UNIVERSITY OF BAMBERG

Computing in the Humanities
Faculty of Information Systems and Applied Computer Sciences

Master's Thesis

XAI in Medicine: Analysis and Evaluation of XAI Tools and Legal Liability for Neural Networks. A Case Study on Tumor Image Classification

BY

Alina Tenne

Supervisor:
Prof. Dr. Christoph Benzmüller and Prof. Dr. Andrea Vestrucci

Bamberg 2024

Chair for AI Systems Engineering

March 2024

URN: urn:nbn:de:bvb:473-irb-966721
DOI: https://doi.org/10.20378/irb-96672

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) have immense potential to revolutionize various fields, especially in the domain of medicine. Deep learning models are increasingly used in the healthcare sector for image classification and disease diagnosis. However, the problem of explainability remains a major concern. It is still unclear how much explanatory power current XAI tools have.

Therefore, this analysis aims to evaluate explainable AI (XAI) tools based on the current Ethics Guidelines of Trustworthy AI from the European Commission. The purpose is to determine the extent to which current explanation algorithms provide trustworthy, transparent, and explanatory support for black box models. The issue of liability arises due to the lack of traceability and increased use of black box models. It is unclear which organization, individuals, or groups of people are liable in the event of a claim. This thesis analyzes the current draft legislation on the AI Act and the Legal Liability Directive with regard to the question of liability. Additionally, it examines the role of XAI tools in this context.

XAI tools currently provide extensive capabilities for visualising model decisions and explaining the factors that are most likely to have contributed to the outcome in an understandable manner. However, the analysis and evaluation of XAI tools revealed that there are some opportunities for improvement. Additionally, the availability of XAI tools heavily influences the issue of liability, as traceability and transparency are crucial elements for the legal implementation of new technologies.

Contents

List of Figures

List of Tables

1 Introduction

The swift advancement of artificial intelligence (AI) is significantly affecting various aspects of our society, especially the healthcare sector. AI technologies hold the potential to transform medical practice, ranging from diagnosing and treating diseases to enhancing patient care [GSM21] [van+22] [Asi+23]. Brain tumors pose a significant challenge in medicine due to the difficulty in achieving early identification and timely treatment, which are crucial for survival [BFM23]. The incidence of malignant brain tumors is increasing globally, with 3.5 out of 100,000 people diagnosed. In 2020, 308,102 people worldwide were diagnosed with malignant brain tumors, a 38% increase from 2002 when 189,582 people were diagnosed [Hir24]. In addition to the high incidence of illnesses, cancer is one of the the leading cause of premature death due to illness.

> According to the WHO and the American Cancer Society (ACS), cancer is the second leading cause of death despite tremendous advances in tumor identification and therapy in recent years [Alr+23].

Given the rapid advancements in modern technology and the increasing challenges in medicine, the integration of AI in disease diagnosis is of utmost importance. AI models can enhance diagnostic accuracy by identifying patterns in medical images that may not be discernible to the human eye. Additionally, AI-based support can lead to significant time savings. Computer-aided diagnosis enables specialists to perform treatment measures remotely, thereby promoting greater workflow efficiency. In the case of brain tumor detection, AI can visualize the relationship between the brain tumor and the surrounding brain structures. This can be helpful in treatment planning [Zei+22] [Pei24].

Despite the impressive advances in AI, the opaque nature of many AI models remains a challenge. This has led to the development of the concept of Explainable Artificial Intelligence (XAI). XAI aims to make AI decisions transparent and understandable, which is particularly important when AI is used in sensitive areas such as medicine [van+22]. The following example is based on a study of AI-driven classifiers by Bach et al. [Bac+15] and illustrates a typical situation that arises in particular due to the black box phenomenon: a neural network was developed for recognizing horse images. As the accuracy was immediately very high, to the surprise of the developers, they wanted to determine the cause of this in order to extend this property to other artificial networks. After the complex arithmetic operations and model relationships could not be deciphered, it was later discovered that the trained network had cheated. Many images in the data set contained watermarks as copyright. Accordingly, the neural network had not learned the properties and characteristics of horses but had cheated by exploiting the watermark. In a targeted manner, the AI leaped only as high as necessary, specifically searching for watermarks rather than similarities with horses [Bac+15] [Lap+19]. In summary, the black box principle refers to the fact that the conditions and results of certain AI models are measurable, but the intermediate processes and decision-making remain hidden. American experts were able to demonstrate another example of the so-called Clever Hans effect by training a logical regression classifier that distinguishes wolves from huskies. The network was provoked by including only wolf images with snow in the background. The network was shown to respond to snow as a classification indicator and consequently identified all images with light backgrounds as wolves [RSG16]. Clever Hans predictors typically achieve very good test results, but as soon as the models encounter new, unknown data, difficulties with the classifications increase. In the real world, where both huskies and wolves live in regions with snow or horses occur in a natural environment without watermarks, these systems would fail completely [SM19] [Lap+19]. Explainable AI illuminates the inner workings of the box, thus it helps to identify and understand weaknesses in a system by revealing biases. Incorrect behavior of classifiers (e.g. the incorrect classification of a husky due to a snow background) can be comprehended through explanations and addressed accordingly. Even the application to a single test image can be sufficient to identify the actual origin for the particular prediction [SM19]. Systems that are backed up with explanations not only offer a head start in terms of transparency and traceability, but can also be improved more easily as it is possible to analyze errors. A good understanding of system behavior provides an overview of the overall processes and is crucial for swift corrections and debugging [SM19].

The terms explainability, transparency and trust are closely linked in relation to AI systems and are mutually dependent in order to promote the understanding and acceptance of AI [Gri23] [Cha+23]. The aim of the interplay is to understand AI systems and agree with their actions in order to enable their use in various areas [Cha+23]. The term explainability refers to explaining the actions and decisions of AI systems to people in an understandable way and comprehending the logic and reasoning behind them. Transparency is tightly intertwined to this, can be achieved through visualizations and helps to increase confidence in the decisions made by an artificial classifier. Transparency ultimately expresses how understandable and comprehensible the internal processes and decisions of AI-based software are [Cha+23]. Explainability and transparency form the basis for trust in AI systems, as this is what broad acceptance and successful implementation are built on. Only with a comprehensible functioning of a system can it be deemed as integrity-preserving and reliable, thus classified as trustworthy [Gri23] [BTR23] [Cha+23]. Trust is a crucial factor that affects people's reliance on certain technologies. People are more likely to accept predictions or decisions made by AI/ML algorithms when backed by logical explanations and scientific information. In other words, if a system is classified as trustworthy, then it is sufficiently explainable through a transparent design [GSM21]. The trustworthiness of a system can therefore be used as an evaluation criterion for assessing explainability. XAI and healthcare form an effective combination of digital technologies. Due to the nature of AI-based diagnostic systems, trusting AI-driven conclusions is a matter of serious concern [GSM21].

> Trust is an important factor for the perseverance of AI in the medical and healthcare segment of the digital industry [GSM21].

It is therefore undisputed that, particularly in the medical field, there is a great need for trustworthy systems that successfully support experts and make reliable and comprehensible decisions. It should be noted that there are logic-based diagnostic models that are inherently comprehensible as the decisions are based on clearly delineated and predefined rules. Whether a particular person has asthma or not is accordingly identified based on a patient's medical history or by employing a model that traverses a checklist of symptoms and assesses the extent to which specific criteria are met [GSM21]. The rigid adherence to explicitly defined rules makes decisions comprehensible and explainable for people. In contrast to this are the black box models in which an "unboxing mechanism" is required to understand and subsequently assess the behavior [Has+22]. In the field of medical diagnosis, black box ML models have been shown to outperform rule-based models in terms of accuracy and efficiency, as they are capable of recognizing features and complex characteristics that are often invisible to the human eye and would therefore be neglected [Luc+22]. Logic-based systems are often limited by the complexity of the rules required to capture the full range of factors that influence a disease. Therefore, black-box models start where rule-based models stop: they capture complex interactions between variables that do not have to be explicitly predefined using rules [GSM21] [Luc+22] [Has+24]. According to several researches, deep learning-based models have shown great promise in a variety of applications, including the diagnosis of lung cancer [SK22], skin cancer [Est+17] or diabetic retinopathy [Pai+19]. Although these are criticized for their lack of explainability, another important advantage is that they can learn from large data sets without prior knowledge. Rule-based approaches, on the other hand, need to be fed with domain-specific knowledge. It is therefore a major concern to promote and further develop XAI tools in order to achieve technical progress [Luc+22] [Bor+23]. These tools include generating visualizations of the model's decision-making process and identifying the most important factors that contribute to AI-driven predictions. Hence, nowadays researches should focus on developing methods for explaining model behavior [GSM21] [Has+24]. As artificial intelligence (AI), including XAI, has the potential to revolutionise medical diagnosis and treatment of diseases, the question of liability becomes crucial [Ces+23]. Legal requirements and claims may change with increasing transparency and trust in AI systems. Despite the promising advantages, AI and XAI raise important questions regarding legal liability. Who is responsible if an AI-based system makes a mistake? Is it the doctor who used the system, the manufacturer of the system or someone else? When it comes to the use of potential life-saving technologies in the treatment of life-threatening diseases like brain tumors, it is important to research the legal consequences for the parties involved [Com21] [Com22b] [Ces+23].

After establishing the context and background of AI in medicine, the importance of explainability and the transparent and traceable implementation of XAI tools become apparent. These tools can enrich the medical field and enhance disease diagnostics with state-of-the-art technology. As a result, the following research questions emerge:

- To what extent do the current explanation algorithms within XAI tools contribute to the explainability of AI systems, particularly concerning black box models?

The research question is addressed particularly by conducting an analysis and evaluation of the effectiveness of existing explanation algorithms within XAI tools. This aims to identify both current strengths and weaknesses, with a specific focus on fostering explainable AI. Building upon the findings of the analysis and evaluation of existing explanation algorithms within XAI tools, the subsequent research questions delve into the mechanisms and implications of liability in the context of AI failures.

- How is legal liability addressed in the event of misbehavior, miscalculations, or misdiagnoses by an artificial system?
- What role do XAI tools play in resolving these liability issues?

The research aims to evaluate the effectiveness of explanatory algorithms in XAI tools to enhance the transparency of AI systems and strengthen trust in them. This involves analyzing the contribution that XAI tools currently make to the explainability of AI systems and exploring ways to improve it. Moreover, the study will analyze liability in the context of AI errors and their effects. In addition to theoretical considerations, this work includes a Python case study. The empirical case study allows for practical application and testing of the concepts to be analyzed. The case study provides a concrete example that illustrates the process of model learning, as well as the necessity and implementation of explanation algorithms.

Chapter 3 provides a more concrete introduction and presentation of the Python case study. The appendix A includes a link to the associated Github repository where the Python case study can be accessed.

The upcoming sections will examine the methodologies and analytical strategies employed, and guide through the field of machine learning and explainable AI. After providing a comprehensive explanation of the theoretical concepts in the field of machine learning, including deep learning and explainable AI tools, the practical implementation will be illustrated and explained through the Python case study. This approach highlights the multifaceted nature of the thesis and ensures a cohesive journey towards addressing and answering the core inquiries outlined above. The section, **Background and Related Work**, provides the historical perspective and developments in the research field. It also outlines the current state of work in comparison with post-studies. The section, **Methodology**, emphasizes the study's structure and design. The section on **Machine Learning** starts with an introduction to deep learning, followed by an overview of the residual network ResNet50, which is implemented in the

Python case study. The section on ResNet50 compares different model architectures and explains why ResNet50 was chosen for the classification problem. It then examines the model training and concludes with a detailed analysis and performance overview of ResNet50. The subsequent section, **XAI Tools**, presents and explains three explanation algorithms, first in theory and then in practical implementation in the case study. The section **Analysis and Evaluation of XAI Tools** combines an ethical analysis of trustworthy XAI with legal liability under the current state of law within the European Union. The section **Evaluation of Implementation** reviews the implementation of the Python case study with user interviews, providing an overview of the understanding and assessment of an external user group. The **Conclusion** offers a retrospective of the thesis, summarises the results, and provides insights for further steps and limitations.

2 Background and Related Work

The purpose of the background and related work section is to offer a general understanding of the topic and present the outcomes of previous research related to categorising the scope of the thesis. The section does not aim to provide an exhaustive list of all relevant research, but rather an overview. The studies presented were selected according to the similarity criterion and thus provide a general introduction to the subject area of artificial intelligence, explainable AI and deep learning methods in the medical field. Additionally, the text follows a schematic structure, progressing from general to specific. It introduces (X)AI and clarifies the meaning of the term *explainable*, before delving into a detailed discussion of specific model forms and research gaps.

Artificial intelligence is a vast interdisciplinary research field that covers advanced topics such as machine learning, computer vision and autonomous systems. It has gained prominence and is currently experiencing heightened visibility. Numerous studies have been conducted, reflecting a rich body of knowledge in this field. While the existing research provides valuable insights, the rapid evolution and widespread implementation of (X)AI underscore its current significance, necessitating ongoing and extensive research to address emerging challenges and explore new dimensions of this dynamic topic [Bac+15] [Lap+19] [GSM21] [Luc+22] [Bor+23] [S B+23]. Currently, AI is revolutionising the medical field with state-of-the-art applications. However, there is still doubt about the trust in these black box phenomena. It is important to shed light on the dark to exploit the strengths of artificial systems and minimise or eradicate the weaknesses [Bor+23] [S B+23]. Medical healthcare is a vast field where new technologies have a significant impact on transforming the delivery of healthcare services, resulting in an efficient patient care system. Today's medical technologies aim to enhance healthcare services and improve patient outcomes. They are used in various areas, including developing treatment plans, diagnosing illnesses, analysing medical images, creating medication, robot-assisted operations, maintenance of medical devices or resource management in hospitals [Fra24].

Although deriving rules from trained neural networks has long been a central concern in AI research, the term XAI was first coined in 2004 by Van Lent et al. [LFM04] [Ali+23]. Their aim was to explain the actions of AI-controlled units in simulation games and reveal their system-specific capabilities. Since then, there has been a renewed focus on explainability in artificial systems, with various research groups exploring the topic. The issue of explainability in AI decision-making is approached from different perspectives, and explanations are subject to different meanings [Ali+23].

> The word *explainability*, in terms of AI concepts, means functional knowledge of the model, the purpose of which is to attempt to describe the model's black-box behavior. It is often used interchangeably with the term interpretability in the literature. Explainability expresses what is occurring in the model by providing a human-readable explanation of the model's decision [Ali+23].

However, defining what qualifies an explanation as *explainable* is challenging. Generally, an explanation involves describing something and providing a response to the question of why that description is correct [Ali+23]. Arrieta et al. offer a widely accepted definition of explainable AI, which emphasises the interaction between the explainer and the explainee. An ideal explanation should be social and collaboratively communicated to users. This underscores the importance of systems providing clear explanations that are tailored to the specific user group. Explanations for developers differ from those for regulators, which in turn differ from those for end-users [Bar+20].

In recent years, different XAI tools have been developed to address the issue of explainability and leverage the extensive capabilities of artificial intelligence. For instance, XAI techniques like Xception CNN and VGG16 with SHAP(XAI) have been employed to detect Covid-19. By using a VGG16 architecture and visual images of lungs captured through computed tomography (CT) and X-rays (RX), it is possible to diagnose whether a patient is suffering from Covid-19 [Alo21] [S B+23]. The model has achieved a high level of accuracy and can distinguish between Covid-19 infection and non-Covid-19 pneumonia, which is crucial for human health. Explanatory algorithms have been introduced to clarify the decision-making process. The GradientExplainer approach was used to calculate the Shapley Additive Explanations (SHAP) values for this purpose. These values measure the influence of specific relevant features in an image on the prediction. Areas of an image with high SHAP values contain relevant information for the classification model. Overall, this model and its explanatory properties are a valuable enrichment for medical diagnosis using visual image files. Additionally, the relevance and potential of deep learning architectures were emphasised, proving once again that these are future-oriented technologies that must continue to be promoted and investigated [Alo21]. In addition, chronic kidney disease can be detected using K-Means Clustering, DB-Scan, I-Forest, and Autoencoder with SHAP [Ant+21] and the combination of Extreme Gradient Boosting (XGBoost) and SHAP has been used to identify mediastinal cysts and tumors [Wan+21].

Despite the fact that the number of studies on XAI is quickly growing

[Ali+23], artificial systems that also provide a tool to explain model behaviour are not yet the standard in studies. Transparent and comprehensible design, along with the use of XAI tools, should go hand in hand with the establishment of AI-based systems. These tools are considered a fundamental prerequisite for trustworthy interactions in society and form the basis for significant safety precautions [Bor+23]. Despite its undisputed importance, this part of machine learning has been mostly neglected to date. The focus so far has been on the successful training of artificial networks, resulting in models that make optimal predictions. Accordingly, once the model evaluation process was completed, these were then established and the step of providing the opportunity to understand the model's decision was omitted [Tal+19] [Reh+20] [AZ21] [Tew+21] [SK22]. It is worth noting that when XAI tools are employed, detailed analysis is not always conducted to determine the most appropriate explanatory method for a specific use case. Furthermore, XAI tools lack specific assessment criteria for categorising their capabilities and skills [Ali+23]. The aforementioned studies confirm these tendencies and thus emphasize the need to investigate XAI tools to determine their effectiveness in terms of explainability and to identify which tool is appropriate for each use case [Alo21] [Ant+21] [Wan+21] [Bor+23]. While there are existing taxonomies for categorising AI systems as trustworthy or non-trustworthy, these have only been applied to XAI tools to a limited extent. One example is the risk-based framework supported by a suite of AI ethical criteria, such as the IEEE $CertifAIEd^{TM}$ program or the IEEE 7000^{TM}2021 standard program, which addresses ethical concerns during system design. The International Organization for Standardisation (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) are recognised as publishers and have pursued the goal of providing a generally valid classification for artificial systems under ethical standards [New23]. The European Commission has also published both an AI Act and Ethics Guidelines for Trustworthy AI. These criteria represent the latest draft legislation for categorising artificial systems within the European Union [Art19] [Com22b] [Par23]. Due to the newness of the topic, there are not yet many studies on the ethical guidelines by the European Commission, especially regarding their direct application to XAI tools, despite their undisputed importance. Research in the area of explainable technologies and XAI tools is of high relevance due to the current draft legislation of the European Commission. It offers a future-oriented and sustainable progress for the development of artificial systems and is in the public interest [New23] [Com22a].

The cited sources [Tal+19] [Reh+20] [AZ21] [Tew+21] [SK22] provide examples for research and implementation of neural networks that have not integrated explanatory algorithms. The focus has been on developing and applying the neural networks themselves, without incorporating explanatory algorithms into the research process.

Convolutional neural networks (CNN) are commonly used in the medical field, especially for image classification focused on diagnosis. The architecture and parameters of the model play a crucial role in determining its performance and the resulting outcomes [SSA19]. While VGG16 and similar models have shown promising results, the ResNet50 architecture has become the preferred choice for brain tumor detection [SK22] [Sha+23] [Asi+23]. More precisely, a study on TCGA-LGG and TCIA dataset for MRI applications demonstrates that ResNet50 generally outperforms CNN models in accurately detecting and classifying tumors. This architecture significantly improves model performance and achieves the most

The Cancer Genome Atlas - Lower Grade Glioma (TCGA-LGG) and the Cancer Imaging Archive (TCIA) datasets consist of medical images of brain tumors used for research and development of imaging procedures and diagnostic tools in the field of magnetic resonance imaging (MRI). A total of 120 patients are included [Asi+23].

valuable accuracy values in the overall analysis [Asi+23]. The developed ResNet50 model by [Asi+23] is able to accurately identify the presence or absence of malignant brain tumors. Furthermore, Ismael and Abdel-Qader [IA18] and Sultan et al. [SSA19] have developed models that can not only detect the presence of a tumor but also classify the medical differences between three types of brain tumors. Their deep learning architectures can accurately identify whether a patient has a meningioma, glioma, or pituitary tumor. The main advantage of CNN architectures, including residual networks such as ResNet50, lies in the manner in which they extract features of visual data samples. Convolutional filters act as feature extractors, allowing for the extraction of increasingly complex features, such as spatial and structural information, by delving deep into distant modelling layers. This is one of the reasons why they are popular for implementing medical AI-driven applications [SSA19]. Moreover, section 4 presents a detailed comparison of specific model architectures and explains how deep networks work, highlighting their advantages.

3 | Methodology

In order to offer insight into the methodological approach, this chapter provides a guide to the systematic implementation of the research questions, an overview of the work and the concept of the research process.

Early diagnosis of brain tumors is crucial for efficient and effective treatment. Non-automated tumor classification by experts is time-consuming, costly, requires physician capacity that cannot be used elsewhere and is not reproducible [AZ21]. For this reason, artificial neural networks (ANNs) are intended to represent a technical supporter, which on the one hand predicts precise disease diagnoses using computer-aided state-of-the-art technology and on the other hand, similar to a doctor, provides methods of explaining how the classifications are made [Tew+21] [Bor+23]. As a future-oriented step, the aim of this thesis is to focus on the explainability of neural networks for medical image classification. Due to the lack of providing XAI tools when using intelligent systems [Tal+19] [Reh+20] [AZ21] [Tew+21] [SK22], there is initially a need for an analysis of XAI tools in order to evaluate their efficiency and potential limitations to ascertain the current state and provide a prospective outlook for a desired future state.

More specifically, based on a Python case study designed for a medical classification problem, a selection of explanation algorithms is applied to an artificial network. This enables an analysis and subsequent evaluation of explanation methods, as well as a final assessment of the properties of XAI tools, to determine their adequacy in providing explanations for the use of artificial systems. Within the case study, the entire process of model training, model evaluation, and the application of selected explanatory algorithms will be outlined. This will serve as a practical example guiding through the creation of an artificial system aimed at diagnosing various types of brain tumors. Moreover, the methods and processes contained within the case study will be theoretically explained and illustrated within the work.

Furthermore, [Ali+23] states the importance of binding the gap between legal requirements of trustworthy AI and current developments and applications in the relevant technological fields. In other words, there is a need to reconcile current legal forms with future systems in order to establish robust AI-driven software.

This enables the deployment of AI systems that are not only

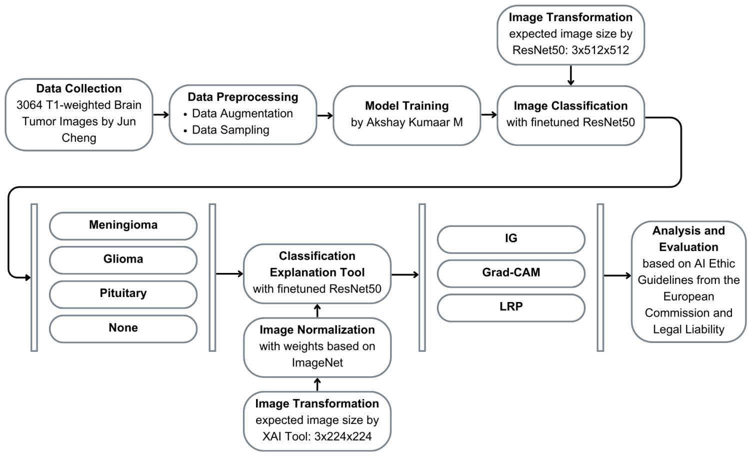


Figure 3.1: Workflow Diagram [1]

technically advanced, but also meet legal and ethical requirements. Addressing this gap facilitates the development of trustworthy AI and ensures that it is used for the benefit of the society [Ali+23].

To achieve this goal, it is necessary, to include legal liability in the analysis and evaluation of XAI tools, to provide XAI systems that respect fundamental values, data protection standards and social norms as well as legal requirements. The Ethics Guidelines of Trustworthy AI are capable of addressing the initial three components, while the final aspect relies on the existing EU law and the AI Act [Art19] [Par23]. While the guidelines focus on ethical principles and values for the classification of trustworthy AI, the AI Act sets out legal provisions and obligations. Combined with the current EU liability framework, they constitute a solid foundation for the analysis and assessment of XAI tools.

