher: Frontiers, DOI: 10.3389/fnins.2019.00966.

[86] D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis", *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017, DOI: 10.1146/annurev-bioeng-071516-044442.

[87] J. R Simpson, J Horton, C Scott, W. J Curran, P Rubin, J Fischbach, S Isaacson, M Rotman, S. O Asbell, J. S Nelson, A. S Weinstein, and D. F Nelson, "Influence of location and extent of surgical resection on survival of patients with glioblastoma multiforme: Results of three consecutive radiation therapy oncology group (RTOG) clinical trials", *International Journal of Radiation Oncology*Biology*Physics*, vol. 26, no. 2, pp. 239–244, May 20, 1993, DOI: 10.1016/0360-3016(93)90203-8.

[88] P. van der Smagt and B. Krose, "An introduction to Networks Neural", The University of Amsterdam, Sep. 14, 1995.

[89] R. Sánchez-Ortiga, E. Boix Carreño, O. Moreno-Pérez, and A. Picó Alfonso, "Glioblastoma multiforme and multiple endocrine neoplasic type 2 A", *Medicina Clinica*, vol. 133, no. 5, pp. 196–197, Jul. 4, 2009, DOI: 10.1016/j.medcli.2008.06.021.

[90] A. A. Thomas, M. S. Ernstoff, and C. E. Fadul, "Immunotherapy for the Treatment of Glioblastoma", *Cancer Journal (Sudbury, Mass.)*, vol. 18, no. 1, pp. 59–68, Jan. 2012, DOI: 10.1097/PPO.0b013e3182431a73.

[91] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[92] S. Tibshirani, H. Friedman, and T. Hastie, *The Elements of Statistical Learning*, 2nd ed. Springer, 2001, [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-21606-5.

[93] C.-L. Tso, W. A. Freije, A. Day, Z. Chen, B. Merriman, A. Perlina, Y. Lee, E. Q. Dia, K. Yoshimoto, P. S. Mischel, L. M. Liau, T. F. Cloughesy, and S. F. Nelson, "Distinct transcription profiles of primary and secondary glioblastoma subgroups", *Cancer Research*, vol. 66, no. 1, pp. 159–167, Jan. 1, 2006, Publisher: American Association for Cancer Research Section: Cell, Tumor and Stem Cell Biology, DOI: 10.1158/0008-5472.CAN-05-0077.

[94] B. Tysnes and R. Mahesparan, "Biological mechanisms of glioma invasion and potential therapeutic targets", *Journal of Neuro-Oncology*, vol. 53, no. 2, pp. 129–147, 2001, DOI: 10.1023/A:1012249216117.

[95] H. Um, F. Tixier, D. Bermudez, J. O. Deasy, R. J. Young, and H. Veeraraghavan, "Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets", *Physics in Medicine & Biology*, vol. 64, no. 16, p. 165 011, Aug. 2019, Publisher: IOP Publishing, DOI: 10.1088/1361-6560/ab2f44.

[96] K. Urbańska, J. Sokołowska, M. Szmidt, and P. Sysa, "Glioblastoma multiforme – an overview", *Contemporary Oncology*, vol. 18, no. 5, pp. 307–312, 2014, DOI: 10.5114/wo.2014.40559.

[97] C. L. Ventola, "Cancer Immunotherapy, Part 1: Current Strategies and Agents", *Pharmacy and Therapeutics*, vol. 42, no. 6, pp. 375–383, Jun. 2017, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440098/.

[98] C. J. Wallace, P. A. Forsyth, and D. R. Edwards, "Lymph node metastases from glioblastoma multiforme.", *American Journal of Neuroradiology*, vol. 17, no. 10, pp. 1929–1931, Nov. 1, 1996, Publisher: American Journal of Neuroradiology, [Online]. Available: http://www.ajnr.org/content/17/10/1929.

[99] M. L. Waskom, "Seaborn: Statistical data visualization", *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 6, 2021, DOI: 10.21105/joss.03021.

[100] R. Weinberg, "How cancer arises.", *Scientific American*, vol. 275, no. 3, pp. 62–70, 1996, DOI: 10.1038/scientificamerican0996-62.

[101] L. M. Weiner, "Cancer immunology for the clinician", *Clinical Advances in Hematology & Oncology: H&O*, vol. 13, no. 5, pp. 299–306, May 2015.