Figure 3.1 provides an overview of the components of the foundational framework in order to visualize the sequence and the individual work steps of the case study. The components are described more detailed in the following sections, as this illustration is intended to provide an outline and categorization of the overall process. First, **data collection** must be performed to identify a dataset that fits the specific classification problem. In this case, since the focus is on diagnosis of brain tumors, the data set of [Che17] is used, which provides MRI image data of three different tumor types. Subsequently, in the **data preprocessing** step, preparations are made for successful data processing in model training, namely enlarging the data collection and dividing it into different data samples. The **model training** is based on an already established AI-driven model of [M20], since this work focuses on implementing XAI tools to explain the model decision process and model training from scratch is time-consuming and requires a lot of expert experience [Tal+19] [Tew+21]. The residual network, which is a special architecture of deep neural networks, is used to **classify the input image** into four possible disease diagnoses after

[1]Figures and tables without references were created independently by the author. This applies to all illustrations in this work without reference.

the input data has been adapted using **image transformation**. By implementing three different **explanation algorithms** that are specifically tailored to the neural network and the particular classification problem, the final **analysis and evaluation** can be carried out based on AI Ethic Guidelines and Legal Liability. In particular, the ethical guidelines present evaluation criteria in relation to fundamental values, data protection standards and social norms, and the presentation of liability when using AI-driven software takes into account the legal basis. Before the explanation algorithms can successfully process the images, the input data must be prepared and standardized using **image normalization** and **image transformation**. Together, these individual components constitute the pipeline of the case study. As a result of this process, the question of which XAI tools are suitable for explaining model decisions and what characteristics they must possess to be considered interpretable and to align with legal requirements, will be answered. The experimental implementation is based on the interactive development environment (IDE) Jupyter Notebook with Python as the programming language, since the application gained significant prominence in medical contexts, particularly in the domain of image recognition utilizing convolutional neural networks [KSH18] [Lau+23].

4 Machine Learning

Machine learning (ML) is a fundamental aspect of contemporary AI and plays a crucial role in various applications and industrial sectors [Bar+20] [Jea+23] [Web+23]. ML algorithms automatically recognise patterns and structures in data and learn from them without explicit programming. This is achieved by adapting models to data to make predictions or decisions [Hin+23]. The aim of this chapter is to introduce key areas of machine learning. This will provide theoretical support for the practical content and concepts of the Python case study, and the necessary background knowledge to understand the implementation transparently. The section 4.1 introduces deep learning, describing how deep neural networks work and introducing a specific form of DL networks, namely CNNs. Additionally, the fundamental concept of model training and limitations of DL processes are covered. After discussing the theoretical aspects of deep learning, the section 4.2 is dedictated to explaining the practical implementation of the ResNet50 model in the Python case study. This expands on the previous theoretical framework by demonstrating how the acquired insights can be applied in a tangible context.

4.1 Deep Learning

Deep learning (DL) is a subfield of ML, which in turn is a subset of AI. The emergence of the DL concept cannot be dated to a specific point in time or traced back to a person. However, Rina Dechter and Geoffrey Hinton played crucial roles in coining the term [Dec86] [HSA06] [Sch15] [Sch16] [Hin18] [Tho22]. Furthermore, it can be noted that the concept has experienced a resurgence since the 1990s with the objective of developing intelligent systems capable of abstracting the functionality of the human brain and possessing the ability to formulate their own predictions and decisions [Sch15] [Sch16]. DL methods employ numerous layers of non-linear processing units for the purposes of deep learning, feature extraction, and transformation. In this case, non-linearity enables the model to gain highly complex and abstract representations of data, which is a key aspect for deep learning models. DL architectures are composed of multiple layers arranged sequentially. These layers, categorized as consecutive, utilize the output value of the preceding layer as their input.

Furthermore, this subfield is divided into supervised learning and unsupervised learning models. The former includes well-known architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Conversely, unsupervised learning encompasses self-organising maps (SOM) or autoencoders (AE). As CNNs are most widely used in deep learning networks for classification processes and they have shown great development in recent years, the study will be based on a CNN architecture [YC20]. Deep residual networks as the popular ResNet50, belong to the family of CNNs. The number after the model name refers to the layer depth; the ResNet50 is therefore 50 layers deep [Tal+19] [Was23].

4.1.1 CNNs

Artificial neural networks have become increasingly popular in literature due to their ability to handle large amounts of data. Specifically, a CNN is an artificial network that includes a convolutional layer and is capable of detecting patterns in data. This feature detection makes CNNs particularly useful for image analysis. Neural networks are composed of interconnected information processing units called neurons. These neurons process information simultaneously and are organized in layers, each with their own learnable weights and biases. The function of neurons is to take in multiple inputs and yield a single output. Neural networks are capable of generalizing from learned examples and predicting new situations based on them. CNNs are distinguished from typical neural networks by their additional convolutional operation, which makes them well-suited for learning from images and image-related data [Hu+21] [CO21] [Sel+19].

Each convolutional layer is associated with a distinct filter kernel. A layer can be considered as a group of neurons that perform the same operation, with identical hyperparameters. Hyperparameters are the parameters of a machine learning algorithm that determine the network structure and the learning process, such as the learning rate. A convolutional layer contains learned kernels that extract features distinguishing different images. The convolutional operation is the summation of element-by-element multiplication of the convolutional neuron and the corresponding neuron of the previous layer. Specifically, when the convolutional layer receives input, the center element of the filter kernel is placed over the source pixel on the input image. Afterwards, the original pixel is substituted with a weighted sum of itself and adjacent pixels. This process is known as convolution, which involves repeatedly applying a convolutional procedure resulting in a feature map that summarises the strength and location of a detected feature in an image. Since images contain multiple edges, shapes, textures, and objects, features identify those distinctive decorations and patterns. After constructing a feature map, each value in the feature activation map is passed through a nonlinearity. The most common nonlinearity used in neural networks is the rectified linear activation function (ReLU), which improves performance and facilitates model training. This default activation function returns the input directly if the result is positive, and outputs zero otherwise. The ReLU equation is given as:

$$f(x) = x^+ = max(0, x) \tag{4.1}$$

where x is the input to a neuron. The overall objective of the ReLU operation is to introduce the property of nonlinearity [Sub18] [Sch20]. In deep neural networks, additional self-normalizing activation functions like the scaled exponential linear unit (SELU) are advantageous. Layers with SELU activation functions have a more stable input distribution, which means that the training is more stable overall. The SELU activation function is given by:

$$selu(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leq 0 \end{cases} \tag{4.2}$$

The input value of a neuron denotes x. If x is greater than 0, x remains unchanged. Otherwise, the SELU transformation is applied, which performs a layer-by-layer normalization and scaling. The values of α and λ are pre-defined constants ($\alpha = 1.67326324$ and $\lambda = 1.05070098$), which must be selected in such a way that the variance and the mean value of the inputs between successive layers are maintained [Kla+17]. If there is a multiple classification, the Softmax function forms the output layer of CNNs. This procedure is given as:

$$softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{4.3}$$

where exp represents the standard exponential function. Exp is applied to each element of the given input which calculates positive values above 0 but not fixed in the range $(0, 1)$ which is required for the calculation of probabilities [Sub18]. The bottom of the formula ensures that all values will sum to one and fall with the range of $(0, 1)$. In other words, softmax function's main purpose is to ensure that the outputs of CNNs add up to one. Each class corresponds to an appropriate probability after passing through this function. The difference between a softmax function and a standard normalization lies in the usage of the backpropagation method for training neural networks. The backward procedure is based on the principle of error elimination, where input patterns are compared with output patterns to calculate the error rate. Backpropagation propagates the errors back to the input layer via the output layer to approximate the desired result. However, the use of the softmax function ensures that the correct answer is given the highest weight, as it is initialized with the largest signal. This is because the max is significantly weighted higher than other logits. Additionally, rescaling operations apply equal weight to every input [SZ15].

Pooling layers, such as Max or Average pooling, are commonly introduced between successive convolutional layers to gradually decrease the spatial extent of neural networks. This reduces the number of parameters and total amount of calculation and can help counteract overfitting through the use of the MAX operation. Max-pooling layers operate independently and slide over each depth slice of the input with a specified kernel size. They select only the largest value at each kernel slice from the input to yield a value for the output [CO21] [SZ15]. Fully connected layers compile the data extracted by the previous layer and classify it into various classes. Unlike convolutional layers, fully connected layers are connected to every activation unit. This means that the calculations of each previous neuron flow into the calculation of the current neuron, resulting in many dependencies that are not all relevant but require a lot of computing power [CO21] [Sel+19] [SZ15].

4.1.2 Model Training

This section focuses on the theoretical application, significance and necessity of the model learning process, providing a clear foundation for understanding the specific aspects and characteristics of the ResNet50 model within a broader theoretical framework. Transparent foundations enable the identification and interpretation of potential strengths, weaknesses, and implications of the model and its learning process.

The training of a neural network is characterized as an iterative process with the objective of optimizing the model for accurate predictions on new data. Achieving optimal results is considered time-intensive, as it necessitates both experimentation and empirical knowledge, alongside a final fine-tuning stage. Training constitutes the pivotal phase in machine learning. During training, the prepared data is fed to the machine learning model for pattern recognition and prediction. This process facilitates the model's acquisition of knowledge from the data, enabling it to execute the designated task [Fas22]. The model training process generally can be partitioned in the following steps:

1. Data collection
2. Data preparation
3. Model selection
4. Model evaluation
5. Parameter tuning
6. Predictions

The initial step establishes the foundation for optimal results, as data selection is of paramount importance to ensure that the model can identify the correct patterns and subsequently classify effectively. It is imperative to utilize data solely from reliable sources, where it can be inferred that the data is neither outdated nor incorrect [Ban23] [Fas22]. The data preparation step ensures that the data to be processed does not contain duplicates, unwanted or missing values. Subsequently, the cleaned data is partitioned into three sets: the training set, validation set and test set. While the training set is only used to train the model, the validation set, also known as the development set is solely for evaluation purposes. This enables us to assess whether the model derives insights from the training data that can be applied to novel data, represented by the validation dataset. On the other hand, if the model were not divided into training and validation data, the evaluation would consistently occur using the same data it was initially trained on. Consequently, the results would be skewed or embellished, as the model would already be familiar with the data, leading to a biased performance.

> One way to understand this situation is that, in a sense, we don't want our model to get good results by "cheating". If it makes an accurate prediction for a data item, that should be because it has learned characteristics of that kind of item, and not because the model has been shaped by *actually having seen that particular item* [Fas22].

However, in most realistic scenarios, the process does not end with constructing the model by training it once and then evaluating it. Instead, a cyclical process occurs where, based on the evaluation, parameters are adjusted, and the model is retrained. This is the point at which the test set comes into play. Just as the validation set is withheld from the training process, the test set is now kept separate from the recurring evaluation. More specifically, the test set represents even more highly reserved data meaning its sole utility emerges for the model evaluation in the final stage of model building but it cannot be employed for model improvement. Essentially, a hierarchy of data subsets can be established, determined by the extent of isolation required from the training and modeling processes: the training data is entirely exposed, the validation data is relatively less exposed, and the test data remains entirely concealed [Fas22]. Typically, the training set comprises the largest portion, accounting for approximately 70% of the original dataset. The remaining portion of the datasets often divides the percentage equally. Of course, the total dataset size significantly influences the proportional data allocation. Primarily, it is essential to ensure that both the validation and training sets have a sufficient quantity of data available to achieve optimal model accuracy [Fas22]. For the model selection step it is imperative to consider whether the model is suitable for the specific problem. A machine learning model defines the result obtained following the execution of a machine learning algorithm on the acquired data. The selection of an appropriate model relevant to the specific task is crucial. Over time, researchers and engineers have created diverse models tailored to distinct tasks such as speech recognition, image recognition, prediction and more [Sch15] [Sch16] [Sel+19]. Model evaluation and parameter fine-tuning are two interrelated components, as parameter settings evolve based on validation, leading to the subsequent evaluation of the finely tuned model. With successive training iterations, the model's predictive capabilities improve. After the training phase, the evaluation measures how the model performs and if the model meets the standards for high accuracy. Then predictions and classifications follow [Ban23].

4.1.3 Limitations of Deep Learning

Despite the prospect of boundless opportunities, deep learning, especially model training, still has some limitations and challenges. Discussing the limitations of DL is crucial to gain an understanding of its applicability and effectiveness in different areas. This also highlights areas that need improvement or overcoming, providing impetus for future research and development [KM23]. It is particularly relevant in the rapidly evolving field of machine learning, where continuous innovation and progress are sought [Jea+23].

Overfitting is one of the most common and problematic challenges as it occurs when a machine learning model is highly attuned to the training data, leading to an overly precise fit that hampers its ability to generalize effectively to new, unseen data [JK15] [Fas22].

Underfitting in contrast refers to the issue where the data is overly tuned to a model, specifically involving only relevant variables and parameters directly related to the problem. This can lead to distorted conclusions and a substantial bias within the model. Where overfitting stands for too much complexity, occurring when the model has been overtrained, resulting in high error rates in test data, underfitting refers to undertraining or lack of complexity results. Mostly, underfitting occurs as a side effect from preventing overfitting because the scenario of training interruption is employed to reduce model complexity [JK15] [Fas22]. Detecting overfitting can prove to be more challenging compared to underfitting, primarily due to the fact that, unlike underfitting, an overfitted model exhibits high accuracy on the training data. To evaluate the algorithm's accuracy, a method known as k-fold cross-validation is commonly employed.

Hyperparameter optimization refers to the several hyperparameters of deep learning models. They typically can be adjusted by trial and error method. Several research studies have identified six most influential hyperparameters that exert an immediate and significant impact on the model's quality and computations. These include: initial learning rate, batch size, dropout, exponential decay factor for learning rate which on the one hand refers to the training procedure of fully connected layers and on the other hand the architecture with its numbers of layers and units per layer can be highly influential [KM23] [Bar23] [FKH18]. Several algorithms have been developed to find and optimize hyperparameters in the context of deep learning. Notable examples include the Metropolis-Hastings (MH) algorithm and Bayesian optimization. The MH algorithm, a Markov Chain Monte Carlo (MCMC) method, provides a probabilistic framework for exploring hyperparameter space. It can efficiently select hyperparameters based on how likely they are to improve model performance. The Bayesian optimization on the other hand, uses probabilistic models to efficiently search for optimal hyperparameters. It balances the exploration-exploitation trade-off by choosing hyperparameters that are likely to yield the best performance given the model and its uncertainties. The goal of hyperparameter optimization is always to achieve robust performance and fast convergence to optimal settings.

Computational power represents a very common and well-known issue in terms of deep learning methods. Due to the complexity and deep net structures, the result computations become intensive and time- consuming especially if the aim is, to tackle a real-world problem. Within the scope of technical prerequisites and performance parameters, it is pertinent to provide a preliminary overview at this point. Further elucidation pertaining to the selected ResNet50 model will follow in subsequent sections.

The ResNet50 model training and finetuning by [M20] has been done using Google Colab which has become highly established as it provides powerful hardware with access to Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), which can significantly accelerate the model development process. Using Ubuntu 18.04 64-bit as the operating system (OS) offers a widely employed system for the scientific context especially in the field of ML and DL. The Random Access Memory (RAM) of 12 GB DDR4 is sufficient to handle various tasks, although specific requirements vary according to the complexity of DL models. Having more RAM can be advantageous when dealing with expansive datasets or complex models. The Nvidia Tesla P100 is a powerful GPU and is well-suited for profound learning errands. GPUs are especially accommodating for computationally seriously assignments such as neural machine learning. Depending on the specific needs of the intelligent system, 40 GB of non-persistent storage can be sufficient. In certain instances, there may arise a need for larger or enduring storage media especially in case of preprocessing a larger amount of data [M20] [Cla22] [KM23].

Gradient-based learning is one of the most frequently used DL processes as it has been approved by many different past studies. An exemplification that involves gradient-based computations is the application of backpropagation. This method calculates the gradients of the loss function relative to model parameters. The neural network architecture is yet another example that follows a common pattern. It consists of multiple layers of neurons, where the weights are continually adjusted through gradient computations. On the downside, gradient-based methods exhibit significant limitations. These include getting trapped in local minimums when dealing with multi-objective cost functions, extensive computation time as a result of calculating gradient information through numerous iterations, and the requirement for continuous cost functions. The considerable advancement in optimizing the structure and parameters of Artificial Neural Networks (ANNs) and DL models using meta-heuristic algorithms can be attributed to the fact that training these models is a NP-hard (nondeterministic polynomial time) optimization problem [KM23]. As a special category within complexity theory, NP-hard comprises decision problems in computer science that are notoriously challenging to solve. The issues at hand fall under the category of NP problems. This implies that if a solution is provided, it can be easily confirmed; however, the process of finding a solution within polynomial time remains undisclosed [Pas16].

The determination of the appropriate type of deep learning architecture can be very challenging since there seems to be an absence of clear explanations regarding which architectural solutions effectively address specific issues [KM23]. CNNs are considered state-of-the-art and excel in classification of plant diseases [HYA20], object detection [Xia+20] or applications in medical image analysis [DVV22]. A large number of DL review papers, mostly focusing on CNNs, have been published in recent years [Tal+19] [YC20] [DVV22]. However, most of them have gaps in the treatment of key aspects such as methods and concepts, technical tools and the application for the implementation and computation of the model. Since 1989, continuous improvements and adaptations have been developed for optimal model architectures. These include: structural reformulation, regularization and parameter optimizations. However, the main reason for the improved performance of the latest CNN models is the reorganization of the process units and the integration of new block units, which are essential for the model calculations. In particular, recent developments have shown that adjustments in model depth are crucial for optimal performance [Alz+21]. Despite this, research is increasingly focusing on the aspect of improving model architectures, and the question of which model architecture is best used for which problem usually remains unanswered [KM23]. However, by studying the architecture features such as input size, depth and robustness, the choice of a suitable architecture for a classification task can be justified [Alz+21]. Although the model architecture is a very large influencing factor for the results and predictions, the size and diversity of the data set is an even more important one. DL models are extremely data-hungry and demand for large data amounts to achieve a well-behaved performance [Alz+21] [KM23].

The dataset amount and availability greatly influence the performance of neural networks since the quality of the models is dependent on the dataset provided for training. However, one of the major hurdles faced by DL architectures is the lack of availability of the dataset [KM23]. To properly address this issue, three key concepts have been identified. The first refers to the widespread concept of transfer learning. It is essential that data from similar tasks is collected and then applied to the current problem. The transferred data does not directly augment the actual data but helps to improve the original input representation and the mapping function. The second refers to the popular technique of fine-tuning in which an already well-trained model of a similar task is used as a basis and the last one or two layers of this model are fine-tuned. Data augmentation is an important component and is performed to expand image data, as common techniques such as image shifting, mirroring and rotation do not normally change the image label [Alz+21]. The ResNet50 model in this work is based on the finetuning concept, as the last layer of the already established ResNet50 architecture of PyTorch was adapted and modified to the current classification problem and data set [Sub18] [M20]. More details of this process will be provided in 4.2.2. The last method incorporates simulated data into the training process in order to increase the volume of the data set. It utilizes the possibility that simulators are defined based on the physical understanding of the process. However, it is essential that the problem is well understood by the model. If this is the case, as much simulated data can be generated as required for well-behaved model performance [Alz+21].

In addition, there are some other limitations that significantly affect the efficiency and accuracy of the model, which can be summarized as follows: the importance to contain heterogeneity in the image dataset, imbalanced data, one-shot learning and lack of results for different DL architectures in the benchmark database. The first and second point refer to the fact that the data must contain various different and heterogeneous images but the predicted classes still need to have enough representatives in the training process, to be classified correctly. Thirdly, to achieve high-quality results, DL architectures for One-Shot Learning necessitate a substantial amount of training data. To illustrate, the Image-Net database boasts over a million images, and effectively classifying them necessitates thousands of instances with the DL architecture. Humans do not require an excessive number of bicycle images in order to learn what a bicycle looks like. When children are presented with a bicycle, they frequently have the ability to identify another bicycle, even if it appears in diverse models, shapes, and colors. Finally, solving various benchmark database or benchmark engineering problems still poses a challenge due to the inadequate outcomes from different DL architectures [KM23].

4.2 ResNet50

This section proceeds with the presentation of the implemented ResNet50 in the Python case study. Initially, an overview of the fundamental operations of residual networks is provided to enhance comprehension of the operational mechanism of the implemented network. Subsequently, an exposition will detail the distinctions of ResNet50 from other models to elucidate why the chosen model is suitable for the problem resolution of tumor image classification. Additionally, the specific model architecture, the model's learning process, and ultimately, the model's performance will be delineated.

Residual Networks, or ResNets, learn remaining capacities with reference to the layer inputs, rather than learning unreferenced functions. The core idea of this network is the so-called residual block [Zha+23]. To make the meaning and procedure of residual mapping more understandable, this is explained in more detail using the following example. Imagine a labyrinth situation: the goal is the way out and the steps are limited. Each step symbolizes a layer in a neural network. Usually, the first intention would be to take the direct path from the starting point to the destination. This would be comparable to each layer in the ResNet trying to find the answer to the classification question. Residual mapping takes advantage of the use of differences. More precisely, only the differences or changes to the previous reference point are considered. Within the labyrinth situation, it can be transferred to the fact that markers are used to guide the way back to the last starting point if a path has been lost. The markings are like a reminder of the original path. Back to the neural network, this means that the classification is based on differences between the starting point and the current or end point, and if calculations become too complex, the memory of the previous position helps to find the answer. This principle makes learning easier for neural networks and enables deep networks to be trained in a more efficient way. To put it in a more mathematical context, in contrast to a normal block, a residual block needs to learn the residual mapping: $g(x) = f(x) - x$ when x denotes the input and $f(x)$ represents the desired classification of the input. Hence, $f(x) = x$ displays the identity mapping. Since $g(x) = 0$, residual blocks allow faster access through residual connections between layers. Basically, residual blocks have two branches, one of which is the identity mapping. As a result, the residual mapping makes the identity mapping easier to learn [Zha+23].

In comparison to traditional neural networks, ResNets enable to skip information from one layer to another which most likely prevents gradients from vanishing or exploding. Additionally, residual networks can converge faster and achieve good results with less training data and they tend to generalize better on new data. This allows more effective application to different tasks and data sets [Zha+23]. Deep neural networks can be prone to vanishing or exploding gradients, making training difficult. Residual mapping allows information to be passed from one layer to another, reducing missing or bursting problems [He+15]. The gradients can be seen as hints, back to the maze example this would mean that the gradients help to recognize which step would be the next one to get to the goal (i.e. to learn better). Gradient fading in this case means that the gradients become very small, so the clues are very faintly recognizable or invisible. The network then has difficulties in learning because the adaptations are barely noticeable. In contrast to this is the gradient exploding, as the gradients become very large, it is as if the clues are too strong. The network then forces very large steps and the challenge is to find the best path. To counteract this, the residual mapping serves as a guiding support so that the network does not run the risk of concentrating on too small or too large leaps, but instead uses the available markers as a guide to arrive at the best solution or to carry out the most effective learning process [He+15] [He+16] [Zha+23]. Overall, ResNets have established themselves as successful models, particularly in the field of image recognition [Tal+19] [M20] [KG20] [SK22].

4.2.1 ResNet50 in Comparison

To provide an overview of Resnet50 capabilities and assess its performance in 4.2.4, this section outlines various studies in the field of deep learning and their results. A comparison with other DL models and research is presented, to demonstrate why ResNet50 is particularly suitable for the problem of tumor classification.

According to a study conducted by [Sha+23], it has been demonstrated that the ResNet50 neural network exhibits superior performance and accuracy, making it a prime choice for application in this work. In comparison to the models such as Inception V3, VGG16 and VGG19, the ResNet50 reached a very high accuracy of 99.75% with a loss rate of 0.33, while the other models achieved 98.16%, 98.47%, and 98.56%, respectively [Sha+23]. The accuracy is a useful metric that measures the number of predictions that are correct as a percentage of the total number of predictions that are made. Therefore, the accuracy of 99.75% simply means that 99.75% of the total data amount has been classified correctly. The loss rate serves as the foundation for subsequent computations in the neural network, leading to the adjustment of weights to finally improve the model predictions in the next iteration. The goal is to minimize the loss value during the training process [Bot21]. Shah et al. substantiate the deficits in model performance in particular with a lack of effective propagation of gradient problems among the other models. Residual mapping from ResNets is in contrast to this, as it facilitates the return flow of gradients during training by skipping certain connections. Shah et al. refer to this term as the concept of connection skipping [Sha+23].

Another recent study by [SK22] comparing the ResNet50 with VGG16 and VGG19 models yielded comparable favorable outcomes for ResNet50 since it produced the most accurate results. The models have been tested for classification of lung cancer types using non-contrast enhanced medical images.

Talo et al. further emphasized the performance of ResNet50, highlighting its efficacy and significant results [Tal+19]. The study compared the neural networks AlexNet, VGG16, ResNet18, ResNet34 and ResNet50 for the detection of multi-class brain diseases using MRI images. As a result, the ResNet50 achieves the highest classification accuracy of 95.23% ± 0.6 better than the other four models. The conducted study utilized the Harvard Medical School MR brain images dataset, comprising all 1074 available images. Each of the five selected models was subjected to testing across five categories, including Accuracy, Precision, Sensitivity (Recall), Specificity, and the F1 Score. Across all categories and also across all five classes (normal/ healthy, cerebrovascular disease, neoplastic disease, degenerative disease, and inflammatory disease), the ResNet50 model demonstrated a significant performance advantage, consistently outperforming all other models [Tal+19]. The commonly used F1-Score also known as the harmonic mean informs about the balance between precision and sensitivity and helps to evaluate the total performance of the classification process. Hence, a high F1 value concludes a delicate equilibrium between precision and sensitivity. Precision is a measure of the reliability of retrieval as it identifies the relevant hits within the overall set. With other words, it reveals the positively predicted values. Specificity outlines the ability of a model, to correctly identify negative results. More precisely, this true-negative rate focuses on the false identified positive results, in this case the fraction of disease-free individuals who also produce a negative test outcome. Hence, it is the counterpart to sensitivity as the latter also known as the true-positive rate measures the percentage of people with a disease which have been correctly identified as non-healthy by the model [DVV22].

It is noteworthy to mention that another study published by [Kha+20] in the year 2020 has reported contrasting findings. A comparison between VGG16, ResNet50, and Inception-V3 yielded accuracy values of 96%, 89%, and 75%. However, even the best result of 96% obtained does not surpass the accuracy measurements of 99.3% reported by [M20] which contributes to the ResNet50 model used in this work. Additionally, it is pertinent to acknowledge that the research conducted by [Kha+20] is predicated upon a comparatively limited dataset comprising 253 MRI brain tumor

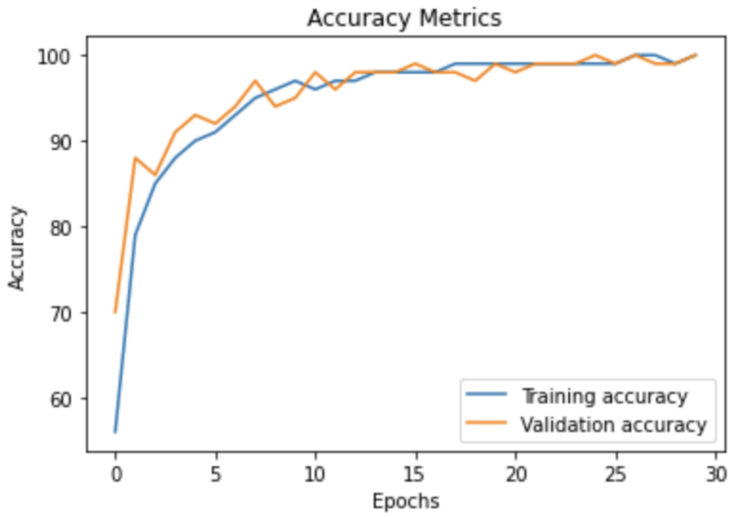


Figure 4.1: ResNet50 Accuracy Overview [M20]

images. Consequently, its inclusion within this discourse is primarily out of a sense of comprehensiveness.

All in all, the ResNet50 by [M20] in this work is rated with a very high accuracy value since an accuracy measure above 90% symbolises an excellent model performance. This is also consistent with industry standards [Hin+23] [Bar22]. Figure 4.1 underscores the training and validation accuracy values approaching 99%, ultimately signifying that nearly every input image has been correctly classified. Within the realm of neural networks, a solitary traversal through the feedback loop is commonly denoted as an epoch. In this context, 30 epochs refer to thirty complete passes or iterations through the entire training dataset. The greater the number of epochs, the longer the neural network was trained. During a single epoch, the neural network processes each data point once and updates its internal parameters, such as weights and biases, based on the observed errors. This iterative process is integral to the network's learning mechanism, as it allows the model to gradually adjust and improve its performance in response to the training data. The number of epochs in the training process is a hyperparameter that can influence the model's convergence and overall performance, with a higher number of epochs potentially leading to improved accuracy but also a risk of overfitting [Kha+20] [Tal+19] [M20]. Hence, it is important to define the right epoch number and to prevent the neural network from overfitting. [MH20] evaluated different numbers of epochs using a recurrent neural network (RNN) with data consisting of the daily opening prices of two stocks in the New York Stock Exchange. The analysis demonstrates that very promising results are achieved even with an epoch count as low as 25, which is immediately reflected in a significantly low loss value (0.001). For this reason, it can be inferred that the epoch value of 30 was selected because it operates both efficiently in terms of performance and effectively minimizes errors. In the realm of critical parameters, the selection of the models batch size stands as another key determinant. In this case, the training and the validation set contain a batch size of four and the

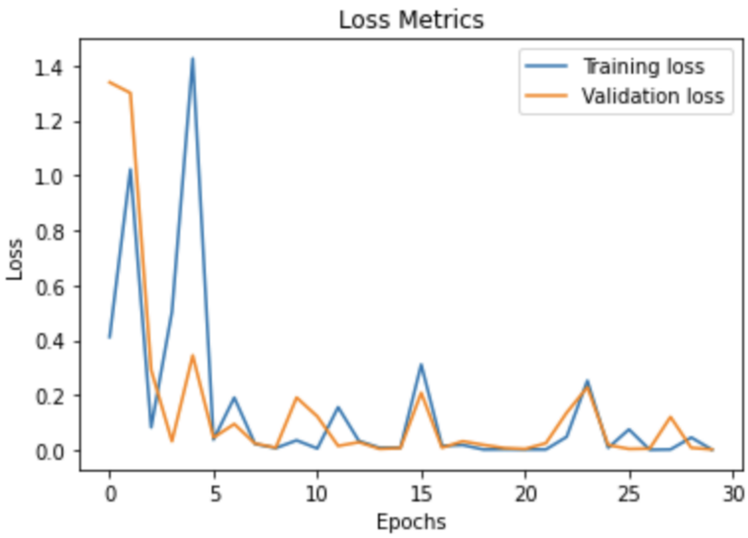


Figure 4.2: ResNet50 Loss Overview [M20]

test set of ten [M20]. The velocity and convergence of model learning are influenced by the selected batch size. A larger batch size may lead to faster convergence of the model, requiring fewer epochs to achieve satisfactory results. Conversely, a smaller batch size may result in slower learning and necessitate a greater number of epochs to attain good performance. Typically values between 32 and 256 are referred to as optimal batch sizes. Aware of the potential limitations of using a large batch size, which may significantly affect speed and memory consumption, it is important to emphasize that training accuracy can also be attained with smaller batch sizes, provided the training duration is extended. Consequently, the ResNet50 model was trained utilizing what are known as mini-batches [Tal+19] [Kha+20] [DVV22].

Figure 4.2 illustrates the cross-entropy loss metric which aims to measure the performance of a classification model by calculating the error rate as it quantifies the error between the model predictions and the actual labels [DVV22]. It is evident that with an increasing number of epochs, the loss values significantly decrease and ultimately approach near-zero levels. More specifically, the smaller the loss value, the better the model. Consequently, the model yields a loss value which indicates an optimal model performance with a minimum of false predictions [M20].

4.2.2 ResNet50 Architecture

The application of deep neural networks requires a profound understanding built on years of experience in the field. Especially the decision of hyperparameters, regularization and network architecture, which are all tightly coupled, is fundamental to the overall network performance [Smi18]. Adhering to Aristotle's principle, it can be observed that a well-begun task is already half-done. Hence, the architectural design establishes the initial conditions and framework for subsequent computations and operations [Smi18].

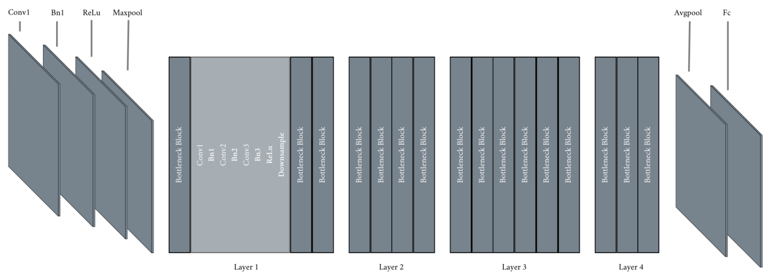


Figure 4.3: ResNet50 Overview

To start with, the ResNet50 model basically follows the model structure of PyTorch's pre-trained ResNet50 model [Sub18] [M20]. The general differences to the original PyTorch model are, on the one hand, that all parameters of each model layer are released for training. This means that all weightings can be adjusted across all layers during the training process. On the other hand, the fully-connected layer is redefined to adapt it to the specific classification problem. For this, the number of input features is extracted to generate an adaptation to the tumor characteristics. Additionally, a sequence of new layers is introduced to better fit the non-binary tumor classification. This includes the introduction of a SELU operation, the modification of droupout layers and the adaptation of the last layer to four outputs and a log-sigmoid operation [M20]. The utilization of this activation function in the context of multiclass classification tasks deviates from conventional practices, but in this order leads to a value range from 0 to 1 for each of the four outputs [Sub18] [Alz+21]. More precisely, each of the four output classes is assigned a probability that is independent of the other classes. Contrary to the Softmax application, the sums therefore do not add up to 1 [Dra19]. In order to create comparability between the classes and for subsequent analyses of the behavior of the model, a softmax operation is performed in addition to the fine-tuned model of [M20]. These calculations based on the Softmax activation function have been supplemented in particular to utilize the advantages described in 4.1.1 and to create interpretability for multiclass labels. Thus, if softmax is applied, in order for the probability of a certain class to increase, that of another class must decrease, resulting in the probability of occurrence in an input image for each of the four classes [Dra19]. Further details on the ResNet50 training process are provided in 4.2.3, as the focus here is on the model architecture.

Figure 4.3 illustrates the fundamental architecture of the ResNet50 model tailored for brain tumor detection. It can be seen that the model primarily comprises four bottleneck layers (*Layer 1 - Layer 4*), each predominantly composed of a varying number of bottleneck blocks, culminating with Average Pooling and the Output Layer (*Fc*). *Layer 1* in 4.3 exemplifies what lies behind an individual bottleneck layer, consisting of a pair of convolutional layer ($Conv_x$) and batch normalization (Bn_x), followed by a ReLU operation. Downsampling occurs exclusively within the first bottleneck block of each layer. The adoptation of batch normalization after each convolution (and before activation) ensures regularization to accelerate model training and promotes model stability, as activations fluctuate less and data flows more consistently through the network during training [Hu+21].

layer name	output size	34-layer	50-layer
conv1	112×112	7×7, 64, stride 2 3×3, max pool, stride 2	
conv2-x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3-x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
conv4-x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
conv5-x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$

Table 4.1: Architecture for ResNet34 and ResNet50. Building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed by conv3-1, conv4-1, and conv5-1 with a stride of 2 [He+15].

Table 4.1 highlights the visual view of the ResNet50 with the underlying numerical values. The table shows the image size to which the downsampling has reduced the input image and how many bottleneck blocks fill a complete bottleneck layer. Further it becomes clear, that for the architecture of 50-layer ResNet, each 2-layer block in the 34-layer net is replaced with 3-layer bottleneck blocks. The advantage lies once again in the principle of residual mapping as every building block is modified to a bottleneck design, resulting in a stack of 3 layers which are 1×1, 3×3 and 1×1 convolutions. The 1×1 layers are responsible for diminishing and after that expanding (reestablishing) dimensions, leaving the 3×3 layer a bottleneck with smaller input/output measurements [He+15]. The distinction between building blocks and bottleneck blocks with residual mapping is further elucidated by figure 4.4: the skip connection (identity shortcut) creates a direct path from the input to the output of a part of the network. It should be emphasized that these connections are parameter-free and therefore no additional weights or parameters are introduced, which in turn contributes to the efficiency of the architecture. More concrete, within the bottleneck block, starting with a 1×1 convolution, the dimension is reduced (and increased again at the end) while the 3x3 convolutional layer processes the actual spatial information. This allows

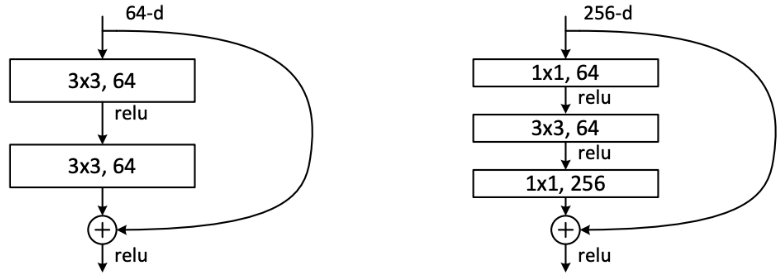


Figure 4.4: Comparison of block design for ResNets. Left: a building block (on 56×56 feature maps) for ResNet34. Right: a *bottleneck* building block for ResNet50 [He+15].

feature learning to be more efficient because fewer parameters are used while maintaining the generality of the data [He+15].

While Figure 4.3 has illustrated the broad architecture, Figure 4.5 provides insights into the specific details within each layer to gain a deeper understanding of the ResNet50 model's structure and further explore its characteristics. The layer column conveys that the model's architecture comprises various layers, including convolutional layers, batch normalization and bottleneck layers. Output shapes denote the output size for each individual layer in the model, which is crucial for comprehending how the network processes data dimensions. In this case, the output shapes are represented in the following format: [batch size, number of channels, height, width], which, for the last convolutional layer $Conv2d - 169$, specifically implies that it contains 2048 channels and that each feature map has a size of 7x7 pixels. The value -1 pertains to the adaptable batch size, which is not fixed and varies according to the data. The parameter count indicates that the model includes approximately 31,9 million parameters, all of which can be optimized during training. Additionally, the model incorporates activation functions such as ReLU and SELU. Dropout layers primarily serve to prevent overfitting by stochastically deactivating some neurons during training. Finally, the model's last layer is represented by logistic regression, as it is a non-binary classification model with four distinct classes. In total, the customized ResNet50 comprises a neural network consisting of 50 layers, culminating in the fully connected layer which produces 4 outputs. This configuration is tailored for a non-binary classification task encompassing 4 distinct classes (no tumor and 3 different tumor types) [M20].

However, the network architecture unveils that the final convolutional layer exhibits an extensive parameter count, exceeding one million parameters. Typically, higher parameter counts suggest the capacity to capture more intricate patterns within the provided data. This augmented model

[1]This graphic for the model summary is a shortened version of the original summary output, which was created using the torchsummary module from the PyTorch library. As the model summary normally shows all layers of the ResNet50, the middle section has been replaced by a placeholder for reasons of clarity and readability.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1         [-1, 64, 112, 112]           9,408
       BatchNorm2d-2         [-1, 64, 112, 112]             128
              ReLU-3         [-1, 64, 112, 112]               0
         MaxPool2d-4           [-1, 64, 56, 56]               0
            Conv2d-5           [-1, 64, 56, 56]           4,096
       BatchNorm2d-6           [-1, 64, 56, 56]             128
              ReLU-7           [-1, 64, 56, 56]               0
            Conv2d-8           [-1, 64, 56, 56]          36,864
       BatchNorm2d-9           [-1, 64, 56, 56]             128
             ReLU-10           [-1, 64, 56, 56]               0
           Conv2d-11          [-1, 256, 56, 56]          16,384
      BatchNorm2d-12          [-1, 256, 56, 56]             512
           Conv2d-13          [-1, 256, 56, 56]          16,384
      BatchNorm2d-14          [-1, 256, 56, 56]             512
             ReLU-15          [-1, 256, 56, 56]               0
       Bottleneck-16          [-1, 256, 56, 56]               0
           Conv2d-17           [-1, 64, 56, 56]          16,384
      BatchNorm2d-18           [-1, 64, 56, 56]             128
             ReLU-19           [-1, 64, 56, 56]               0
           Conv2d-20           [-1, 64, 56, 56]          36,864
      BatchNorm2d-21           [-1, 64, 56, 56]             128
             ReLU-22           [-1, 64, 56, 56]               0