[102] M. Weller, N. Butowski, D. D. Tran, L. D. Recht, M. Lim, H. Hirte, L. Ashby, L. Mechtler, S. A. Goldlust, F. Iwamoto, J. Drappatz, D. M. O'Rourke, M. Wong, M. G. Hamilton, G. Finocchiaro, J. Perry, W. Wick, J. Green, Y. He, C. D. Turner, M. J. Yellin, T. Keler, T. A. Davis, R. Stupp, J. H. Sampson, N. Butowski, J. Campian, L. Recht, M. Lim, L. Ashby, J. Drappatz, H. Hirte, F. Iwamoto, L. Mechtler, S. Goldlust, K. Becker, G. Barnett, G. Nicholas, A. Desjardins, T. Benkers, N. Wagle, M. Groves, S. Kesari, Z. Horvath, R. Merrell, R. Curry, J. O'Rourke, D. Schuster, M. Wong, M. Mrugala, R. Jensen, J. Trusheim, G. Lesser, K. Belanger, A. Sloan, B. Purow, K. Fink, J. Raizer, M. Schulder, S. Nair, S. Peak, J. Perry, A. Brandes, M. Weller, N. Mohile, J. Landolfi, J. Olson, G. Finocchiaro, R. Jennens, P. DeSouza, B. Robinson, M. Crittenden, K. Shih, A. Flowers, S. Ong, J. Connelly, C. Hadjipanayis, P. Giglio, F. Mott, D. Mathieu, N. Lessard, S. J. Sepulveda, J. Lövey, H. Wheeler, P.-L. Inglis, C. Hardie, D. Bota, M. Lesniak, J. Portnow, B. Frankel, L. Junck, R. Thompson, L. Berk, J. McGhie, D. Macdonald, F. Saran, R. Soffietti, D. Blumenthal, S. B. C. M. André de, A. Nowak, N. Singhal, A. Hottinger, A. Schmid, G. Srkalovic, D. Baskin, C. Fadul, L. Nabors, R. LaRocca, J. Villano, N. Paleologos, P. Kavan, M. Pitz, B. Thiessen, A. Idbaih, J. S. Frenel, J. Domont, O. Grauer, P. Hau, C. Marosi, J. Sroubek, E. Hovey, P. S. Sridhar, L. Cher, E. Dunbar, T. Coyle, J. Raymond, K. Barton, M. Guarino, S. Raval, B. Stea, J. Dietrich, K. Hopkins, S. Erridge, J.-P. Steinbach, L. E. Pineda, Q. C. Balana, B. B. Sonia del, M. Wenczl, K. Molnár, K. Hideghéty, A. Lossos, L. Myra van, A. Levy, R. Harrup, W. Patterson, Z. Lwin, S. Sathornsumetee, E.-J. Lee, J.-T. Ho, S. Emmons, J. P. Duic, S. Shao, H. Ashamalla, M. Weaver, J. Lutzky, N. Avgeropoulos, W. Hanna, M. Nadipuram, G. Cecchi, R. O'Donnell, S. Pannullo, J. Carney, M. Hamilton, M. MacNeil, R. Beaney, M. Fabbro, O. Schnell, R. Fietkau, G. Stockhammer, B. Malinova, K. Odrazka, M. Sames, G. Miguel Gil, E. Razis, K. Lavrenkov, G. Castro, F. Ramirez, C. Baldotto, F. Viola, S. Malheiros, J. Lickliter, S. Gauden, A. Dechaphunkul, I. Thaipisuttikul, Z. Thotathil, H.-I. Ma, W.-Y. Cheng, C.-H. Chang, F. Salas, P.-Y. Dietrich, C. Mamot, L. Nayak, and S. Nag, "Rindopepimut with temozolomide for patients with newly diagnosed, EGFRvIII-expressing glioblastoma (ACT IV): A randomised, double-blind, international phase 3 trial", *The Lancet Oncology*, vol. 18, no. 10, pp. 1373–1385, Oct. 1, 2017, DOI: 10.1016/S1470-2045(17)30517-X.

[103] A. Widjaja, H. Mix, C. Gölkel, P. Flemming, R. Egensperger, A. Holstein, J. Rademaker, H. Becker, M. Hundt, S. Wagner, and M. P. Manns, "Uncommon Metastasis of a Glioblastoma Multiforme in Liver and Spleen", *Digestion*, vol. 61, no. 3, pp. 219–222, 2000, Publisher: Karger Publishers, DOI: 10.1159/000007761.

[104] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", *Climate Research*, vol. 30, no. 1, pp. 79–82, Dec. 19, 2005, DOI: 10.3354/cr030079.

[105] K. K. Wong, R. Rostomily, and S. T. C. Wong, "Prognostic gene discovery in glioblastoma patients using deep learning", *Cancers*, vol. 11, no. 1, p. 53, Jan. 2019, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, DOI: 10.3390/cancers11010053.

[106] J. Wu, A. T. Mayer, and R. Li, "Integrated imaging and molecular analysis to decipher tumor microenvironment in the era of immunotherapy", *Seminars in Cancer Biology*, Dec. 5, 2020, DOI: 10.1016/j.semcancer.2020.12.005.

[107] L. Zhen, C. Yufeng, S. Zhenyu, and X. Lei, "Multiple extracranial metastases from secondary glioblastoma multiforme: A case report and review of the literature", *Journal of Neuro-Oncology*, vol. 97, no. 3, pp. 451–457, May 1, 2010, DOI: 10.1007/s11060-009-0044-9.

[108] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel, R. Gillies, O. Gevaert, and R. Gatenby, "Radiomics in brain tumor: Image assessment, quantitative feature descriptors, and machine-learning approaches", *American Journal of Neuroradiology*, vol. 39, no. 2, pp. 208–216, Feb. 1, 2018, Publisher: American Journal of Neuroradiology Section: Adult Brain, DOI: 10.3174/ajnr.A5391.

[109] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, DOI: 10.1111/j.1467-9868.2005.00503.x.

[110] F. Çelebi, F. Agacayak, A. Ozturk, S. Ilgun, M. Ucuncu, Z. E. Iyigun, Ordu, K. N. Pilanci, G. Alco, S. Gultekin, E. Cindil, G. Soybir, F. Aktepe, and V. Özmen, "Usefulness of imaging findings in predicting tumor-infiltrating lymphocytes in patients with breast cancer", *European Radiology*, vol. 30, no. 4, pp. 2049–2057, Apr. 1, 2020, DOI: 10.1007/s00330-019-06516-x.