                                     ⋮

          Conv2d-166            [-1, 512, 7, 7]       2,359,296
     BatchNorm2d-167            [-1, 512, 7, 7]           1,024
            ReLU-168            [-1, 512, 7, 7]               0
          Conv2d-169           [-1, 2048, 7, 7]       1,048,576
     BatchNorm2d-170           [-1, 2048, 7, 7]           4,096
            ReLU-171           [-1, 2048, 7, 7]               0
      Bottleneck-172           [-1, 2048, 7, 7]               0
AdaptiveAvgPool2d-173           [-1, 2048, 1, 1]               0
          Linear-174                 [-1, 2048]       4,196,352
            SELU-175                 [-1, 2048]               0
         Dropout-176                 [-1, 2048]               0
          Linear-177                 [-1, 2048]       4,196,352
            SELU-178                 [-1, 2048]               0
         Dropout-179                 [-1, 2048]               0
          Linear-180                    [-1, 4]           8,196
      LogSigmoid-181                    [-1, 4]               0
================================================================
Total params: 31,908,932
Trainable params: 31,908,932
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.57
Forward/backward pass size (MB): 286.64
Params size (MB): 121.72
Estimated Total Size (MB): 408.94
----------------------------------------------------------------
```

Figure 4.5: ResNet50 Model Summary [1]

complexity also renders it more susceptible to overfitting, a concern exacerbated by limited training data. Within this context, the substantial parameter count signifies the layer's ability to capture high-dimensional features within the data and acquire knowledge, particularly beneficial for the extraction of abstract features [DVV22] [M20]. Consequently, to prevent overfitting, this model contains dropout layers to randomly deactivate neurons during training in order to reduce inter-neuron dependencies and secondly, the batch normalization layers aim to stabilize activations during training [DVV22] [M20]. On the other hand, the existing model complexity ensures that there is no underfitting. The model architecture, including the number and type of layers, is particularly important here. As this network is considered a deeper bottleneck architecture, it can be assumed to have sufficient model complexity [He+15]. With almost 32 million parameters, this network ranks with a higher parameter value. This value influences the capacity and expressivity in the model and these properties are particularly important for more complex tasks or larger data sets [Hu+21].

4.2.3 ResNet50 Training

After introducing the general method of model training, comparing the ResNet50 structure and its basic properties with different classification models and describing the basic architecture of the ResNet50 deep learning model, the following section focuses on the training process of the ResNet50 model based on [M20], to provide an insight in the initialization strategies and the regularization techniques.

Following the established structure in 4.1.2, the key elements of the model training process, namely data collection, data preparation, model selection, model evaluation, parameter tuning and predictions will be outlined. Thus, for the effective model training, the foundational dataset is based on [Che17] and records an extensive utilization in a variety of research studies [Reh+20] [KG20] [AZ21] [Haq+23]. Due to its versatility and widespread application, the dataset has proven to be a robust and relevant choice. In addition, there is a high demand for current medical MRI images, but only limited access to research data [Saj+19]. To prepare the data for the brain tumor classification problem, a custom dataset class augments the input images into eight different angles, including 0, 45, 90, 120, 180, 270, 300 and 330 degrees [M20]. Resulting in a total of 24.512 augmented brain tumor input images based on the original number of 3.064 MRI images of meningioma, glioma and pituitary tumors, the dataset has expanded eightfold to finally be divided into training, validation and test sets [Che17] [M20]. The model selection is basically attributed to the established utilization and successful application of ResNet50 deep learning structures in the medical field [Tal+19] [M20] [KG20] [SK22]. After the iterative interplay between the process of model evaluation and parameter finetuning, [M20] finalized the ResNet50 training configurations with the model's loss function as Cross Entropy and a Stochastic Gradient Descent (SGD) as the optimization algorithm with a momentum of 0.9 and a learning rate of 3e-4 based on the recommendation of deep learning experts [M20] [Kar23]. Overall, [M20] established a standard training pipeline for neural networks using the PyTorch library, including training, validation, and checkpoint saving to finally result in a reusable finetuned model specialised on brain tumor classification. The SGD optimization approach has recently received much attention in the context of large-scale learning and is responsible for updating the model parameters based on the the best fit between the prediction and the actual data. The learning rate determines the size of the update or moving step, and the momentum helps to overcome local minima and accelerate convergence. A process converges when parameters or variables reach a stable state and no longer change significantly [Ped+11] [Sto23]. Setting the momentum hyperparameter to the value 0.9 is a common and established method [Sub18] [LGY20]. The learning rate was chosen on the basis of Andrej Karpathy, a specialist in deep learning and computer vision [Kar23]. It is important to emphasize that no *best* learning rate can be identified, as this is highly dependent on the model architecture and dataset. Nevertheless, the learning rate is considered the most important hyperparameter for training artificial neural networks, as it influences which network parameters (weights and biases) the model learns [Mor18] [Pra22]. In other words, if the learning rate is set too low, the parameters can only change in very small steps, resulting in the model taking too long to converge. If the learning rate is too high, on the other hand, there is a risk of skipping low spaces of the loss function, which can ultimately lead to the model never reaching a convergent state [Mor18]. The loss function is a metric in machine learning that measures, how well the model performs on the training data within a certain classification problem. The goal is to achieve a low loss function, as a low loss value means that the model can predict the training data well [Sub18] [DVV22]. However, the main goal of neural networks is not to achieve optimal performance on the training data, but on the test data and on new, previously unknown data. This is achieved by finding not only low spots of the loss function but also wide spaces during training. The wide space refers to the fact that the model operates in a large area of the possible solution space and therefore reacts more robustly and generalized to new data [Mor18]. In terms of classifying the three brain tumor types, a wide solution space would mean that the model can recognize general patterns of specific brain tumors. This is particularly important as anatomies are not necessarily similar but, biologically speaking, are individually different. If the model was only adapted to the specific training data, it could have difficulties in predicting tumors that have slight variations in shape or location. In summary, the model was developed through the training process and the repeated parameter tuning based on the ResNet50 of the PyTorch library, resulting in a classifier for Glioma, Meningioma and Pituitary brain tumors.

4.2.3.1 Dataset

In the realm of machine learning and artificial intelligence, the significance of datasets cannot be overstated. Datasets serve as the lifeblood that nourishes the algorithms, providing the raw material for training and validation. They represent the foundational reservoir from which models draw insights, patterns, features and knowledge. A well- constructed dataset not only holds the potential to unlock the full power of cutting-edge algorithms but also plays a pivotal role in shaping the accuracy, generalization, and real-world applicability of machine learning solutions. Dataset selection and preprocessing is paramount to ensure the reliability of the models predictions. The advantages of DL are particularly evident when the training data is labeled as a large data set. With smaller data sets, DL performs similarly to other ML methods [SSG18]. With regard to the benefits of large data sets, the question now arises as to when a data volume can be defined as large. However, in addition to the quantity of data, the quality is also of great importance. This is because the sheer volume of data is not yet decisive for good model performance as long as the images are of poor quality and the learning effect therefore remains low. The point at which the quality is considered sufficiently good depends on what is to be learned with the data, but the quality characteristic *reliability* is a measuring factor. The term reliability refers to the degree to which the data can be trusted. A model that has been trained on a reliable data set is more likely to provide useful predictions than a model that has been trained on unreliable data [SSG18] [Dev22]. The answer to the question of how big a data set should be is highly controversial. There is agreement that large datasets perform better than small datasets, but there are differing opinions on size. On the one hand, it is stated that large data sets far exceed regular storage capacities as they enter parameter dimensions of millions. On the other hand, data points of over 1000 can already represent a large amount of data [BII16] [Dor22] [Dev22].

As this work especially focuses on AI in the medical field, more precisely brain tumors, the ResNet50 model is trained on the brain tumor dataset by Jun Cheng [Che17]. This dataset comprises 3064 T1-weighted contrast-enhanced images derived from 233 patients, encompassing three distinct types of brain tumors: meningioma (708 slices), glioma (1426 slices) and pituitary tumor (930 slices). During data preparation, as is customary, the dataset was partitioned into test, validation and training sets. Subsequently, the samples in each set were augmented with artificially generated data. This process, commonly referred to as data augmentation, aims to augment the volume of total data by introducing variations to existing input images, thereby enhancing the model's robustness and generalization capabilities [M20]. Finally, the augmented training set consists a total of 3680 input files. To measure the reliability of the data, the error rate of the labels can be used [Dev22]. Since the data was saved as Matlab files (.mat files), each data point contains the following consistent information structure: label (1. Meningioma, 2. Glioma, 3. Pituitary), patient ID (PID), image data, and tumor border. However, since no invalid labels or omitted or incorrect values are known, the dataset can be assumed to be complete [Che17]. In light of the confusion matrix in figure 4.6, it becomes evident that the ResNet50 model encounters the least difficulty in making predictions for Pituitary tumors, followed by the classification of Glioma tumors. The least accurate results are obtained in the determination of Meningioma tumors. Figure 4.6 further accentuates that images pertaining to the Glioma tumor type represent

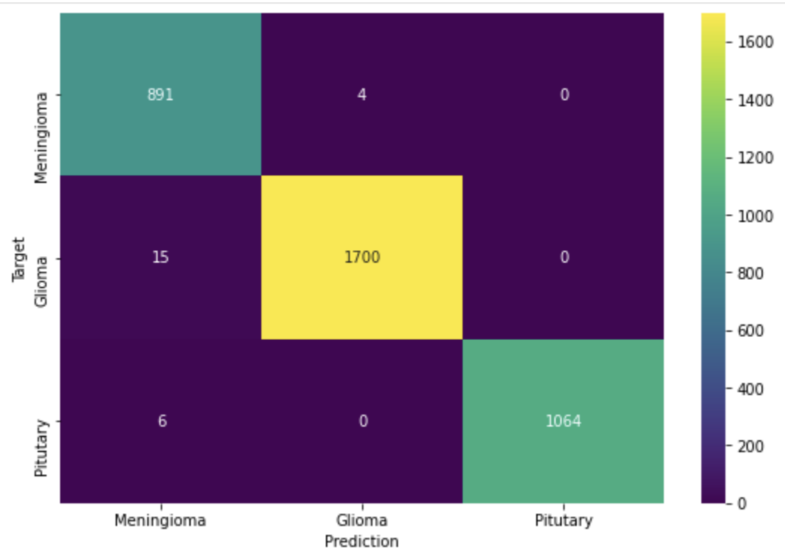


Figure 4.6: ResNet50 Confusion Matrix [M20]

the predominant portion of the dataset, followed by images associated with the Pituitary tumor, while the Meningioma category contributes the smallest proportion. This observation aligns with the dataset distribution reported by Cheng [Che17].

4.2.3.2 Meningioma, Glioma and Pituitary Explained

In order to gain a more comprehensive understanding of the various tumor types included in the dataset, it is necessary to provide a more detailed explanation of each one. The benignancy or malignancy of a tumor in general depends on the behavior and invasiveness of the disease in the affected area in the body. The three different types of brain tumors are depicted from figure 4.7.

The meningioma tumor is known as the most common primary tumor in adults including 35.8% of primary central nervous system (CNS) tumors and over 53% of all benign CNS tumors, with an even higher frequency in autopsy and imaging research. More specifically, diagnosis based on postmortem examinations or imaging techniques register increased occurrence of meningiomas. As the tumor grows slowly, symptoms are usually noticed relatively late when the tumor has already reached a large size. This underscores the importance to detect tumors early in order to start treatment quickly and increase the chances of survival. Benign meningiomas have a favorable prognosis. In general, meningioma tumors record a promising treatment either with surgery, radiation therapy, or a combination of both [Go 23].

> Although most meningiomas can be treated with complete excision, it is often difficult to achieve gross total surgical resection and is life-threatening [Go 23].

The classification based on location is of significant clinical importance. A meningioma occurs when the cells of the arachnoid cap, a protective layer of the brain's membranes, undergo a transformation. Usually, the meningioma starts its growing process on the surface of the brain above the hemispheres known as *convexity meningiomas*, or in the region of a fibrous partition between the hemispheres (*falx meningiomas)*. Meningioma located at the base of the skull are often more challenging to operate due to their location compared to the previous two tumor types. Meningiomas have the potential to extend not only within the meninges of the brain, but also into the bone or even originate from the bone and invade surrounding soft tissues. Up to 20% of patients have been found to not only have single but multiple meningiomas and 10% to 20% grow faster than average or regrow after the treatment [Sim]. Neuropathology distinguishes between different types of meningiomas classified by the World Health Organisation (WHO). The most frequently WHO Grade I meningioma is benign. This slow-growing brain tumor is noncancerous, has distinct borders and is less likely to invade adjacent brain tissue, spread to other locations throughout the body (metastasize) or recur after tumor resection. In contrast, the rarer meningioma type defined as grade II by WHO, is considered aggressive with a high rate of coming back, requiring several surgeries and often leading to a decline in the patient's neurological health. Grade II tumors have characteristics that fall in between benign and malignant behaviors as they exhibit increased tissue and cell abnormalities. Lastly, WHO Grade III meningiomas are regarded as malignant or anaplastic. This type of aggressive brain tumor grows and spreads very quickly and has a tendency to invade the areas of the brain that are in close proximity to the tumor [Mah+22].

The glioma tumor is known as the most common malignant primary brain tumor as it accounts for more than 80% of all harmful brain tumor diagnoses. This tumor type affects roughly 3-5 people per 100 000 persons per year. Life anticipation for people with *glioblastoma* diagnosis, even with standard treatment is typically limited to approximately 12 to 15 months. Gliomas are most likely to impact males, the elderly, Caucasians, and people with particular uncommon hereditary disarranges. Even with the latest developments in therapy, the outlook for glioma patients is still bleak when using modern treatments such as chemotherapy, surgery, radiation, or a combination of these methods. Hence, the discovery of a novel target for the treatment of glioblastomas (GBM) could lead to better progression-free survival and overall survival rates [Tew+21] [Alr+23]. Besides the functional level and the age of affected individuals, tumor size and location depict the major part of the clinical decision-making process prior to surgery. The so-called *astrocytic gliomas*, as the name suggests, have their origin in astrocytes, a type of glial cells in the brain. They are classified in two major groups. Firstly, the astrocytic tumors that invade diffusely into the brain tissue, including diffuse astrocytoma (WHO Grade II), anaplastic astrocytoma (WHO Grade III), and glioblastoma (WHO Grade IV), wherein the grade reflects the aggressiveness of the tumor. Secondly, astrocytic tumors with a more well-defined advancement plan relative to brain tissue, encompassing pilocytic astrocytoma (WHO Grade I) as its essential representative. Tumors within this group essentially happen in children and youthful grown-ups, develop relatively slowly and tend to have a low risk for harmful change (<1%). The first group

(a) Meningioma (b) Glioma (c) Pituitary

Figure 4.7: Brain tumors from the dataset by [Che17], representing each of the three tumor types with exemplary input images[2]

in contrast mostly affects adults, develops more malignant recurrent tumors and as a result records significant worse prognoses [Ton10]. WHO Grade I pilocytic astrocytomas are the most common brain tumors of childhood. About half of the tumors are located in the area of the anterior visual pathway and hypothalamus, and about a third in the cerebellum. Diffuse astrocytomas on the other hand, occur most frequently in young adults and in the white matter of the cerebrum. The WHO Grade II tumor grows slowly and without a sharp boundary to the adjacent brain tissue. Anaplastic astrocytomas (WHO Grade III) can either develop from a preexisting diffuse astrocytoma or arise directly from a diffuse astrocytoma. This tumor manifests signs of anaplasia which refers to fine tissue characteristics of malignancy and leads to rapid tumor growth. Lastly, the glioblastoma (WHO Grade IV) accounts for more than half of all gliomas and is preferentially located in the cerebrum of adults [Ton10].

The pituitary tumor derives from an abnormal growth of cells within the pituitary gland, a small, bean-sized gland located at the base of the brain below the hypothalamus. The pituitary organ regulates a framework of hormones within the body that control development, the digestion system, the push reaction, and capacities of the sex organs by means of the thyroid organ, adrenal organ, ovaries, and testicles. Hence, these tumors provoke a hormonal reversal by either causing too many or too few hormones to be produced. However, most pituitary tumors are benign and therefore non-cancerous, develop gradually and do not necessarily spread to other parts of the body. Pituitary tumors frequently stay undiscovered as the symptoms caused by the tumor resemble those of so many other more common diseases. For this reason, it would be beneficial if the diagnoses were more accessible and easier to carry out as early determination and treatment are key to a great forecast [Hea23]. In contrast to meningioma or glioma, pituitary tumors tend to have a high survival rate, as the probability of a malignant tumor is less than one percent. Despite the generally benign nature of the tumor, the hormonal changes can also lead to severe impairments and have a significant impact on everyday life and quality of life [Coh23].

[2] The tumor borders marked in red color are not part of the original MRI images and have been created based on ground truth data and with the associated jupyter notebook of this work.

4.2.4 ResNet50 Performance Analysis

The evaluation and performance analysis of neural networks is crucial for assessing the efficacy of a ML model, providing insights into its predictive capabilities and potential areas for improvement. Additionally, it is important to establish the model properties in order to relate the final results with the overall model performance.

Table 4.2 shows the classification report for the ResNet50 model predictions. The metrics precision, sensitivity/ recall and the F1-value are indicated for all three tumor types. The support value refers to the images included in the calculations, where the augmented samples are represented in this case. The *macro avg* values give the average of all metrics (precision, recall, F1) for every class of the model. In contrast to the *weighted avg* values, the class frequencies are not taken into account within the *macro avg*. The *weighted avg* adds weight to classes with larger datasets. This is useful in situations where classes are not evenly distributed [Ped+11]. At a glance it is evident, that across all metrics and tumor classes, the

	Precision	Recall	F1	Support
Meningioma [1]	0.98	1.00	0.99	895
Glioma [2]	1.00	0.99	0.99	1715
Pituitary [3]	1.00	0.99	1.00	1070
accuracy			0.99	3680
macro avg	0.99	0.99	0.99	3680
weighted avg	0.99	0.99	0.99	3680

Table 4.2: Classification Report of ResNet50 [M20].

values converge towards 1, indicating a very high model performance. In general, the F1-score is a widely spread and commonly known metric to measure model performance [Fin+23]. The harmonic mean of precision and recall with a range of [0,1], is given as:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{4.4}$$

whereas 1 indicates perfect precision and recall. In this case, the metric states very high values for input images predicting Meningioma, Glioma and Pituitary tumors. In other words, the model makes a small number of false positive (precision) and false negative (recall) predictions. This is an indicator of a well-balanced model that is both precise and comprehensive in its classification [Ped+11]. Lastly, the Jaccard Index is used to measure the similarity between the predicted and actual results. In classification problems, the Jaccard index indicates how well the set predicted by the model matches the actual positive instances. As the score natively applies to binary classes but this model classifies three different tumor types, the index needs to be aggregated to obtain a multiclass problem [Ped+11]. The Jaccard similarity coefficient with a ground truth label set y

and predicted label set $\widehat{y}$, is defined as:

$$J(y, \widehat{y}) = \frac{|y \cap \widehat{y}|}{|y \cup \widehat{y}|}. \tag{4.5}$$

More precise, by computing the index set-wise, it can be extended to multi-label and multi-class averaging [Ped+11]. The Jaccard Index of the ResNet50 model measures 0.99, which underlines a very high model performance [M20]. Overall, it can be stated that both the performance metrics and the test value for equality of two data sets demonstrate a high level of model effectiveness. Predictions made by the model correspond to 99% of the ground truth values, which in turn means that the model identifies the correct tumor type in almost every case.

5 | XAI Tools

In order to meet increasing production volumes and, consequently, a growing user base, the demand to understand what intelligent systems explicitly do and how they arrive at a specific decision has witnessed a significant surge. The request for AI in various fields and, in parallel, the need for explainable visualization techniques have exponentially increased. For this reason, there has been a clear acceleration of the topic in recent times. Particularly in the early 2020s, awareness of the importance of explainable AI heightened [Kra+21]. The increasing complexity in AI-based systems, such as image classification or facial recognition, equally leads to a decreasing interpretability of the algorithms. CNNs often seem like black boxes, efficiently taking inputs and producing outputs with high accuracy, all while keeping their internal mechanisms hidden from view. In order to make these systems explainable, various XAI tools have been established, which differ on the one hand in their representation and on the other hand in the way in which the calculations are carried out. Due to the scope of this thesis, not all existing XAI methods can be implemented and subsequently analysed, thus a selection was made of the three explanatory algorithms that are the most widespread and therefore the most relevant for tumor classification. Van der Velden et al. conducted a study in which research papers were grouped and listed thematically according to the medical diagnosis and the associated explanation tool. The study revealed that the GradCAM, LRP and IG methods were commonly used for image classification of brain tumors based on MRI images [van+22]. For this reason, this thesis relies on three explanation algorithms as the fundamental XAI tools. The following sections will describe the functionality of GradCAM, LRP and IG, and explain how they were implemented in the case study.

5.1 GradCAM

Gradient-weighted Class Activation Mapping (GradCAM) is one of the most widely used and well-known visualization techniques introduced by Selvaraju et al., aimed at highlighting the most relevant features of an input image that have significantly contributed to the output decision. In the case of X-ray images of brain tumors as input, GradCAM would generate a heatmap that highlights the pixels that have contributed the most to the decision regarding the type of tumor present. More precisely, by incorporating gradient information, the activation classes on

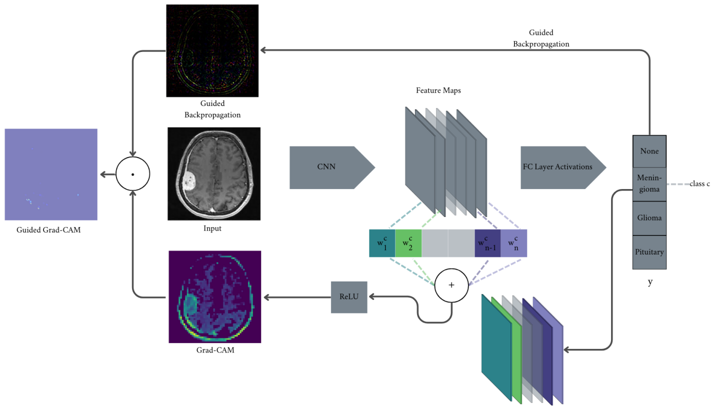


Figure 5.1: GradCAM Overview [1]

the input-received images are depicted. This explains the name of this method [Sel+19].
In detail, GradCAM utilizes the gradients of any target concept, such as Pituitary tumor, which flow into the final convolutional layer to generate a coarse localization map, highlighting the regions in the image that are crucial for predicting the concept. Naturally, convolutional layers provide more insights into spatial features compared to fully-connected layers. Therefore, the final convolutional layers offer the optimal balance between high-level semantics and highly detailed spatial information as neurons in these layers search for semantic class-related information or more precisely, object parts. The approach predominantly adheres to a rather broadly applicable concept and can thus elucidate various layers within a neural network. Nevertheless, the use of the final convolutional layer remains the most prevalent. Hence, the primary focus is clearly on explaining output layer decisions [Sel+19]. The class-discriminative localization map GradCAM $L^c_{GradCAM} \in \mathbb{R}^{\upsilon \times \nu}$ depicted in 5.1 with width υ and height ν for a given class c, is generated by first calculating the gradients of the class score, referred to as y^c with respect to feature map activations A^k of a convolutional layer, culminates in $\frac{\partial y^c}{\partial A^k}$. The gradients backpropagated through global average pooling, as indicated by the indices i and j corresponding to height and width, offer valuable insights into the influence of the neurons. The GradCAM method's procedural sequence starting with the neuron importance weights α^c_k, can be succinctly encapsulated as following:

$$\alpha^c_k = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \qquad \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backpropagation}} \tag{5.1}$$

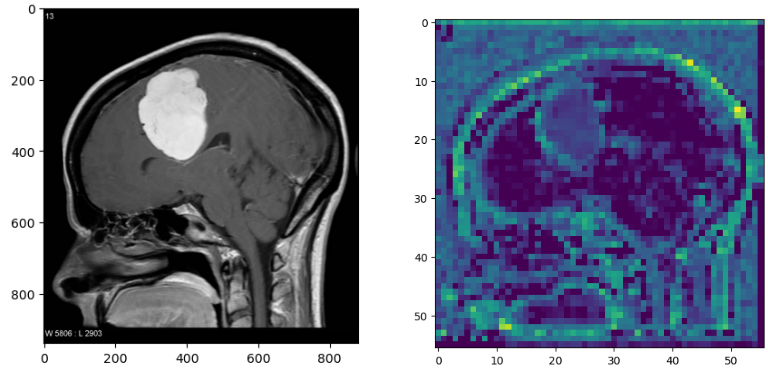
[1]The calculations of the Grad-CAM, Guided Grad-CAM and Guided Backpropagation visualizations are based on the associated jupyter notebook of this work. The input image used is based on [Ket24] and illustrates a meningioma tumor.

In summary, the computation of α_k^c occurs during the backpropagation of gradients, involving a sequential matrix multiplication of weight matrices and these gradients with respect to their activation functions, culminating in the last convolutional layer. Therefore the weight α_k^c portrays a partial linearization of the deep neural network downstream from A, and encapsulates the 'significance' of feature map k for a target class c. As only the selected (target) class holds significance, the focus is on features that exert a positive influence on the intelligent system's decision. To enhance the value of y^c , the pixel intensity associated with the class of interest is emphasized. In contrast, negative pixels, largely belonging to different categories in the input image, do not impact the target class. Hence, localization maps would display additional, irrelevant features without the application of ReLU, consequently resulting in an inferior performance. The calculation can be summarized in the following formula:

$$L^c_{GradCAM} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}} \tag{5.2}$$

This yields a coarse heatmap with dimensions consistent with those of the 7x7 sized convolutional feature map of the last convolutional layer from the modified neural network. In summary, a so-called saliency map encompasses a visualization especially when using GradCAM as the basis for the computations, the map highlights the regions and features which contribute the most to the final model prediction [Sel+19].

The Python case study implemented the GradCAM method using the TorchCAM library, an extension of the PyTorch open-source library, and its *SmoothGradCAMpp* method [Fer20]. Prior to checking the input image for relevant pixels using the explanation algorithm GradCAM, an image transformation process must occur to ensure the image is of the expected size of $3x224x224$. This is followed by an image normalization process using weights from the ImageNet dataset. The weight selection was based on the pre-training of the ResNet50 model using the ImageNet dataset from the PyTorch library [Sub18]. Normalization is crucial to ensure consistent input data during training and inference by normalizing image pixel values using averages and standard deviations (ImageNet weights in this case) [DVV22] [Sub18]. The image preparation steps are applicable to all three explanation algorithms. *SmoothGradCAMpp* uses hooks, which can execute a forward call using a forward hook or trigger a backward phase using a backward hook. Hooks are mechanisms that enable the user-defined code to enter the forward and backward traversal state of a neural network. This allows the monitoring of activations, modification of gradients, or collection of information. A forward hook is called during the propagation of inputs through ResNet50 to generate predictions about the tumor type. Backward hooks are called during the backward run when the gradients are calculated to update the network's parameters [Sub18] [Fer20]. The CAM objects in the TorchCAM library act as a wrapper around the model. By default, the target layer is the



(a) Original Image [Osm23]   (b) GradCAM Calculations

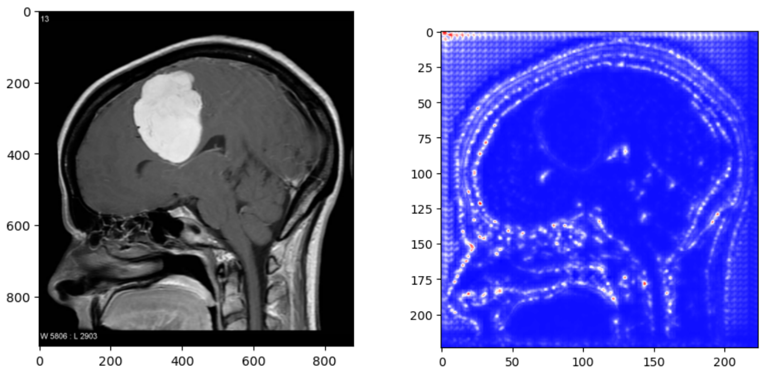
Figure 5.2: GradCAM Saliency Map

The GradCAM saliency map in figure 5.2b is based on layer1 of ResNet50, which is the first layer of the model. Therefore, the visualization tends to reflect the earlier learning processes, which focus in particular on the coarse features such as edge detection [Sel+19]. In comparison to saliency maps that use deeper layers of ResNet50 in the case study, this visualisation highlights significantly more pixels and therefore tendentially better illustrates the algorithm's functionality.

last non-reduced convolutional layer, which was also used in the overall calculation and for evaluating robustness (further details in section 6.1.2) [Fer20]. In contrast, figure 5.2b displays a visualization of the first layer, which represents the most finely porous layer and provides a more informative image in this case. Therefore, the GradCAM saliency map, emphasized in 5.2, visualises the class relevant features for the detection of a Meningioma in the first layer of ResNet50. In order to demonstrate the ResNet50 capabilities, external testing was performed using 5.2a randomly sourced from the internet [Osm23]. This allowed a visualization of the model performance on data outside the original dataset, showcasting its ability to handle diverse input sources. While figure 5.2a shows the original tumor image, which has been classified correctly by ResNet50 as Meningioma with 99.99% confidence, figure 5.2b highlights the relevant features calculated by GradCAM. Figure 5.2b initially suggests certain assumptions that will be verified later in the investigation. It is evident that the GradCAM calculations do not exclusively highlight the features relevant to the correctly identified meningioma tumor, but instead mark a large number of pixels in the image with colour. This phenomenon will be examined in detail later in the thesis, specifically in Chapter 6.1.2 by analysing the robustness of XAI algorithms. At this stage, it can be provisionally stated that the saliency map colour-codes numerous pixels on the basis of the *SmoothGradCAMpp* method. The coloured pixels highlight those features that contributed significantly to the model decision.

5.2 LRP

The Layer-wise Relevance Propagation (LRP) was firtly introduced in year 2015 by Bach et al. [Bac+15]. The LRP algorithm's key element is its use of relevance values. When given an input image x representing the first layer, along with the pixel features of the specific image x and the last layer which constitutes the classifier f and produces a real-valued prediction output, the l layer is modeled as a vector $z = (z_d^{(l)})_{d=1}^{V(l)}$ with the dimenionality $V(l)$. The relevance score R_d^{l+1} for each dimension z_d^{l+1} of the vector z at layer $l+1$ is essential for the LRP calculation as the



(a) Original Image [Osm23] **(b)** LRP Calculations

Figure 5.3: LRP Saliency Map

main goal is to identify the relevance score R_d^l for each dimension z_d^l of the vector z at the next layer l which is effectively closer to the input layer. Therefore, the following formula is given:

$$f(x) = \cdots = \sum_{d \in l+1} R_d^{l+1} = \sum_{d \in l} R_d^l = \cdots = \sum_d R_d^l \tag{5.3}$$

Through the process of iteration, shown in formula 5.3, from the last layer (the classifier output $f(x)$) to the input layer (the input image x with the input pixels) it leads to the following formula:

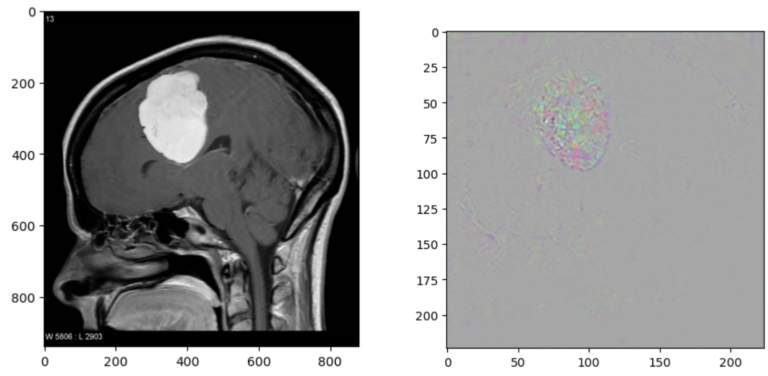
$$f(x) \approx \sum_{d=1}^{V} R_d \tag{5.4}$$

If $R_d > 0$ then there is evidence, that there exists a learning structure for the model. Otherwise, if $R_d < 0$ there is evidence of the absence of responsible pixels (features), resulting in the prediction for a particular outcome. The overall principle is the pixel-wise decomposition where the key is to understand the contribution of a single pixel of an input image x, to the prediction $f(x)$ made by a classifier f. As a result, individually for each input image x the LRP calculates to what extend which image pixel contributes negatively or positively to the classification result. In summary, when presenting a saliency map calculated by the LRP method, the visualization can be used to interpret the relevance score for each single input feature to the model classification result [Bac+15].

After transforming and normalising the input images by default, the LRP method calculated an LRP-type attribution for the ResNet50 using the Zennit framework based on PyTorch [And+21]. This was achieved by combining the *EpsilonGammaBox*, *SequentialMergeBatchNorm* and *Gradient* methods. To generate the rule-based attributions, hooks were registered to perform a backward pass based on the LRP calculations. Additionally, *Composites*, *Canonizers* and *Attributors* were introduced. The *composites* are responsible for correctly applying the appropriate hook to the corresponding layer. There are several *Composites* available that assign specific hooks for certain layers. In this case study, the *EpsilonGammaBox* concept is utilised, which applies particular rules for gradient calculation to convolutional and fully connected layers, thereby defining the hooks for the network layers. The *Canonizers* helper function converts the ResNet50 model into a canonical form to recognize special requirements and structures. The *SequentialMergeBatchNorm* class merges the BatchNorm layers following the linear layers in residual network architectures, simplifying the model. Finally, the *Attributors* helper function applies the gradient-based attribution method to the ResNet50. This function also defines the *Composite* parameter, which significantly influences the hooks [And+21]. In the end, the LRP calculations are presented by a color saliency map in figure 5.3. The figure highlights certain features of edges and corners, but it does not colour-code any pixels in the tumor-specific area. Based on the fundamental assumptions and background of the LRP calculations, it can be inferred that the pixels with the most significant impact on the classification result were not the ones relevant to the tumor, but rather those representing the structure of the human head.

5.3 IG

The Integrated Gradients (IG) is an explanation technique introduced by Sundararajan et al. in 2017 on the basis of two fundamental axioms [STY17]. These two so-called desirable characteristics encompass *Sensitivity* and *Implementation Invariance* and state significant attributes which should be satisfied. Sensitivity refers to the ability of a method to react to different inputs. If an input and baseline differ in a feature and therefore do not have the same prediction, the feature that is different should be assigned a non-zero weighting. If the results of the neural network cannot be traced back to a specific feature, the assignment is always zero. Implementation invariance refers to the situation where two neural networks are functionally identical if they produce the same outputs for all input variables, even if they have different implementations. It is important to always adhere to this principle, as functionally equivalent implementations should have the same associated explanation techniques [STY17]. The sensitivity axiom introduces the baseline, which is a crucial component of the IG method. A baseline is defined as the absence of a feature in an input. The definition may be initially confusing, but it can be clarified by describing the baseline as "input from the input space that produces a neutral prediction" [Erd22]. Therefore, the baseline can serve as an input for generating a counterfactual explanation by observing the model's behaviour as it transitions from the baseline to the original image. Sundararajan et al. exemplary describe the black image as the baseline for visual models and the zero embedding vector for textual models. In both cases, an equivalent value is chosen to indicate when a certain property or variable has no influence on the output. *Integrated gradients* are the cumulated gradients along the straight-line path from the reference input x' to the current input x, determined by the path integral. The formula for integrated gradients along the i^{th} dimension and the gradient $\frac{\partial F(x)}{\partial x_i}$



(a) Original Image [Osm23] **(b)** IG Calculations

Figure 5.4: Integrated Gradients Saliency Map

of $F(x)$ for an input x and a baseline x' is defined as follows:

$$IntegratedGrads_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha. \quad (5.5)$$

When used correctly, the model achieves the principle of completeness. This is because the system's influences and contributions on various pieces of information directly result in the prediction of the program. More precisely, the difference between the output of F at the input x and the baseline $x\prime$ add up to the attributions [STY17]. This can be succinctly encapsulated as following:

$$\sum_{i=1}^{n} IntegratedGrads_i(x) = F(x) - F(x'). \quad (5.6)$$

For most deep learning models it is possible to generate a baseline for which the prediction would be calculated near zero $F(x') \approx 0$. For visual models based on image input, the black image baseline indeed satisfies this property, signifying the absence of objects [STY17]. To sum up, a saliency map calculated by the IG method highlights the relevant features contributing to the model classification output and differ significantly from the baseline. The Python case study implemented the IG using the open-source deep-learning PyTorch Library and the Captum Library, which provides additional functions and methods for model interpretation and analysis. Specifically, the Captum Library includes the attribution-based *IntegratedGradients* method [Pyt23]. The XAI tool expects each input image to be transformed to the size of $3x224x224$ by default. After the transformation, the image is normalized using the ImageNet weights with a transform function before the explanation algorithm applies the calculations [Sub18] [Pyt23]. If a user-specific baseline is not defined, the method uses a default zero scalar for each input tensor [Cap24]. Additionally, the model output is attributed with the target index of the predicted class, which in this case are 0 for None, 1 for Meningioma, 2 for Glioma, and 3 for Pituitary. The method is approximated in 200 steps, as specified. This parameter setting was adopted after the

PyTorch example which successfully uses 200 steps for the interpretation process with IG [Pyt23]. The calculations can finally be visualized using a colour heat map, shown in 5.4. Compared to GradCAM and LRP, the IG algorithm's figure highlights the class-relevant features within the meningioma tumor area. The coloured markings in the figure represent the exact size and shape of the tumor, demonstrating that the model decision was based on the tumor-specific pixels.

6 Analysis and Evaluation of XAI Tools

The multitude of modern explanation tools for unboxing the black box raises several questions: Which tool for which application? How well do the XAI strategies perform or how well do they have to perform to be considered sufficiently explained? Hence, the main challenge when faced with a wide range of options is selecting the one that is most appropriate for the specific problem at hand [Ali+23] [Bor+23]. To illuminate the obscurity, the analysis of XAI tools clarifies the performance and effectiveness of explanatory algorithms in order to enable continuous improvement of AI-driven systems. The subsequent evaluation promotes a clear depiction of the current state of XAI methods and highlights potential areas for development.

Due to the digital transformation and the rapidly changing technical requirements, the European Commission has issued an AI vision that focuses on adaptable, safe, ethical and cutting-edge AI-specialized systems in Europe. In order to create a framework for the implementation and value assessment of intelligent systems, the European Commission formulated so-called Ethics Guidelines for Trustworthy AI. These guidelines provide orientation for the implementation of ethically acceptable, socially norm-compliant and secure modern software. The European Commission represents the executive of the European Union and is therefore an official and legally anchored institution. Due to its significant political influence and its key role in the formulation of legal framework conditions, the direct involvement of the European Commission provides an insight into current developments and political decisions [Art19] [Art20]. In addition to the ethical guidelines, there is an upcoming AI Act that categorizes AI systems and sets out corresponding obligations and regulations. On December 9 2023, the EU Parliament reached a provisional agreement with the Council on the AI Act. When the regulation comes into force, it will be legally binding for all EU member states and will be applicable EU law [Par23]. In contrast, the Ethics Guidelines do not possess legally enforceable impact, but serve as an orientation and classification of systems into trustworthy assistants. Since the focus of this work is particularly on the analysis and evaluation of explanation algorithms in order to find out how explainable and transparent XAI tools developed to date are, the guidelines form a cornerstone for the analysis. The AI Act is taken into account in the subsequent legal classification of the systems in terms of liability in order to clarify the question of what happens in the event of unpredictable behavior or an incorrect diagnosis and to what extent explainability plays a role. In addition, trust was found to be a solid evaluation criterion for XAI tools. In the context of explanatory AI tools, trust is considered as the perception of integrity for their effective deployment and adoption. Recognizing trust as an important metric highlights its central role in evaluating XAI tools [GSM21] [Gri23] [BTR23] [Cha+23].

6.1 Ethics Guidelines for Trustworthy AI

The Ethics Guidelines for Trustworthy AI have been published 2019 by a high-level expert group on artificial intelligence (AI HLEG) set up by the European Commission [Art19]. Based on these Ethics Guidelines, in 2020 the Assessment List for Trustworthy Artificial Intelligence (ALTAI) has been released, to help business and organizations to self-assess the trustworthiness of their AI systems under development [Art20]. The Ethics Guidelines introduced the concept of Trustworthy AI, based on seven key requirements:

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. environmental and societal well-being
7. accountability

The key requirements provide guidance on how trustworthy AI can be realized and which specifications AI-driven systems should meet. Beforehand, it should be clarified what the term trustworthy AI means in this case. The European Commission describes trustworthy AI with three key components: lawful, ethical and robust. Ideally, these components work harmoniously and overlap in their functions. If discrepancies nevertheless arise in practice, the society should endeavour to minimize or reconcile them. The first point on lawfulness refers to the utilization of AI-driven software under current legal requirements and regulations. The second addresses the consideration of ethical values and principles, while the last point refers to the technical and social robustness that is guided by good intentions and does not cause unintended harm. The guidelines are limited to the ethical perspective and robust design and, in this sense, initially neglect the legal requirements for trustworthy AI [Art19]. Therefore, the aspect of legal liability and the examination of the legal situation with regard to the use of artificial systems will then shed light on this point in order to present a complete analysis for trustworthy AI. The European Commission distinguishes between a horizontal and sectorial approach, in other words, a broad, overarching approach that can be applied to AI systems in general and a problem-specific approach. Within these guidelines, the former is considered and could be complemented by a sectorial approach if needed due to the context-specificity of AI-driven software [Art19]. Trustworthiness is the prerequisite to enable "responsible competitiveness" in assuring that the design, development, and use of AI systems adhere to principles of lawfulness, ethics, and robustness. All in all, the guidelines aim to promote sustainable and responsible AI systems in Europe. Therefore, ethical principles constitute the core pillar in the development of a unique approach to AI, that benefits, empowers and protects human prosperity and social well-being [Art19].

6.1.1 Human Agency and Oversight

Intelligent systems are primarily intended to function as support and promote the human ability to make decisions and be autonomous. It is important that control is not vested in the system but rather human oversight is facilitated. AI-based technologies should be developed with a focus on fair and democratic application and respect fundamental rights.

Human agency refers to the overall principle of users autonomy and the associated central role of embedding this in modern systems. It is crucial to provide humans with system-relevant knowledge and tools to satisfactorily understand and interact with AI systems and, where possible, enable them to reasonably self-evaluate or challenge the system.

Human oversight helps to adherence human autonomy in order to provide governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches. While HITL focuses on human intervention in each decision cycle, HOTL limits intervention to the design cycle and monitoring. Finally, the HIC approach incorporates the ability to oversee the overall activity of the system. This includes the ability to set criteria for disregarding the system's decision, to refrain from system decisions in certain situations or to define levels of human discretion during runtime. All in all, it can be stated, that the more testing mechanisms and restrictions are embedded, the less human oversight is required [Art19].

Since LRP performs a layer-by-layer accumulation to filter out the relevance of the neurons, the algorithm works on each layer and users can intervene and make adjustments on each layer [Fin+23]. The IG shows a similar interactive behavior, since the difference of a baseline input to the actual input is calculated, the user has the possibility to adjust the baseline value individually [STY17]. The GradCAM algorithm, on the other hand, refers to the final output layers and therefore offers fewer options for influencing the overall process. In summary, GradCAM provides less granular information by focusing on the final layers and generating a coarse locality map, resulting in less possible intervention. All in all, IG and LRP promote system interaction more than the GradCAM algorithm [Bac+15] [STY17] [ABW22].

The HITL approach involves the continuous integration of human users into the learning process of the system. Depending on the controlling instance of the learning process, different interaction levels can be identified to HITL. In Active Learning (AL), the system treats the user as an oracle, e.g. to correctly label data that is unclear or would make a major contribution to the learning process. The algorithms offer the possibility of using HITL, since they provide visual representations to show which features have been particularly relevant for the prediction. Therefore, based on the input features utilized by the model, data can be labeled and enhanced, contributing to the overall adaptation of the model [Mos+23] [Vyh+23]. In terms of the HOTL approach, all three algorithms can help identify anomalies or unexpected predictions where human intervention would be necessary. In other words, the approach emphasizes human monitoring and intervention when needed, which IG, GradCAM and LRP enable [Vyh+23]. Finally, the HIC approach refers to the user's complete control over the system and the functioning of the AI software as support for the decision-making process. The three algorithms provide insights into how the model works and how it runs, helping users to make informed decisions [Vyh+23].

Overall, it can be stated that the explanation algorithms differ in particular with regard to flexible system intervention. LRP and IG offer more options for parameter settings during runtime and thus enable simplified influence on the visual output. The GradCAM algorithm, on the other hand, has fewer parameter settings but offers the option of layer selection [Sub18] [Sel+19]. In addition, the algorithm works on the basis of hooks, essential callback functions that capture the characteristics of different layers. The way the hooks are generated can basically be controlled by human interaction with the system [Sub18] [ABW22]. In terms of human agency and oversight, the characteristics of the explanation algorithms tend to contribute to users being able to intervene increasingly with the system and thus gain more control.

6.1.2 Technical Robustness and Safety

Technical Robustness and Safety is subdivided into a multitude of components including resilience to attack and security, fallback plans and general safety, accuracy, and reliability and reproducibility [Art19]. The first two points relate in particular to the fact that the systems are protected against external attacks or have mechanisms that counteract hackers and mitigate vulnerabilities. In general, these algorithms are vulnerable and can be manipulated, because wherever inputs can be made manually and are not standardized, they can also be arbitrarily adapted, extended or changed. Attackers could undertake evasion, poisoning, data protection or misuse attacks and thus influence the behavior and decision of the intelligent systems [Gül23] [Bou24]. Nevertheless, it must be emphasized that the relationship between AI and cybersecurity is closely interlaced, as AI frameworks are progressively being utilized to assist distinguish, anticipate and react to cyber dangers [Gül23]. In particular, XAI algorithms contribute to the safety of AI-driven systems as they increase transparency and traceability. As a result, anomalies, weak points and error developments can be traced more quickly and easily, the causes of unexpected system behavior can be identified and manipulative influences can be reduced. Overall, this creates greater system security [Sri+22]

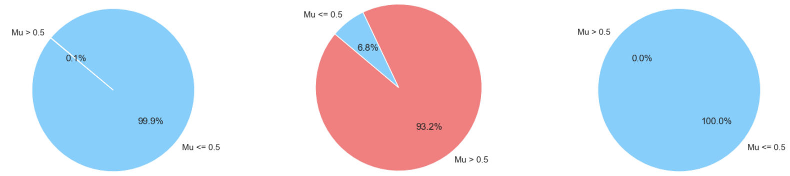
[Gül23]. Conclusively, it can be stated that although XAI algorithms themselves are vulnerable and may have technical gaps, they still contribute to improving the overall security of artificial systems by creating a basis for tracing vulnerabilities and subsequently correcting errors [Sri+22]. A fallback plan also includes the legal challenge of an AI decision. It is therefore important that the explanation algorithms follow guidelines on legal liability so that the legal framework is regulated in the event of a wrong decision [EV24].

The remaining components, accuracy as well as reliability and reproducibility in particular emphasize robust development and evaluation of AI systems to mitigate potential risks and support effective functionality [Art19]. Overall, the ability to make accurate decisions, avoid undesirable outcomes and perform reliably under different conditions helps an AI system to be more resilient to challenges and unexpected situations. Robust AI systems can maintain performance and operate efficiently even in the face of unexpected or variable data. Therefore, accuracy, reliability and reproducibility are key factors in the development of robust AI systems [Art19]. To point out, robust intelligent systems must not only make accurate predictions, but also work reliably in different environments and in the presence of different inputs.

In order to evaluate the robustness of visual explanations more precisely, quantitative methods have demonstrated their efficacy [Fin+23]. Based on ground truth labels and domain specific knowledge, polygons that form the tumor boundaries have been generated individually for each input image within the associated case study. The knowledge about the tumor coordinates allows the specific location, size and shape of the tumor to be determined and demarcated. This made it possible to calculate the number of relevance values in the overall image and to quantify the representation of relevant features within the polygon which illustrates the tumor borders. The results provide final information on how many class-relevant tumor regions were included in the area marked as relevant by the respective algorithm. Given the ratio μ, the following formula calculates the amount of positive relevance inside the polygon or boundary (R_{in}) and the positive overall relevance in the image (R_{tot}):

$$\mu = \frac{R_{in}}{R_{tot}}. \tag{6.1}$$

Ranging from value 0 (no positive relevance inside the tumor border) to 1 (all positive relevance inside the tumor border) the $\mu - value$ indicates how much the classification output is based on class-relevant parts of the image [Fin+23]. If the μ-value is above 0.5, the majority of relevance aggregates inside the tumor borders. Assuming that the diagnosis of a meningioma is correctly predicted by the model, a value above 0.5 would indicate, that the model tends to decide based on the tumor-specific features. The μ-value has been calculated for all three explanation algorithms on the entire dataset of 3064 T1-weighted contrast-inhanced input images. The results are encapsulated in figure 6.1. Of particular significance is the μ-value distribution of 6.1b, since the explanation based on integrated gradients shows in 93.2% of the total input images that the majority of

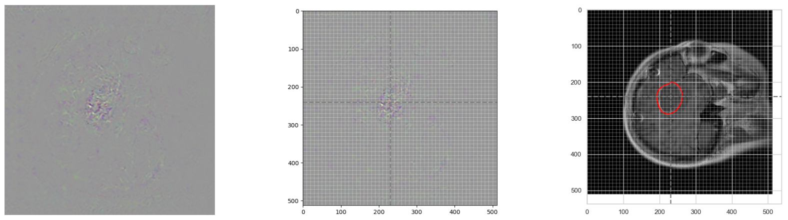


(a) GradCAM **(b)** IG **(c)** LRP

Figure 6.1: μ-value Distribution of Explanation Algorithms

relevance aggregates inside the tumor borders. In other words, on almost every input image, the IG algorithm identifies the decisive characteristics for the tumor prognosis within the defined tumor boundaries and thus directly on the specific tumor. In contrast, both the GradCAM and the LRP algorithm do not present visualizations in the case study in which the relevance values are above the average in the range of the tumor coordinates. This means that the algorithms do not attribute the decision of the ResNet50 model to the tumor features. Figure 6.2 addresses the class-relevant visual representation of the IG explanation algorithm and provides the original tumor image to compare the relevance values with the actual tumor, which is marked and identified by the tumor coordinates. Incontestably, the outcomes underscore that the relevance values in 6.2a and 6.2b derived from the IG algorithm reflect the tumor in 6.2c, thus the positive relevance aggregates inside the tumor borders, resulting in a visualization that clearly attributes the model decision to the tumor-specific features.

In summary, although the ResNet50 model displays high F1 scores for all three tumor classes (Meningioma, Glioma, Pituitary), as shown in Chapter 4.2.4, the relevance is not in the expected areas for the GradCAM and the LRP algorithm because most of their μ-values indicate no positive relevance inside the tumor borders. The analysis reveals that GradCAM and LRP perform suboptimally when assessing the robustness of algorithms in the Python case study. Specifically, these explanatory algorithms do not produce accurate and reliable visualizations that represent the tumor-specific class-relevant features. The visual explanations based on IG, on the other hand, demonstrate a high degree of robustness, thereby signaling their accuracy, reliability as well as reproducibility and enable the robust development of AI-driven systems. The limitation of the robustness calculation based on the μ-values is that it does not consider normalization of the values according to the size of the regions. This is important because different sized tumor regions exist for each input image and some tumor boundaries may have a more stringent effect on the relevance distribution, which affects the model performance. Therefore, for future evaluation it is suggested that a weighted value calculation is implemented that takes into account the overall distribution of relevance in the image through thresholding [Fin+23].



(a) Simplified IG Output

(b) Enhanced IG Output with Coordinates and Tumor Borders

(c) Original Tumor Image with Tumor Borders and Coordinates

Figure 6.2: Comprehensive illustration of the Integrated Gradients (IG) analysis process, showcasing different levels of output: simplified IG calculations (a), output with coordinates and tumor borders (b), and the original tumor image [Che17] with borders and coordinates (c).

6.1.3 Privacy and Data Governance

The prevention of harm and the fundamental right to privacy are closely linked. To ensure privacy, data must be appropriately managed based on quality and integrity, and processed only in a way that respects privacy. This includes ensuring data protection and security throughout the entire life cycle of a system, as well as maintaining data quality and integrity. Data protection and security include all information provided by the user, as well as any user information and data collected during the system's lifetime. This is especially important for minimizing inaccuracies and errors and improving the performance of AI systems. When dealing with self-learning systems, it is crucial to ensure the integrity of the data to prevent undesirable behavior. Finally, it is crucial to implement security protocols for data access to prevent unauthorised access [Art19].

GradCAM, LRP and IG are explanation algorithms and no closed AI systems, thus indicating an evolving nature. When categorizing privacy and data governance, individual explanation algorithms are of limited help as they only represent algorithmic components in the overall system. Therefore, the ResNet50 model and the system as a whole are increasingly used to address this problem. The ResNet50 model was trained using publicly accessible data and as a result, no special mechanisms were implemented to protect user privacy and sensitive data security. However, if sensitive user information is collected, it is necessary to meet special data protection requirements. Moreover, security protocols should be implemented to prevent unauthorized access and ensure the unrestricted integrity and security of user data [BP22]. Security is a significant concern when using AI systems in healthcare. It is crucial to maintain the patients anonymity and protect personal health information. Therefore it is essential to establish clear guidelines to mitigate malicious data leaks, attacks and to maintain patient trust and privacy in AI systems [GJW24]. The importance of secure data storage is exemplified by the recent hacking of over 30,000 treatment files from a psychotherapy centre in Finland. As a result, patients were blackmailed with demands for payment and their files are now publicly accessible [Sch21]. This breach highlights the need for highly secure data storage to protect sensitive information. Employers could potentially gain an unfair advantage by viewing the medical data of candidates. Similarly, curious citizens could search the database for prominent individuals and check for possible mental health conditions. It is crucial to continuously address these challenges and contradictions and enhance AI systems to the point where such significant data breaches are no longer possible [Sch21].

6.1.4 Transparency

Transparency is essential for an AI system to be trustworthy and aligns with the principle of explicability. It refers to providing clear and understandable access to the data, system and business models. This requirement pertains not only to technical aspects (data and system) but also to economic aspects that should clarify how the system operates and what values it produces to achieve its goal [Art19]. The focus is on three key components: traceability, explainability and communication. The initial component pertains to thoroughly documenting the data sets and processes, including those of data gathering and data labelling that inform the modelling decision to ensure traceability. This is crucial for identifying the causes of incorrect AI decisions and preventing future errors. The second component of explainability refers to comprehending the technical processes and potential human decisions. It is possible for explainability to compete with accuracy, requiring trade-offs between improving explainability and achieving highly accurate results. It is important to provide adequate explanation and disclosure of the decision-making process, particularly for AI systems that have a significant impact on human life. This ensures objectivity and transparency in the decision-making process. Finally, in terms of communication, it is important to respect fundamental rights and the freedom of individuals to interact with an artificial system. Therefore, it is crucial to clearly label AI systems as such and explicitly communicate their limitations, as they do not involve human communication. As a solution, AI-driven technologies could provide output criteria that include the accuracy level and limitations of the decision-making process, to ensure that end users are aware of the system's capabilities [Art19].

Transparent, comprehensible and understandable access to the system's decision-making process is inherent to the nature of explanation algorithms, regardless of their specific mode of operation and application. Therefore, it could be assumed that the principle of transparency is covered by all three explanation algorithms. The components accuracy and explainability often contradict each other because complex models tend to be simplified to make them explainable. However, these explanations may not reflect the complex interactions in the deeper layers of the model, leading to a degradation of accuracy [Art19] [Pet22]. The case study presented in this thesis illustrates the inherent dilemma associated with balancing explainability and accuracy in AI systems for the medical field. The tumor diagnosis is highly accurate, but the visual explanations do not yield the desired results. The technical robustness evaluations indicate that the relevant features contributing to the decision-making process are only found within the tumor boundaries in rare cases (see 6.1.2). The accuracy of the model is excellent, however, the visual explanations do not emphasize the expected and medically relevant areas. This makes it difficult to explain the prediction, which can have serious consequences as both healthcare professionals and patients require explanations to make informed decisions [GSM21] [Bor+23]. The IG method has been successful in providing robust visual representations and highly accurate results. Conversely, the other two explanatory algorithms, GradCAM and LRP, depict the conflict between explainability and accuracy, as the robustness values are in contrast to the accuracy level of the model in the case study. Moreover, the model communicates the accuracy of disease diagnoses by indicating the probability with which a particular tumor was identified. Users can evaluate the model's capabilities by classifying the decision's confidence based on the probability. However, future XAI tools could improve by providing not only the model prediction with a probability but also the individual explanation algorithms. This would indicate how likely the respective visualizations are to be displayed correctly [Art19] [Art20].

In general, it can be concluded that the three explanation algorithms apply to the key components (traceability, explainability, communication) of the transparency criterion to varying degrees. The IG method is both traceable and explainable according to the categorical definition, as the visual results contribute significantly to understanding the modelling decision. The GradCAM and LRP algorithms provide traceability, as data is labelled to ensure the traceability of the system process, similar to the IG. However, the aspect of explainability is still questionable for GradCAM and LRP, as their visualizations have not yet produced robust results and do not provide the expected representation of relevant features inside the tumor boundaries. It is assumed that the model's accuracy is compromised due to its complexity, and that a simplified explanation of the modelling decision creates a conflict [Pet22] [S B+23]. However, it is important to note that just because the explanatory visualizations make sense to humans, does not necessarily mean that they convey what is actually happening within the model. Evaluating and assessing visualization methods to determine their validity in the model is still a significant challenge. As a result, researchers are actively seeking improved benchmarks that incorporate human input to draw more reliable conclusions [Fel+23]. The robustness score is a useful initial indicator of the model's response to changes and its average performance [S B+23] [Fin+23]. However, additional performance evaluation methods are necessary to determine the extent to which GradCAM and LRP fail to provide adequate explainability. It cannot be concluded that there is a problem with explainability simply because the expected results are not emphasized. Nonetheless, there is reason to suspect that this may be the case [Pet22] [Fel+23] [Fin+23]. In terms of the communication aspect, the overall system provides a transparent representation of its capabilities. However, the individual explanation algorithms do not offer the option to display the probability of the correctness of the visualization strategy.

6.1.5 Diversity, Non-discrimination and Fairness

The principle of fairness enlights the important role of non-biased and non-discriminative AI system's lifecycle. To maintain trustworthy AI systems, data collection, processing, and use must not be based on the social demographic characteristics of users. Instead, integrity, diversity, and fairness should be fundamental values throughout the entire system. Inclusive access to system processes and equal and fair treatment of data are crucial. In order to avoid biased data it is relevant, to detect and control historical discrimination, proxy variables, unbalanced datasets or algorithm choices leading to unfair results [Art19].

AI in medicine is particularly susceptible to challenges referring to ethics and inequality as biased training data and decision-making processes can lead to disparities in healthcare outcomes. Underrepresented patient subgroups and shortcoming of healthcare data typically result in unbalanced and biased data, creating high potential risks for incorrect medical assumptions and results [GJW24]. In a clinical context, algorithmic discrimination can have a significant impact on the under- or over-treatment of patients or provoke serious treatment errors. The problem is exacerbated by the black-box nature of DL systems, as decisions cannot be transparently traced, weightings are not scrutinised and it is therefore difficult to detect errors and biases [Sch21].

All three explanatory algorithms were trained on the same standardised medical data. In total there are 3064 different tumor images, including 708 meningioma samples, 1426 glioma samples and 930 pituitary samples [Che17] [M20]. The input data forms the fundamental learning base of the algorithms and thus the final disease types that the model can identify. The patient as an individual is not considered as a case-by-case decision by DL systems, but in the context of a stereotyped group element, which can lead to incorrect conclusions due to generalization. In other words, if there is an unequal gender distribution in the tumor dataset, for example, gender-specific health conditions may be missed or misdiagnosed. AI systems developed to detect skin cancer can be highly sensitive to differences in skin colour, predicting a completely different outcome for darker skin types with the same disease [Sch21]. The aim is therefore to generate a highly diverse distribution of socially relevant groups in terms of age, gender, ethnicity or social background. The [Che17] dataset does not provide detailed survey information to verify the diversity of the data. However, it can be critically noted that the disease profiles (meningioma, glioma, pituitary) are represented in varying quantities. The model could therefore be biased, as it has reviewed and learned a higher number of glioma tumors than the remaining two tumor groups. It should also be emphasized that the ResNet50 model was trained by [M20] on four classes (None, Meningioma, Glioma, Pituitary), but the unassigned class *none* was not considered in the dataset or training process and therefore represents an unbalanced residual class. It is common for data sets to have varying numbers of data points for each class. However, if the difference becomes too large, it can significantly impact the preferential treatment of majority classes over minority classes. Overall, this can result in a skewed class-by-class modelling performance [Web+23]. In summary, the *none* class presents a significant challenge as it lacks specific training and cannot be traced. The explanation algorithms are unable to attribute data samples belonging to the *none* class to relevant features, as these have not been specified and are therefore unavailable.

Another issue is the absence of information about the labelling process for the tumor data [Che17] [Web+23]. Depending on the data domain, experts should either manually identify and label the data or automatic data annotation should be used. As [Che17] does not provide any further details regarding data collection, this aspect should be considered in future work as it has a significant impact on data quality and model performance. In the worst-case scenario, the three explanatory algorithms may fail to produce accurate visualizations that attribute the model's decision to the learned features of the brain tumor due to incorrect labelling of the data [Web+23]. However, since this is a widely known dataset, it is not assumed [Reh+20] [KG20] [AZ21] [Haq+23].

6.1.6 Environmental and Societal well-being

In the technical field, it is also important to consider environmental impacts and use sustainable resources. The goal is to enrich society not only for the current generation but also for future generations. Therefore, modern technologies should be implemented with the future in mind, being environmentally friendly, sustainable and socially impactful [Art19]. AI-driven solutions should address areas of global concern, such as for instance the Sustainable Development Goals. Overall, the focus is to create beneficial AI solutions which also take the circumstances of the current environmental situation into account and do not blindly connect state-of-the-art systems that would have a lasting negative impact on the environment [Art19].

At the macro level, explanatory algorithms initially improve environmental and sustainability aspects by systemizing systems more efficiently, reducing computing power usage. Image-based models generally consume more energy and computational power than models that analyse numerical input. Furthermore, power consumption is heavily dependent on the hardware and software used, as well as the parameters and other adjustable variables. Image processing models typically rely on computationally intensive and powerful technical systems, resulting in high energy requirements. Additionally, the volume and complexity of data also significantly impact power consumption [Jea+23]. As this work was implemented on a small scale, no large server rooms with high computing power were used. Additionally, the original model from [M20] was reused and carried out on an energy-efficient system [Cla22] [KM23]. Therefore, it can be assumed that the overall system is tendentious environmentally friendly. It should be added that the positive social impact of XAI tools and the associated benefits and gains for society are undisputed [SM19] [Gri23] [BTR23] [Cha+23].

From a microeconomic perspective, it can be concluded that the computing and energy requirements for individual explanation algorithms depend heavily on their implementation. In comparison, the GradCAM algorithm offers a relatively high resolution of visual explanations and is known for its efficient computing power, requiring only one forward pass and occasionally one backward pass per input image [Sel+19]. IG and LRP are more computationally intensive and require more time per run [Mar+15] [Mon+19]. The IG has established itself due to its broad and efficient applicability, as it can be applied to different AI models and offers ease of implementation, theoretical justifications, and computational efficiency [Mar+15]. Overall, there is limited research on the environmental impact of the three explanation algorithms. Future studies could analyse this in more detail to determine the extent to which the algorithms can be classified as sustainable and performance-efficient. For now, it is initially noted that the explanation algorithms contribute to a large-scale increase in efficiency and have a significantly positive influence on social well-being, due to their enrichment in terms of traceability and transparency.

6.1.7 Accountability

The requirement for accountability pertains to ensuring responsibility and accountability for the results, both before and after the implementation of AI systems. This involves making the system process assessable, particularly for applications that affect fundamental rights, which should be subject to external and independent review. Reporting on systematic impacts should be enabled and encouraged as a matter of principle to strengthen trust, allow for legitimate concerns to be reported and create an overall system assessment to ultimately minimize negative impacts. Trade-offs, which refer to decisions made as a compromise when certain objectives or principles conflict, must be properly justified and documented based on current technology. Additionally, the principle of redress should be applied to counteract any unjustified negative impacts and provide mechanisms for redress, particularly for disadvantaged groups [Art19].

The focus is on ensuring traceability, verifiability, and assessability of AI systems to enable transparent reporting, auditability, and legal documentation [Art19]. To evaluate the implementation and overall assessment of the ResNet50 model, an external user analysis and survey identified potential weaknesses or risks of the AI system (further details are discussed in section 7). Furthermore, this analysis and evaluation of the XAI tools also documents and identifies conflicts between the seven categories of the Ethics Guidelines for Trustworthy AI. Therefore, it can be concluded that transparent documentation is available and the ethical principles and explanations of the model's decisions have been taken into account and presented. However, it is recommended to provide explicit training for third parties, such as end-users, healthcare professionals, or test subjects. Additionally, establishing an AI committee for the independent assessment of potential ethical grey areas or the general accountability of the model is suggested [Art20]. To ensure reliability across diverse scenarios, it would be beneficial to organize an external group of experts to evaluate and categorize the implementation of ethical principles in future work in this area [Art19] [Art20]. In conclusion, explanation algorithms have a positive influence on achieving accountability criteria by improving transparency and traceability, enabling documentability. However, there is potential for improvement in the implementation of external consulting, professional training in the use of the specific AI system and the establishment of a risk management process.

6.2 Legal Liability

In the context of the discussion on AI, the question of the prerequisites and requirements of the legal framework inevitably arises, due to the complexity and far-reaching effects of advanced applications of AI-driven systems. The increased use of algorithmic models, particularly in the field of machine learning, has led to an acceleration of the political debate [Hac+20] [Ric+23]. This discussion focuses specifically on the legitimacy and accountability of the advanced application of machine learning. The importance of the topic is further emphasized by the proliferation of artificial intelligence technologies and their impact on different sectors and areas of society. In this context, it is necessary to clarify legal terms and requirements to ensure proper regulation and legal liability.

> Algorithmic models are increasingly emerging as the subject of an intense policy debate surrounding the legality of, and liability for, advanced machine learning applications [Hac+20].

In jurisprudence, the focus is particularly on the justifiability of a particular decision, so that it is not based on arbitrariness, but on general socially recognized guidelines anchored in law. As the law is a large, extensive topic, it is necessary to narrow down which legal aspects are of primary importance when using and implementing artificial systems. In the field of machine learning besides to data protection and the principle of equal treatment, the question of responsibility and liability arises in particular, as it is important to know what happens in the event of a malfunction or wrong decision [ČGS15] [Hac+20]. The legal system is not only diverse in itself, but also contains significant differences depending on the country and state system, which is why this work is limited to European law in relation to the handling of artificial intelligence. The European Commission's proposal to enforce an AI Act provides a strong foundation for discussing the current legal basis and liability criteria related to AI systems [Com22b]. The Ethics Guidelines of Trustworthy AI have already provided information on the ethical categorization of the ResNet50 and the explanation algorithms. It is important to additionally consider the legal implications and requirements when evaluating the effectiveness of current XAI tools. This includes addressing the potential consequences of a wrong decision made by the system, even when explanatory tools have been used.

The AI Act was created specifically to address the fact that 43% of companies intending to use AI were hesitant to do so due to the absence of a legal framework governing liability [Com22b]. Due to the rapid advancement of technology and the increasing use of autonomous systems, individuals who have been harmed by an artificial error of judgement face significant challenges. This is due to the lack of clear regulations regarding the use of AI systems to claim damages or seek redress for harm caused by a systematically incorrect decision. The misdiagnosis may be attributed to a lack of burden of proof, as the system's high complexity and low traceability make it nearly impossible to comprehensively prove that the system was the cause [Com21] [Com22b].

The AI Act adopts a risk-based approach that classifies AI systems into four risk levels: unacceptable risk, high risk, limited risk, and minimal risk.
Unacceptable risk. All AI systems that pose a threat to general safety or society are prohibited without exception. This includes for example the production of children's toys that encourage dangerous behaviour.
High risk. These systems are subject to strict obligations and requirements as they are used in sensitive situations. This includes applications in healthcare, product safety, critical infrastructure and education. The risk level encompasses areas of people's lives crucial to their personal development or health. In other words, systems that could have an immediate and direct impact on human life. As high-risk systems categorize applications in the medical field and systems that impact human health, ResNet50 and its associated explanation methods are also considered high-risk systems.
Limited risk. AI systems that are required to meet transparency obligations pose a limited risk. For instance, chatbots must clearly indicate that the interaction is with an artificial system and not a human.
Minimal risk. Most AI-based applications currently available fall into this category, such as AI-enabled video games or spam filters. All of these items share the characteristic that a malfunction does not result in significant harm or negative consequences [Com22b] [Com24].

As the tumor diagnosis system implemented in this work is classified as high-risk, the issue of liability regarding high-risk technologies is of utmost importance. The ResNet50 model and its associated implementations aim to detect brain tumors in patients and subsequently visualize the model's behaviour and decision-making process. It is considered a high-risk system according to Article 6(2) and the systems listed in Annex III, particularly Section 5(c) of the AI Act. While for non-high-risk AI systems, only very limited transparency obligations are imposed, for high-risk AI systems, it is strictly necessary to meet requirements for high-quality data, documentation, traceability, transparency, human oversight, accuracy and robustness to mitigate the risks to fundamental rights and safety. Compliance with legal requirements for information, documentation, and logging of high-risk systems is crucial for determining claims for damages and disclosing evidence in case of a claim. Article 13 of the AI Act defines the transparency requirements and states that meeting these requirements is a prerequisite for implementation, legal judgment, and evidence. In the event of non-compliance with this duty of care, the owners of the AI system can be held accountable if this can be proven by the plaintiff [Article 4(2), AI Liability Directive]. Similarly, end users must also demonstrate responsible use of AI-based systems and ensure that they follow the instructions for using the system [Article 4(3), AI Liability Directive].

The legislative proposal of the AI Liability Directive is based on a tiered approach and the first tier is based on the principle of rebuttable presumption [Com22b]. Rebuttable presumptions are common in national liability regulations and aim to balance the interests of plaintiffs and defendants, as well as ensure that all parties fulfil their duty of care. The second stage of the draft law considers the incorporation of future technological, regulatory, and legal developments. The examination will determine if there is a need to harmonise other elements of compensation claims or liability claims instruments [Com22b]. More specifically, the rebuttable presumption addresses the principle of causality. It ensures that persons claiming damages when an AI system was involved enjoy the same level of protection as in cases where no AI systems were used. This is because it can be problematic for injured parties seeking compensation to prove that fault is attributable to a causal chain of systematic malfunctions. For this reason, the rebuttable presumption applies if it can be proven in advance that, firstly, damage has occurred and, secondly, there is a presumption of a causal connection [Article 4(1), AI Liability Directive]. Regarding situations where non-compliance with the duty of care or legal requirements can be proven, or where it can be shown that the user did not use the system in accordance with regulations, a rebuttable presumption may be drawn that there is a causal link between the proven error and the resulting damage. The accused party retains the right to defend against this presumption and prove otherwise [Com22b].

In the case of a misdiagnosis and resulting patient harm with the ResNet50 tumor diagnosis system, it is necessary to first clarify whether all high-risk system requirements under Chapter 2 of the AI Act, Articles 8 to 15 and safety measures were fulfilled and whether the owner of the AI system lawfully observed their duty of care. Subsequently, it must be determined whether the user used the system lawfully and in accordance with the instructions. If a patient has suffered harm due to an incorrect model prediction, the plaintiff must prove that a breach of the duty of care has occurred, or that there has been non-compliance with the legal requirements. Alternatively, it must be proven that the user of the model did not act in accordance with the system instructions. Only then will the rebuttable presumption apply. Therefore, if an error that has been previously proven to cause damage is identified, it is assumed to be the cause of the misdiagnosis. Ultimately, the responsible party, whether an individual or an organization, will be held liable for failing to meet their duty of care and legal obligations, or for incorrect use or non-compliance with system instructions [Com22b]. The explainability of systems is crucial, as it is a fundamental requirement for high-risk systems and is also essential for the legal assessment of a claim [Com21] [Com22b].

6.3 Discussion

A preliminary assessment is possible due to the step-by-step comparison of the explainable algorithms and review of the AI-driven system to determine whether it meets ethical requirements for trustworthy AI and to what extent it can be legally classified.

The analysis of XAI tools based on the seven characteristics for trustworthy AI has revealed the extent to which explanation algorithms, specifically GradCAM, LRP, and IG, currently cover the criteria and where there is potential for improvement. The criteria are fulfilled to varying degrees and in different ways by the three explanation algorithms. The initial criterion, human agency and oversight, is addressed by all three algorithms, but LRP and IG perform better as they allow for more flexibility in parameter settings during system runtime. Regarding the second criterion, technical robustness is currently only completely addressed by the IG algorithm, while the other two explanation algorithms still have room for improvement in this area. Furthermore, it is imperative to enhance the security measures of all three algorithms and the entire system to safeguard against external attacks and manipulation. Privacy and data governance were afforded limited consideration, as publicly accessible data was utilized and the system is still in its developmental stage. All explanation algorithms primarily fulfil the principle of transparency by providing a clear representation of the model decision, thus counteracting the black box phenomenon. However, differences in the actual explainability of the algorithms can still be identified. Accordingly, the IG algorithm provides visualizations that accurately represent relevant characteristics within the class-related tumor boundaries in this specific application. However, diversity, non-discrimination, and fairness present challenges due to the lack of information about the data collection process. It is important to note that potential bias in specific classes is a significant concern in this context. Therefore, it is uncertain whether this criterion is fulfilled. While explanatory algorithms typically enhance efficiency and social well-being in terms of environmental and social well-being, it is still unclear to what extent the algorithms themselves promote environmental sustainability. Finally, the explanation algorithms have a positive effect on accountability, but are not solely responsible for meeting the criterion. To prepare for the future, AI system users should receive specific training, risk management should be integrated and external committees should be established to provide expert advice.

Overall, the explanation algorithms require improvement as they do not yet meet all the criteria for trustworthy AI. In the context of this application, the IG method is initially preferred as it achieves the most accurate results based on ethical guidelines. Therefore, the efficiency of current XAI tools and their ability to make models transparent and explainable depends on their context-specific application. However, there is still room for improvement, particularly in ensuring that the explanation algorithms provide accurate explanations that are not only comprehensible to humans but correctly precise.

After comparing the current legal situation, two main liability scenarios can be identified in relation to the implemented system:

- If the injured party can prove that the owners of the AI system failed to fulfil their duty of care in accordance with the legal requirements of the AI Liability Directive, and that this failure is also clearly attributable to the misdiagnosis, then a claim can be made. In relation to this case study, it would mean that a misdiagnosis of a brain tumor has occurred and it can be proven that the error is due to the lack of, for example, system robustness which directly caused the claim. In this case, the party responsible for non-compliance with the necessary requirements in accordance with the AI Liability Directive is liable [Article 4(1) and 4(2), AI Liability Directive].
- If the owners of an AI system can prove that any damage caused is due to the fact that the users did not follow the correct system instructions or did not use the system lawfully, the rebuttable presumption also applies. It is assumed that this misuse constitutes a causal link for the damage. In relation to this case study, this would mean that an incorrect tumor diagnosis was made and the AI owners can prove that the end users used the ResNet50 incorrectly. For instance, if the system instructions require software updates within a specific timeframe to ensure accurate diagnosis, and the end users fail to comply, the party responsible for the system updates will be held liable for any misdiagnosis that occurs due to the lack of updates, provided that all legal requirements and due diligence have been met [Article 4(1) and 4(3), AI Liability Directive].

When discussing the role of XAI tools in the event of a damage claim and associated liability, it is clear that the use of explanatory methods has a significant impact. This is because explanatory mechanisms are necessary for transparent and comprehensible decision-making, and their absence can affect legal liability criteria.

7 Evaluation of Implementation

As part of quality assurance, performance optimisation and to ensure maintainability and scalability, it is essential to evaluate the underlying case study, paying particular attention to the implementation of the three explanation algorithms IG, GradCAM and LRP. Furthermore, user interviews can provide additional insights to support the evaluation and analysis of the explanation algorithms.

7.1 Methodology

Qualitative surveys, including user interviews, provide valuable insights into user needs and can aid developers in understanding requirements and preferences. User interviews are an important tool for planning explanation algorithms, as they provide direct assessments of understanding and comprehensibility. This allows for better tailoring of tools to social needs. It serves to identify problems and improve the user experience. In 1993, Jakob Nielsen and Thomas Landauer developed a mathematical model to determine the number of users needed for a sufficient sample [NL93]. This model remains valid and widely used today [Nie00] [TLN06] [All23]. The research states that only five test subjects are needed to reveal 80% of the problems [NL93] [Nie00]. Therefore, five test subjects were selected with diverse backgrounds for the user interviews to achieve a variance in the sample. The initial aim was to obtain a general assessment of the case study. For future research in this area, expert interviews should be conducted to gain a more focused impression of the target group [HKB15]. In the field of brain tumor research and image classification, the target group could be radiologists or medical professionals [Tal+19] [GSM21].

The Interview Guideline and the Consent Form can be found in Appendix B

Table 7.1 presents an overview of the test subjects, including their sociodemographic characteristics and background knowledge. The subjects vary in age and expertise in the field of machine learning. However, all participants received a thematic introduction to neural networks, the black box phenomenon and the purpose of explanatory algorithms, regardless of their expertise. The introduction aimed to familiarise the interviewee with the complex and diverse subject area of artificial intelligence and create a comfortable environment. The interviews consisted of a pre-interview, main interview, and post-interview, in total lasting approximately 20 minutes. The pre-interview aimed to obtain general information about the test person and familiarize them with the interview situation. The main interview was concerned with evaluating the individual explanation algorithms. To introduce variance into the visualisations and avoid the possibility that certain statements only relate to the selected output image, two different example images were used. Each test subject received either example image I or example image II consistently. The first image in the collection is a representation of a meningioma tumor and was sourced from publicly available medical X-ray images by [Osm23]. It was used as an input image that was unknown to the ResNet50 model. The second image is a glioma tumor from the test set that was used to train the model by [Che17]. The post-interview focused on evaluating the explanatory algorithms and personal classification preferences. Participants were informed about the privacy policy and asked to sign a consent form before the interview was recorded using MS Teams. All interviews were conducted remotely via MS Teams. Participants received an electronic invitation to their email address, stating that participation was voluntary and could be cancelled at any time. Additionally, the study's scope was explained, its benefits described and the consent form was sent. The study was anonymised, and participants were assigned an ID as can be seen in 7.1.

ID	Age	Profession	Interested in AI	Knowledge about NNs	Example Image
01	26	Sociology	Yes	No	I
02	25	Computer Science	Yes	Yes	II
03	51	Head of department in electrical company	Yes	No	I
04	49	Specialist for anaesthesia and intensive care	Yes	No	II
05	25	Computer Science	Yes	Yes	I

Table 7.1: Interview Testperson Overview

7.2 Interview Outcome

The main interview focussed on the following three questions:

1. What would be your expectations for an explanation tool?
2. How would you interpret the visualization?
3. Are there missing informationen in the visualization?

Question 1 was asked before evaluating the individual explanation algorithms. Questions 2 and 3 were repeated for each algorithm. Overall, the expectation for explanation tools is to provide a clear understanding of the result and facilitate work. The majority of test subjects found it easy to interpret the saliency map based on the IG. They immediately recognised differences in pixel colour and that the colour-coded pixels highlighted the brain tumor. However, two test subjects were unable to associate the pixel conspicuities with the tumor. This was independent of the example image, as both subjects were shown the opposite image. These two test subjects identified a lack of information regarding the spatial categorisation of the image, specifically in terms of orientation on the IG saliency map. Additionally, the other test subjects noted a lack of written explanations, particularly regarding the contextual classification of the visualisation. When interpreting the GradCAM visualisation, it was evident that participants with technical background knowledge, specifically those who were familiar with neural networks or had heard of them before, rated the algorithm significantly worse than those without such knowledge. The participants who were unfamiliar with neural networks found the GradCAM visualisation to be more informative as it highlighted more pixels and made the overall image of the head and brain more recognizable. However, participants with background knowledge criticized this aspect as the algorithm highlighted too many pixels as relevant features for decision-making. The presentation was criticised for lacking written information. The survey on the LRP explanation algorithm showed that participants, regardless of their background knowledge, struggled to interpret the visualisation due to inadequate pixel highlighting. The illustration was generally considered unclear, making it difficult for test subjects to understand what they were supposed to interpret. All test subjects cited the missing intention of the illustration and the request for more detailed information.

The post interview focussed on the following three questions:

1. Which visualization did you like most and why?
2. Which visualization did you like least and why?
3. Were your expectations of explanation tools met?

In the first question, the IG was identified as the preferred explanation algorithm. It is noteworthy that participants who claimed to have knowledge of neural networks favoured the IG because it highlights the pixels that are relevant to the tumor. These test subjects stated that the visualisation based on the LRP algorithm was the least effective, as it does not project any relevant features, making it impossible to trace back the basis on which the model arrived at a tumor classification. However, no single worst visualisation could be clearly identified, as all explanation algorithms were placed in this category at least once by different test subjects. The question of whether the expectations of explanation algorithms were met was mainly answered in the negative. The reasons given were the absence of supplementary written information to facilitate a better understanding of the results and the fact that the interpretations are primarily subjective. Finally, the evaluation of explanation algorithms revealed that they did not meet the expected standards of the test subjects.

7.3 Discussion

It is important to note that the user interview was conducted as a survey of a broadly diversified group of people, and therefore no expert interviews were conducted to represent the direct target group. However, this approach provided a general insight into the evaluation of explanation algorithms, as the study group was diverse and had varying levels of knowledge. Furthermore, it is recommended that non-experts evaluate systems beforehand to determine the system's epistemic trustworthiness based on its effectiveness, adherence to standards, and integrity of intentions. Subsequently, experts can provide their assessment [HKB15].

In summary, individuals prioritize differently based on their contextual knowledge. Those who were familiar with neural networks tended to focus on highlighting pixels that were as class-specific as possible, while those without prior knowledge of neural networks aimed to recognize the original image as accurately as possible. The thematic introduction, which was identical for all test subjects, did not appear to have any impact. It is therefore not surprising that the IG algorithm was identified as the favourite explanation algorithm by the majority of test subjects, closely followed by the GradCAM algorithm due to its colourful saliency map. However, the test subjects expressed a desire for textual explanations in addition to the visual representations, as it became clear that their expectations of explanatory algorithms were not met due to a lack of information and details. Therefore, it is reasonable to assume that visual explanations alone may not be entirely comprehensible. To reduce the possibility of misinterpretation, a textual description should be included to provide clarity on how the model arrived at the result.

8 Conclusion

This section focuses on reviewing the research, summarizing the results and providing a future outlook. The previous sections have already discussed the analysis and evaluation of XAI tools, legal liability and the impact of XAI tools, and results-oriented presentation of implementation evaluation.

8.1 Retrospective and results

This research explores artificial intelligence, specifically machine learning and neural networks. The analysis of the ResNet50 model architecture and performance confirms its suitability for classifying meningioma, glioma, and pituitary brain tumors with high accuracy. Within the model training the residual class None has been identified as a weakness due to a significant bias in the data and unclear training features underlying this class. As most research ends after solving the classification problem, it is essential to begin at this point and use explanatory algorithms. These algorithms are vital for tracing model predictions. In other words, to ensure traceability of model predictions, explanatory algorithms were used starting from the point of solving the classification problem. The implemented case study illustrates the entire process of model training, subsequent classification, and explanation of model behaviour. Furthermore, the practical example could be used to analyse and evaluate the XAI tools, allowing for consideration of the technical capabilities of each algorithm. The analysis of XAI tools is crucial due to new requirements and legal changes regarding the use of AI-based technologies. It is important to take into account the recently published Ethics Guidelines for Trustworthy AI, the AI Act, and legal liability for future research in this area. There is room for improvement in assessing the current state of explanatory algorithms in terms of ethical requirements and social desires. Although the algorithms can provide insight into the model behaviour and classification process, visualisations alone are insufficient for achieving maximum explainability. As a result, it is important in future research to determine which XAI tools can be used either as a supplement to visualisations or as stand-alone tools in order to maximise the explanatory content. In the user interviews in particular, an increased need for textual supplements was identified. Explanatory methods that enable written documentation and presentation of the results therefore appear promising. The evaluation criterion for trust and its analysis, carried out using the Ethics Guidelines of Trustworthy AI, revealed that the level of explainability differs depending on the explanation algorithm. The IG method, based on the Python case study, demonstrates the greatest technical robustness and the most informative visualisations, while GradCAM and LRP produce significantly inferior results. Compared to the other two algorithms, the IG algorithm stands out for highlighting class-specific features within tumor boundaries. This was evaluated both quantitatively for technical robustness and qualitatively through user interviews. The saliency maps under the calculations based on the GradCAM and LRP algorithms within the case study showed little to no class-relevant features within the tumor area. Consequently, test subjects in the user interviews had difficulty understanding the saliency maps and their interpretation. Methods introduced to clarify and make results transparent fail their purpose if end users have as much room for interpretation after using them as before. However, the problem was particularly evident in the case study involving the GradCAM and LRP-based algorithms. The visualisations did not clearly indicate that the ResNet50 model had performed the classification based on the tumor-specific pixels.

All three explanation algorithms have improved the properties of human agency and oversight, transparency, environmental and social well-being, and accountability according to the ethical guidelines. GradCAM and LRP provide more flexible parameter settings, resulting in an increasing emphasis on the first property. However, only the IG algorithm has achieved outstanding results in the criterion of technical robustness. In conclusion, the algorithms used in the case study adhere to the majority of ethical guidelines. They can be classified as methods to transform AI-based models into trustworthy systems. The research question:

- To what extent do the current explanation algorithms within XAI tools contribute to the explainability of AI systems, particularly concerning black box models?

can primarily be answered through the use of XAI tools. Their application is necessary for a transparent breakdown of the results. The analysis of the XAI tools has shown that the explanation algorithms largely fulfil the criteria for trustworthy AI. AI systems can be considered trustworthy if they have a transparent design and are ultimately explainable. The three algorithms used for visualising the ResNet50 model behaviour of tumor image classification play a significant role in achieving explainability.

The use of explanatory algorithms is necessary for the lawful introduction of AI-based software, considering the current legal situation regarding legal liability. This is due to the explainability and transparent design of such algorithms, which provide added traceability and information for the end user. To ensure that liability can be established in the event of a claim, specific requirements must be met in advance for high-risk systems, including classification models in the medical field. The AI Act specifies the requirements for high-risk systems from Article 8 to Article 15, which cover some aspects of the ethical guidelines. These requirements include human oversight, accuracy, robustness, cybersecurity, and transparency. Therefore, the answer to the research questions:

- How is legal liability addressed in the event of misbehavior, miscalculations, or misdiagnoses by an artificial system?
- What role do XAI tools play in resolving these liability issues?

depends on the characteristics of XAI tools. These characteristics are crucial and essential for the lawful integration of AI-based high-risk systems. Regarding the case study, liability can be determined based on the extent to which legal requirements were met in advance in the event of a malfunction. In case of damage, liability depends on the extent to which the parties have fulfilled their duty of care, complied with legal requirements under the AI Act and the Legal Liability Directive, and used the system properly according to instructions. Two main cases have been identified: when product owners neglect their duty of care or fail to comply with legal requirements, and when system users use the system in a way that goes against the instructions, resulting in a claim due to unlawful use. In both cases, the party that has violated its obligations is liable if it is reasonable to assume that the damage can be attributed to them, as the rebuttable presumption applies. The availability of XAI tools has a significant impact on the issue of liability. Traceability and transparency are crucial elements for the legal implementation of new technologies. Therefore, XAI tools play a vital role in resolving liability issues.

Overall, this work highlights the importance of developing XAI tools that provide meaningful insights for the successful implementation of new AI-based technologies. Additionally, by addressing the issue of liability, it introduces the complex topic of the legal system. Understanding the consequences and repercussions of damages is crucial for the safe use of AI systems. At the same time, this fosters an understanding of the legal requirements and enables the introduction of preventative measures to minimise the probability of damage and ensure the lawful and safe implementation of new AI systems. Explainability is a crucial tool for implementing new systems. It should be integrated into the development process to provide transparent insight into the system's process, results, and ultimately trustworthy support in the form of a technical system. This ensures that users, developers, and providers can understand the system's behaviour.

8.2 Limitations and Outlook

This work offers an overview of the current state of selected explanation algorithms and can serve as a starting point for further research in this area. Future studies should aim to implement or develop additional explanation algorithms to determine the extent to which other algorithm-specific properties influence the explanatory aspect of AI-based systems. It should be noted that this study only examined the GradCAM, IG, and LRP algorithms, and as a result, the findings are restricted to these XAI tools. This limitation emphasises the necessity for additional analysis using other XAI tools to obtain more universal and generally applicable results. Additionally, the visualisations of the Python case study do not typically highlight any class-specific features within the tumor boundaries. Therefore, the illustrations may not accurately reflect the model's behaviour. Despite the high probability that ResNet50 has correctly identified the tumor, the saliency maps do not clearly reflect this as no class-relevant features of the tumor are highlighted. This contradicts the assumption that ResNet50 has identified the specific tumor based on its class-specific features within the tumor region.

There is a need to implement an explanation algorithm that produces more precise visualisations for the tumor-specific classification problem with the ResNet50, thus achieving more informative results. The aim of this work and research framework is to achieve maximum transparency and trust in an AI-based system by providing clear explanations. After analysing the outcome of the case study it becomes clear that, there is potential to improve the level of explainability. One possibility would be by making incremental changes to the fundamental structure of the model classification process. The methodologies used in the case study are just one approach, and problem-specific modifications, such as adjusting parameter configurations or using different explanation algorithms as a basis, may lead to noticeable improvements. Furthermore, using a different neural network for tumor classification could be a sensible measure. This would allow testing the model's suitability for the problem not only theoretically but also practically during implementation. Additionally, altering the data basis could have a significant impact on the model's behaviour during training, as the data used to teach the model to classify plays a crucial role.

The Python case study is based on the ResNet50 model from [M20], which was already established. The model training and the selected dataset were not altered due to the high model accuracy and promising classification results. It is unclear why the visualisations, in the form of saliency maps, of the various explanatory algorithms in the case study show little to no class-specific features in the relevant areas of the tumor regions despite the high model accuracy. Furthermore, as a next step, expert interviews should be conducted based on user interviews, to generate a comparative value and to address the specific requirements and needs of the user group or individuals involved in the specific medical field. As a final step, once an extensive analysis of explanatory algorithms has been conducted and a suitable algorithm has been identified or developed that covers not only a large part of the ethical guidelines but all of these criteria, the parameter settings are checked and another dataset is used as a basis to ensure comparability, a user-friendly interface should be implemented to ensure a high-quality user experience during model use. Furthermore, it is important to note that any changes or adjustments to the AI Act or the AI Liability Directive cannot be ruled out, as these are draft laws still in development and testing phases. Future research based on this work could involve legal experts to supplement technical assessments and answer questions of liability related to artificial systems.

A Case Study: GitHub Repository

The GitHub Repository for the associated Case Study on Brain Tumor Image Classification and Explanation can be found here: Click here, to get to the GitHub Repository or follow this link: https://github.com/alinatenne/xai_in_medicine.git.

B Evaluation of Implementation: Interview

Interview guideline

The interviewer must answer all questions included in the semi-standardized guide, while maintaining the necessary information. Deviations in wording and order are allowed as long as all information is obtained.

Date: ______________________________

Interview Code:_____________________

Example Image No:__________________

☐ Welcome and Information
Welcome test person and provide information about the procedure.
☐ Declaration of Consent
Inform test person about the declaration of consent and ask him/ her to sign.
☐ Questions
Ask test person for questions before starting the interview.
☐ Pre Interview
Focus on general questions to familiarize test person with interview situation
☐ Short Introduction
Provide a short introduction to the field of neural networks and model predictions
☐ Interview

- *Focus on evaluation/ understanding/ interpretation of every single explanation algorithm*
- *Switch between example image I and example image II for each test subject so that they will receive either I or II for all explanation algorithms*
- *Ask people to take as much time as they would like, to carefully view the visualizations, to be honest not polite*

☐ Post Interview
Focus on evaluation of whole interview and the overall opinion of explanation algorithms
☐ Thanks and Goodbye
Thank test person for the time and participation

Einverständniserklärung
Declaration of Consent

Das Interview, an dem du teilnimmst, dient der Evaluation von Erklärungsalgorithmen, die im Rahmen meiner Masterarbeit zum Thema „XAI in Medicine: Analysis and Evaluation of XAI Tools and Legal Liability for Neural Networks. A Case Study on Tumor Image Classification“ entstanden sind.

The interview you are taking part in is for the evaluation of explanatory algorithms that were developed as part of my Master's thesis on "XAI in Medicine: Analysis and Evaluation of XAI Tools and Legal Liability for Neural Networks. A Case Study on Tumor Image Classification".

Bestandteile des Interviews
Du wirst gebeten,

- Fragen zu beantworten
- Visualisierungen von Erklärungsalgorithmen anzuschauen

Hinweis: dieses Interview ist freiwillig und darf jederzeit ohne die Angabe von Gründen abgebrochen werden.

Components of the interview
You be asked to,

- answer questions
- look at visualisations of explanation algorithms

Note: this interview is voluntary and can be cancelled at any time without giving reasons.

Datenschutzerklärung
Der Verlauf des Interviews wird digital durch deine Audiospur aufgezeichnet. Dein Name und Deine Kontaktdaten werden nicht an Dritte weitergegeben, sondern ausschließlich anonymisiert genutzt. Du hast jederzeit das Recht, Deine Einwilligung ohne Angabe von Gründen zu widerrufen.

Ich habe die obige Einverständniserklärung zum Interview gelesen und erkläre mich mit allen Punkten einverstanden.

Privacy policy
The course of the interview will be digitally recorded by your audio track. Your name and contact details will not be passed on to third parties, but will only be used in anonymised form. You have the right to withdraw your consent at any time without giving reasons.

I have read the above declaration of consent for the interview and agree to all points.

____
Vor- und Nachname (Druckbuchstaben)
First and last name (block capitals)

______________________________	____
Datum *Date*	Unterschrift *Signature*

Pre Interview

The main purpose of these questions is to confirm the criteria of the test person and to familiarise them with the interview situation.

1. Wie alt bist du?
   How old are you?
2. Was machst du beruflich / Was studierst du?
   What is your profession / What are you studying?
3. Interessierst du dich für künstliche Intelligenz?
   Are you interested in AI?
4. Hast du schon mal von Neuronalen Netzwerken gehört?
   Have you ever heard of neuronal networks?

Short Introduction

The aim of the short topic introduction is to present the field of artificial intelligence and machine learning in order to give the test person the opportunity to familiarize themselves with the subject area and to provide background information on the explanation tools.

Inhaltliche Darstellung des Textes

Im maschinellen Lernen werden Neuronale Netzwerke verwendet, um Modelvorhersagen zu generieren. Ein solches Modell habe ich implementiert um drei verschiedene Tumortypen vorherzusagen, nämlich Meningioma, Glioma und Hypothysen Tumore. Da diese Modelle sogenannte Black-Box Phänomene darstellen, kommen Erklärungsalgorithmen zu Einsatz, um Licht ins Dunkle zu bringen. Black-Box-Phänomene zeichnen sich dadurch aus, dass nicht nachvollziehbar ist anhand welcher Kriterien und Eigenschaften das Model zu der gewissen Vorhersage bzw. dem Ergebnis gekommen ist. Dem sollen Erklärungsalgorithmen entgegenwirken.

Content presentation of the text

In machine learning, neural networks are used to generate model predictions. I have implemented such a model to predict three different tumour types, namely meningioma, glioma and pituitary. Since these models represent so-called black-box phenomena, explanatory algorithms are used to shed light on the dark. Black-box phenomena are characterised by the fact that it is not possible to understand which criteria and properties the model used to arrive at a certain prediction or result. Explanatory algorithms are designed to counteract this.

Main Interview

Before proceeding with the repetitive and consistent questions for all explanation algorithms, an initial question about expectations should be asked. This question will only be asked once for later reference in the post-interview.

1. Was wären deine Erwartungen an ein Erklärungstool?
   What would be your expectations for an explanation tool?

The following questions pertain to all three explanation algorithms (GradCAM, LRP, and IG) and aim to assess the degree to which the implementation achieves its goal of elucidating the model's predictions.

2. Wie würdest du diese Visualisierung interpretieren?
   How would you interpret the visualization?
3. Fehlen dir Informationen in der Darstellung?
   Are there missing information in the visualization?

Note: There are two example visualizations. Each test subject will receive either Example Image I or Example Image II consistently throughout the interview. However, it is important to note that between individual test runs, the example image will change, and thus these questions will be asked with regard to either one or the other image.

Post Interview

The main objective is to ask for an overall impression after the test person has seen three exemplary visualizations of all explanation algorithms.

1. Welche Visualisierung hat dir besonders gut gefallen und warum?
   Which visualization did you like most and why?
2. Welche Visualisierung hat dir am wenigsten gefallen und warum?
   Which visualization did you like least and why?
3. Wurden deine Erwartungen an Erklärungstools erfüllt?
   Were your expectations of explanation tools met?

References

[ABW22] MU Alam, JR Baldvinsson, and Y Wang. Exploring LRP and Grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography. In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. 2022, pp. 258–263 (pages 51, 52).

[AZ21] AM Alhassan and WMNW Zainon. Brain tumor classification in magnetic resonance image using hard swish-based RELU activation function-convolutional neural network. In: *Neural Computing and Applications* 33:(2021), 9075–9087 (pages 9, 11, 32, 59).

[Ali+23] S Ali et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. In: *Information Fusion* 99:(2023), 101805. https://www.sciencedirect.com/science/article/pii/S1566253523001148 (pages 7–9, 11, 12, 49).

[All23] R Allabarton. Conduct a heuristic evaluation: Step-by-step tutorial and tips. url https://blog.logrocket.com/ux-design/conduct-heuristic-evaluation-tutorial/. 2023 (page 67).

[Alo21] A Alorf. The Practicality of Deep Learning Algorithms in COVID-19 Detection: Application to Chest X-ray Images. In: *Algorithms* 14(6):(2021). https://www.mdpi.com/1999-4893/14/6/183 (pages 8, 9).

[Alr+23] AZ Alrosan et al. The effects of statin therapy on brain tumors, particularly glioma: a review. In: *Anti-Cancer Drugs* 34(9):(Oct. 2023), 985–994 (pages 1, 36).

[Alz+21] L Alzubaidia et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In: *Journal of Big Data* 8(53):(2021) (pages 22, 28).

[And+21] CJ Anders et al. Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. In: *CoRR* abs/2106.13200:(2021) (pages 45, 46).

[Ant+21] L Antony et al. A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction. In: *IEEE Access* 9:(2021), 126481–126501 (pages 8, 9).

[Art19] HLEG on Artificial Intelligence. *Ethics Guidelines For Trustworthy AI*. Brussels: European Commission, 2019 (pages 9, 12, 49–53, 55–61).

[Art20] HLEG on Artificial Intelligence. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. Brussels: European Commission, 2020 (pages 49, 50, 57, 60, 61).

[Asi+23] AA Asiri et al. Brain Tumor Detection and Classification Using Fine-Tuned CNN with ResNet50 and U-Net Model: A Study on TCGA-LGG and TCIA Dataset for MRI Applications. In: *Life* 13(7):(2023). https://www.mdpi.com/2075-1729/13/7/1449 (pages 1, 9, 10).

[Bac+15] S Bach et al. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In: *CoRR* abs/1512.00172:(2015). arXiv: 1512.00172. http://arxiv.org/abs/1512.00172 (pages 1, 2, 7, 44, 45, 51).

[Ban23] M Banoula. Machine Learning Steps: A Complete Guide. url https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps. 2023 (pages 18, 19).

[Bar22] K Barkved. How To Know if Your Machine Learning Model Has Good Performance. url https://www.obviously.ai/post/machine-learning-model-performance. 2022 (page 26).

[Bar+20] A Barredo Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In: *Information Fusion* 58:(2020), 82–115. https://www.sciencedirect.com/science/article/pii/S1566253519308103 (pages 8, 15).

[BP22] MC Barton and J Pöppelbuß. Prinzipien für die ethische Nutzung künstlicher Intelligenz. In:

HMD Praxis der Wirtschaftsinformatik 59:(2022), 468–481 (page 55).

[Bar23] T Bartz-Beielstein. Hyperparameter Tuning and Optimization Applications. In: *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*. Ed. by E Bartz et al. Singapore: Springer Nature Singapore, 2023, pp. 165–175. https://doi.org/10.1007/978-981-19-5170-1_6 (page 20).

[BTR23] J Bhagya, J Thomas, and ED Raj. Exploring Explainability and Transparency in Deep Neural Networks: A Comparative Approach. In: *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2023, pp. 664–669 (pages 2, 50, 59).

[BFM23] D de Biase, E Franceschi, and G Marucci. Editorial: Advances in brain tumors diagnosis and treatment. In: *Sec. Pathology* 10:(May 2023) (page 1).

[BII16] A Borji, S Izadi, and L Itti. iLab-20M: A Large-Scale Controlled Object Dataset to Investigate Deep Learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (page 34).

[Bor+23] K Borys et al. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. In: *European Journal of Radiology* 162:(2023), 110786. https://www.sciencedirect.com/science/article/pii/S0720048X23001006 (pages 3, 7, 9, 11, 49, 57).

[Bot21] M Both. Die Verlustfunktion künstlicher neuronaler Netze – Die Basis für ein erfolgreiches Training und die Initialzündung der Backpropagation. url https://www.tecislava.com/blog/verlustfunktion. 2021 (page 25).

[Bou24] C Boutin. NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems. url https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems. 2024 (page 52).

[Cap24] Captum. Integrated Gradients. url https://captum.ai/api/integrated$_g$*radients.html*. 2024 (page 47).

[ČGS15] P Čerka, J Grigienė, and G Sirbikytė. Liability for damages caused by artificial intelligence. In: *Computer Law and Security Review* 31(3):(2015), 376–389. https://www.sciencedirect.com/science/article/pii/S026736491500062X (page 61).

[Ces+23] C Cestonaro et al. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. In: *Frontiers in Medicine* 10:(Nov. 2023) (pages 3, 4).

[CO21] M Ceylan and AH Ornek. Explainable Artificial Intelligence (XAI): Classification of Medical Thermal Images of Neonates Using Class Activation Maps. In: *Traitement du Signal* 38(5):(2021), 1271–1279 (pages 16, 18).

[Cha+23] V Chamola et al. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). In: *IEEE Access* 11:(2023), 78994–79015 (pages 2, 50, 59).

[Che17] J Cheng. Brain Tumor Dataset. url https://doi.org/10.6084/m9.figshare.1512427.v5. 2017 (pages 12, 32, 34, 35, 37, 55, 58, 59, 68).

[Cla22] M Clayton. How to Pick the Best Graphics Card for Machine Learning. url https://towardsdatascience.com/how-to-pick-the-best-graphics-card-for-machine-learning-32ce9679e23b. 2022 (pages 21, 59).

[Coh23] A Cohen-Gadol. What You Should Know About Surviving Pituitary Tumors. url https://www.aaroncohen-gadol.com/en/patients/pituitary-tumor/survival/overview. 2023 (page 37).

[Com21] E Commission. Regulation of the European Parliament and of the Council. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative Acts. In: *COM* 206:(2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (pages 4, 62, 63).

[Com22a] E Commission. Ethics guidelines for trustworthy AI. url https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. 2022 (page 9).

[Com22b] E Commission. Proposal for a: DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). In: *COM* 496:(2022). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0496 (pages 4, 9, 61–63).

[Com24] E Commission. Regulatory framework proposal on artificial intelligence. url https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. 2024 (page 62).

[DVV22] M Dabass, S Vashisth, and R Vig. A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images. In: *Computers in Biology and Medicine* 147(1):(June 2022) (pages 21, 22, 25, 27, 32, 33, 43).

[Dec86] R Dechter. Learning While Searching in Constraint-Satisfaction-Problems. In: Jan. 1986, pp. 178–185 (page 15).

[Dev22] G for Developers. The Size and Quality of a Data Set. url https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality?hl=en. 2022 (page 34).

[Dor22] E Dorfman. How Much Data Is Required for Machine Learning? url https://postindustria.com/how-much-data-is-required-for-machine-learning/. 2022 (page 34).

[Dra19] R Draelos. Multi-label vs. Multi-class Classification: Sigmoid vs. Softmax. url https://glassboxmedicine.com/2019/05/26/classification-sigmoid-vs-softmax/. 2019 (page 28).

[EV24] P Ehrlich and B Vaughan. Fallback plan and general safety. url https://guidelines.panelfit.eu/ai/general-exposition/technical-robustness-and-safety/fallback-plan-and-general-safety/. 2024 (page 53).

[Erd22] K Erdem. XAI Methods — Integrated Gradients Dive into Integrated Gradients method. How the values are calculated? What are the different baselines? url https://medium.com/@kemalpiro/xai-methods-integrated-gradients-6ee1fe4120d8. 2022 (page 46).

[Est+17] A Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. In: *Nature* 542:(2017), 115–118 (page 3).

[FKH18] S Falkner, A Klein, and F Hutter. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. 2018. arXiv: 1807.01774 [cs.LG] (page 20).

[Fas22] Fastai. Fastbook. url https://github.com/fastai/fastbook.git. 2022 (pages 18–20).

[Fel+23] T Fel et al. Don't Lie to Me! Robust and Efficient Explainability With Verified Perturbation Analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 16153–16163 (page 57).

[Fer20] FG Fernandez. TorchCAM: class activation explorer. https://github.com/frgfm/torch-cam. Mar. 2020 (pages 43, 44).

[Fin+23] B Finzel et al. Domain-Specific Evaluation of Visual Explanations for Application-Grounded Facial Expression Recognition. In: *Machine Learning and Knowledge Extraction*. Ed. by A Holzinger et al. Cham: Springer Nature Switzerland, 2023, pp. 31–44 (pages 38, 51, 53, 54, 57).

[Fra24] Fraunhofer-Institut. Künstliche Intelligenz in der Medizin. url https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz/kuenstliche-intelligenz-medizin.html. 2024 (page 7).

[Go 23] KYZ Go Kyeong-O. Brain Invasion and Trends in Molecular Research on Meningioma. In: *btrt* 11(1):(2023), 47–58. eprint: http://www.e-sciencecentral.org/articles/?scid=1516081742. http://www.e-sciencecentral.org/articles/?scid=1516081742 (page 35).

[GSM21] P Gohel, P Singh, and M Mohanty. Explainable AI: current status and future directions. In: *CoRR* abs/2107.07045:(2021). arXiv: 2107.07045. https://arxiv.org/abs/2107.07045 (pages 1, 3, 7, 50, 57, 67).

[Gri23] S Grimmelikhuijsen. Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. In: *PAR: Public Administration Review* 82(2):(2023), 241–262 (pages 2, 50, 59).

[GJW24] A Grzybowski, K Jin, and H Wu. Challenges of Artificial Intelligence in Medicine and Dermatology. In: *Clinics in Dermatology*:(2024). https://www.sciencedirect.com/science/article/pii/S0738081X23002651 (pages 55, 58).

[Gül23] K Gülen. AI and Ethics: Balancing progress and protection. url https://dataconomy.com/2023/01/16/artificial-intelligence-security-issues/. 2023 (pages 52, 53).

[Hac+20] P Hacker et al. Explainable AI under contract and tort law: legal incentives and technical challenges. In: *Artifcial Intelligence and Law*:(2020) (page 61).

[Haq+23] EU Haq et al. An MRI-based deep learning approach for efficient classification of brain tumors. In: *Journal of Ambient Intelligence and Humanized Computing* 14:(2023), 6697–6718 (pages 32, 59).

[HYA20] RI Hasan, SM Yusuf, and L Alzubaidi. Review of the State of the Art of Deep Learning for Plant Diseases: A Broad Analysis and Discussion. In: *Plants* 9(10):(2020). https://www.mdpi.com/2223-7747/9/10/1302 (page 21).

[Has+22] E Hasanpour Zaryabi et al. Unboxing the Black Box of Attention Mechanisms in Remote Sensing Big Data Using XAI. In: *Remote Sensing*

14(24):(2022). https://www.mdpi.com/2072-4292/14/24/6254 (page 3).

[Has+24] V Hassija et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. In: *Cognitive Computation* 16:(2024), 45–74 (page 3).

[He+15] K He et al. Deep Residual Learning for Image Recognition. 2015. arXiv: 1512.03385 [cs.CV] (pages 24, 29, 30, 32).

[He+16] K He et al. Identity Mappings in Deep Residual Networks. In: *Computer Vision – ECCV 2016*. Ed. by B Leibe et al. Cham: Springer International Publishing, 2016, pp. 630–645 (page 24).

[Hea23] NI of Health (NIH). Pituitary Tumors. url https://www.ninds.nih.gov/health-information/disorders/pituitary-tumors. 2023 (page 37).

[HKB15] F Hendriks, D Kienhues, and R Bromme. Measuring Laypeople's Trust in Experts in a Digital Age: The Muenster Epistemic Trustworthiness Inventory (METI). In: *PLOS ONE* 10(10):(Oct. 2015), 1–20. https://doi.org/10.1371/journal.pone.0139309 (pages 67, 70).

[Hin+23] F Hinterwimmer et al. Prediction of complications and surgery duration in primary TKA with high accuracy using machine learning with arthroplasty-specific data. In: *Knee Surg Sports Traumatol Arthrosc* 31:(2023), 1323–1333 (pages 15, 26).

[Hin18] G Hinton. Deep Learning—A Technology With the Potential to Transform Health Care. In: *JAMA* 320(11):(Sept. 2018), 1101–1102. eprint: https://jamanetwork.com/journals/jama/articlepdf/2701666/jama_hinton_2018_vp_180096.pdf. https://doi.org/10.1001/jama.2018.11100 (page 15).

[HSA06] G Hinton, O S., and TY A. A fast learning algorithm for deep belief nets. In: *Neural Comput* 18(7):(July 2006), 1527–1554 (page 15).

[Hir24] D Hirntumorhilfe. Zahlen und Fakten über Hirntumoren weltweit. url https://www.hirntumorhilfe.de/hirntumor/statistiken/zahlen-weltweit/. 2024 (page 1).

[Hu+21] X Hu et al. Model complexity of deep learning: a survey. In: *Knowl Inf Syst* 63:(Aug. 2021), 2585–2619 (pages 16, 29, 32).

[IA18] MR Ismael and I Abdel-Qader. Brain Tumor Classification via Statistical Features and Back-Propagation Neural Network. In: *2018 IEEE International Conference on Electro/Information Technology (EIT)*. 2018, pp. 0252–0257 (page 10).

[JK15] HK Jabbar and RZ Khan. Methods to avoid over-fitting and under-fitting in supervised machine leanring (Comparative Study). In: *Computer Science, Communication and Instrumentation Devices*:(2015), 163–171 (page 20).

[Jea+23] C Jean-Quartier et al. The Cost of Understanding: XAI Algorithms towards Sustainable ML in the View of Computational Cost. In: *Computation* 11(5):(2023). https://www.mdpi.com/2079-3197/11/5/92 (pages 15, 20, 59).

[Kar23] A Karpathy. Deep Neural Nets. url https://github.com/karpathy. 2023 (pages 32, 33).

[KSH18] TA Kass-Hout, LM Stevens, and JL Hall. American Heart Association Precision Medicine Platform. In: *Circulation* 137:(Feb. 2018), 647–649 (page 13).

[KG20] T Kaur and T Gandhi. Deep convolutional neural networks with transfer learning for automated brain image classification. In: *Machine Vision and Applications* 31(20):(2020) (pages 24, 32, 59).

[KM23] M Kaveh and M Mesgari. Application of Meta-Heuristic Algorithms for Training Neural Networks and Deep Learning Architectures: A Comprehensive Review. In: *Neural Process Letters* 55:(2023), 4519–4622 (pages 20–23, 59).

[Ket24] R Ketter. Meningeome. url https://www.uniklinikum-saarland.de/de/einrichtungen/kliniken$_i$*nstitute/neurochirurgie/patien*
2024 (page 42).

[Kha+20] HA Khan et al. Brain tumor classification in MRI image using convolutional neural network. In: *Mathematical Biosciences and Engineering* 17(5):(Sept. 2020), 6203–6216 (pages 25–27).

[Kla+17] G Klambauer et al. Self-Normalizing Neural Networks. 2017. arXiv: 1706.02515 [cs.LG] (page 17).

[Kra+21] T Kraus et al. ERKLÄRBARE KI: Anforderungen, Anwendungsfälle und Lösungen. In: *Künstliche Intelligenz als Treiber für volkswirtschaftlich relevante Ökosysteme*:(2021), 1–100 (page 41).

[Lap+19] S Lapuschkin et al. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. In: *CoRR* abs/1902.10178:(2019). arXiv: 1902.10178. http://arxiv.org/abs/1902.10178 (pages 2, 7).

[Lau+23] L Launet et al. Federating Medical Deep Learning Models from Private Jupyter Notebooks to Distributed Institutions. In: *Appl. Sci.* 13(2):(Jan. 2023) (page 13).

[LFM04] M van Lent, W Fisher, and M Mancuso. An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. In: *IAAI EMERGING APPLICATIONS*:(2004), 900–907 (page 7).

[LGY20] Y Liu, Y Gao, and W Yin. An Improved Analysis of Stochastic Gradient Descent with Momentum. In: *Conference on Neural Information Processing Systems* 34:(2020) (page 33).

[Luc+22] A Lucieri et al. Erklärbare KI in der medizinischen Diagnose – Erfolge und Herausforderungen. In: *Künstliche Intelligenz im Gesundheitswesen: Entwicklungen, Beispiele und Perspektiven*. Ed. by MA Pfannstiel. Wiesbaden: Springer Fachmedien Wiesbaden, 2022, pp. 727–754. https://doi.org/10.1007/978-3-658-33597-7_35 (pages 3, 7).

[M20] AK M. Neural Black. url https://github.com/aksh-ai/neuralBlack. 2020 (pages 12, 21, 22, 24–28, 30, 32, 34, 35, 38, 39, 58, 59, 74).

[Mah+22] S Mahajan et al. World Health Organization Classification of Tumors of the Central Nervous System 5th Edition (WHO CNS5): What is new? In: *Indian Journal of Pathology and Microbiology* 65(5):(2022), 5–13. eprint: https://www.ijpmonline.org/article.asp?issn=0377-4929;year=2022;volume=65;issue=5;spage=5;epage=13;aulast=Mahajan;t=6. https://www.ijpmonline.org/article.asp?issn=0377-4929;year=2022;volume=65;issue=5;spage=5;epage=13;aulast=Mahajan;t=6 (page 36).

[Mar+15] Martín Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. 2015. https://www.tensorflow.org/ (page 60).

[MH20] A Moghar and M Hamiche. Stock Market Prediction Using LSTM Recurrent Neural Network. In: *Procedia Computer Science* 170:(2020), 1168–1173 (page 26).

[Mon+19] G Montavon et al. Layer-Wise Relevance Propagation: An Overview. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by W Samek et al. Cham: Springer International Publishing, 2019, pp. 193–209. https://doi.org/10.1007/978-3-030-28954-6_10 (page 60).

[Mor18] B Morris. Mastering the Learning Rate to Speed Up Deep Learning. url https://brandonmorris.dev/2018/06/24/mastering-the-learning-rate/. 2018 (page 33).

[Mos+23] E Mosqueira-Rey et al. Human-in-the-loop machine learning: a state of the art. In: *Artificial Intelligence Review* 56:(2023), 3005–3054 (page 52).

[New23] J Newman. *A Taxonomy of Trustworthiness for Artificial Intelligence: CONNECTING PROPERTIES OF TRUSTWORTHINESS WITH RISK MANAGEMENT AND THE AI LIFECYCLE*. UC Berkeley: CLTC WHITE PAPER SERIES, 2023 (page 9).

[Nie00] J Nielsen. Why You Only Need to Test with 5 Users. url https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. 2000 (page 67).

[NL93] J Nielsen and TK Landauer. A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: Association for Computing Machinery, 1993, pp. 206–213. https://doi.org/10.1145/169059.169166 (page 67).

[Osm23] MA Osman. Case Courtesy rID:23037. url https://radiopaedia.org/articles/meningioma. 2023 (pages 44, 45, 47, 68).

[Pai+19] R Paisan et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. In: *NPJ Digital Medicine* 2(1):(Dec. 2019). https://www.proquest.com/scholarly-journals/deep-learning-versus-human-graders-classifying/docview/2528862371/se-2 (page 3).

[Par23] E Parliament. EU AI Act: first regulation on artificial intelligence. url https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. 2023 (pages 9, 12, 49).

[Pas16] V Paschos. An overview on polynomial approximation of NP-hard problems. In: *Yugoslav Journal of Operations Research* 19(1):(2016). http://yujor.fon.bg.ac.rs/index.php/yujor/article/view/325 (page 21).

[Ped+11] F Pedregosa et al. Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12:(2011), 2825–2830 (pages 33, 38, 39).

[Pei24] E Peinsipp-Byma. KI-Modelle verstehen und nachvollziehen. url https://www.iosb.fraunhofer.de/de/kompetenzen/bildauswertung/interaktive-analyse-diagnose/erklaerbare-ki.html. 2024 (page 1).

[Pet22] D Petkovic. It is not "accuracy vs. explainability" – we need both for trustworthy AI systems. 2022. arXiv: 2212.11136 [cs.LG] (pages 56, 57).

[Pra22] R Pramoditha. How to Choose the Optimal Learning Rate for Neural Networks. url https://towardsdatascience.com/how-to-choose-the-optimal-learning-rate-for-neural-networks-362111c5c783. 2022 (page 33).

[Pyt23] Pytorch. Model understanding with Captum. url https://pytorch.org/tutorials/beginner/introyt/captumyt.html. 2023 (pages 47, 48).

[Reh+20] A Rehman et al. A Deep Learning-Based Framework for Automatic Brain Tumors Classification Using Transfer Learning. In: *Circuits, Systems and Signal Processing* 39:(2020), 757–775 (pages 9, 11, 32, 59).

[RSG16] MT Ribeiro, S Singh, and C Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778 (page 2).

[Ric+23] KM Richmond et al. Explainable AI and Law: An Evidential Survey. In: *MOBILE - Center of Excellence for Global Mobility Law* 8:(2023) (page 61).

[S B+23] S S Band et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. In: *Informatics in Medicine Unlocked* 40:(2023), 101286. https://www.sciencedirect.com/science/article/pii/S2352914823001302 (pages 7, 8, 57).

[Saj+19] M Sajjad et al. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. In: *Journal of Computational Science* 30:(2019), 174–182. https://www.sciencedirect.com/science/article/pii/S1877750318307385 (page 32).

[SM19] W Samek and KR Müller. Towards Explainable Artificial Intelligence. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019, pp. 5–22 (pages 2, 59).

[SK22] PL Saveri and S Kumar. Classification of Cancerous Lung Images by using Transfer Learning. In: *2022 8th International Conference on Signal Processing and Communication (ICSC)*. Dec. 2022, pp. 298–303 (pages 3, 9, 11, 24, 25, 32).

[Sch16] J Schmidhuber. Deep Learning. In: *Encyclopedia of Machine Learning and Data Mining*. 2016, pp. 298–303 (pages 15, 19).

[Sch15] J Schmidhuber. Deep learning in neural networks: An overview. In: *Neural Networks* 61:(2015), 85–117. https://www.sciencedirect.com/science/article/pii/S0893608014002135 (pages 15, 19).

[Sch20] J Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. In: *The Annals of Statistics* 48(4):(2020), 1875–1897. https://doi.org/10.1214/19-AOS1875 (page 17).

[Sch21] I Schneider. Diskriminierungsgefahren und Regulationsansätze bei der medizinischen Nutzung von KI. In: *Zeitschrift für medizinische Ethik* 67(3):(2021), 327–350. https://brill.com/view/journals/zfme/67/3/article-p327_4.xml (pages 55, 56, 58).

[Sel+19] RR Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *International Journal of Computer Vision* 128(2):(Oct. 2019), 336–359. https://doi.org/10.1007%2Fs11263-019-01228-7 (pages 16, 18, 19, 42–44, 52, 60).

[Sha+23] SR Shah et al. Comparing Inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A Case Study on Early Detection of a Rice Disease. In: *Agronomy* 13(6):(2023). https://www.mdpi.com/2073-4395/13/6/1633 (pages 9, 24, 25).

[Sim] M Simon. Meningeome. url https://www.dgnc.de/gesellschaft/fuer-patienten/meningeome/ (page 36).

[SZ15] K Simonyan and A Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. arXiv: 1409.1556 [cs.CV] (pages 17, 18).

[SSG18] S Singaravel, J Suykens, and P Geyer. Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction. In: *Advanced Engineering Informatics* 38:(2018), 81–90. https://www.sciencedirect.com/science/article/pii/S1474034617305359 (page 34).

[Smi18] LN Smith. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1 - Learning Rate, Batch Size, Momentum, and Weight Decay. In: *US Naval Resaerch Laboratory Technical Report* 5510(026):(Apr. 2018) (page 27).

[Sri+22] G Srivastava et al. XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions. 2022. arXiv: 2206.03585 [cs.CR] (pages 52, 53).

[Sto23] M Stojiljković. Stochastic Gradient Descent Algorithm With Python and NumPy. url https://realpython.com/gradient-descent-algorithm-python/. 2023 (page 33).

[Sub18] V Subramanian. *Deep Learning with PyTorch. A practical approach to built neural network models using PyTorch*. Birmingham, Mumbai: Packt, 2018 (pages 17, 22, 28, 33, 43, 47, 52).

[SSA19] HH Sultan, NM Salem, and W Al-Atabany. Multi-Classification of Brain Tumor Images Using Deep Neural Network. In: *IEEE Access* 7:(2019), 69215–69225 (pages 9, 10).

[STY17] M Sundararajan, A Taly, and Q Yan. Axiomatic Attribution for Deep Networks. In: *CoRR* abs/1703.01365:(2017). arXiv: 1703.01365. http://arxiv.org/abs/1703.01365 (pages 46, 47, 51).

[Tal+19] M Talo et al. Convolutional Neural Networks for Multiclass Brain Disease Detection Using MRI Images. In: *Computerized Medical Imaging and Graphics* 78:(Oct. 2019) (pages 9, 11, 12, 16, 22, 24–27, 32, 67).

[Tew+21] IA Tewarie et al. Survival prediction of glioblastoma patients—are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential. In: *Neurosurgical Review* 44:(2021), 2047–2057 (pages 9, 11, 12, 36).

[Tho22] M Thomas. The History of Deep Learning: Top Moments That Shaped the Technology. url https://builtin.com/artificial-intelligence/deep-learning-history. 2022 (page 15).

[Ton10] JC Tonn. Astrozytome und Glioblastome. In: *Oncology of CNS Tumors*. 2010 (page 37).

[TLN06] CW Turner, JR Lewis, and J Nielsen. Determining Usability Test Sample Size. In: *International Encyclopedia of Ergonomics and Human Factors* 3:(2006), 3084–3088 (page 67).

[van+22] BH van der Velden et al. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. In: *Medical Image Analysis* 79:(2022), 102470. https://www.sciencedirect.com/science/article/pii/S1361841522001177 (pages 1, 41).

[Vyh+23] E Vyhmeister et al. A responsible AI framework: pipeline contextualisation. In: *AI and Ethics* 3:(2023), 175–197 (page 52).

[Wan+21] X Wang et al. A radiomics model combined with XGBoost may improve the accuracy of distinguishing between mediastinal cysts and tumors: a multicenter validation analysis. In: *Ann Transl Med* 9(23):(Dec. 2021) (pages 8, 9).

[Was23] B Wasike. How to Build a Deep learning model with Keras and ResNet-50. url https://medium.com/@bravinwasike18/building-a-deep-learning-model-with-keras-and-resnet-50-9dd6f4eb3351. 2023 (page 16).

[Web+23] L Weber et al. Beyond explaining: Opportunities and challenges of XAI-based model improvement. In: *Information Fusion* 92:(2023), 154–176. https://www.sciencedirect.com/science/article/pii/S1566253522002238 (pages 15, 58, 59).

[Xia+20] Y Xiao et al. A review of object detection based on deep learning. In: *Multimedia Tools and Applications* 79:(2020), 23729–23791 (page 21).

[YC20] M Yildirim and A Cinar. Classification of Alzheimer's disease MRI images with CNN based hybrid method. In: *Ingénierie des Systèmes d'Information* 25(4):(2020), 413–418 (pages 16, 22).

[Zei+22] RA Zeineldin et al. Explainability of deep neural networks for MRI analysis of brain tumors. In: *International Journal of Computer Assisted Radiology and Surgery* 17:(2022), 1673–1683 (page 1).

[Zha+23] A Zhang et al. Dive into Deep Learning. url http://preview.d2l.ai/d2l-en/master/. 2023 (pages 23, 24).