INTERPRETABILITY OF MACHINE LEARNING MODEL FOR BRAIN TUMOR ASSESSMENT

by

HAMZAT OLAMIDE KESHINRO

Thesis submitted to University of Plymouth

in partial fulfilment of the requirements for the degree of

MSc Artificial Intelligence

University of Plymouth

Faculty of Science & Engineering

September 2023

Copyright statement

Masters Dissertations: Consent Form

I Hamzat Keshinro grant to The University of Plymouth the nonexclusive right to create a digital version of the above-named publication and to make my dissertation available as part of the University Libraries' digital repository. I understand that the full text of my dissertation will be available to university staff and students in digital form, and I give my permission for the University Library to reproduce, distribute, display, and transmit in order to make it available online. I affirm that my work does not, to the best of my knowledge, infringe or violate any rights of others and any third party copyright material quoted or utilised within the dissertation is appropriately referenced.

Signed:………………………………………………………………………………………

Date:14/09/23

ABSTRACT

This dissertation investigates the application of machine learning algorithms, specifically Random Forest, Logistic Regression, and Decision Tree models, to the field of radiomics, focusing on the diagnosis and characterization of meningioma tumours. A dual-dataset approach is employed, incorporating both T1-segmented and T2-featured medical images to provide a comprehensive analysis. The study is driven by five main objectives, including the development and evaluation of novel methods for interpretability and explainability in radiomics machine learning models. Quantitative measures reveal impressive model performances, with a Random Forest Classifier achieving an accuracy of 75% and an AUC score of 0.90. However, the study places equal emphasis on qualitative measures to assess the real-world applicability and interpretability of these models. Employing techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations), the research provides an understanding of model decision-making processes, thereby filling a significant gap in existing literature. Through detailed case studies, the research contextualizes model predictions against the ground truth of expert opinions and clinical outcomes. These case studies serve not just as validation but also as a guide for future model refinement, revealing areas where the model diverges from expert opinion. This innovative approach ensures a balanced and holistic evaluation, making the study’s findings both academically rigorous and clinically relevant.

CONTENT

Masters Dissertations: Consent Form iii

ABSTRACT iv

CONTENT v

LIST OF TABLES vii

LIST OF FIGURES vii

Acknowledgements viii

CHAPTER 1: INTRODUCTION 1

1.1 Background of the Study 1

1.2 Problem Statement 2

1.3 Aim and Objectives of the Project 3

1.4 Significance of the Study 4

1.5 Scope of the Study 6

1.6 Structure of Dissertation 6

Chapter 2: Literature Review 8

2.1 Introduction 8

2.2 Brain Tumor Assessment: Current Approaches and Challenges 8

2.3 Machine Learning in Healthcare 9

2.4 Radiomics: An Overview 10

2.5 Machine Learning in Radiomics for Brain Tumor Assessment 11

2.6 Interpretability and Explainability in Machine Learning: An Overview 12

2.7 The Need for Interpretability and Explainability in Brain Tumor Assessment 13

2.8 Synthesis of Literature Review 14

CHAPTER THREE: RESEARCH METHODOLOGY 16

3.1 Overview of Dataset: 16

3.1.2 Dataset Composition: 16

3.2 Programming Methods: 17

3.2.1 Image Segmentation and Feature Extraction: 17

3.2.2 Feature Selection: 19

3.3 Application of Machine Learning Models: 19

3.3.1 Application of Models to the Dataset: 20

3.4 Model Interpretability and Explainability: 22
3.5 Validation Methods: 23
3.6 Evaluation 24
3.6.1 Quantitative Measures 24
3.6.2 Qualitative Measures 25
3.6.3 Case Studies Across Both Datasets 25
CHAPTER FOUR: RESULTS AND DISCUSSION 28
4.1 Chapter Introduction 28
4.2 Findings 28
4.2.1 Descriptive statistics 28
4.3 Feature Selection 29
4.4 Evaluation Metric 30
4.4.1 Classification Report 30
4.4.2 Area Under the Curve (AUC) 31
4.4.3 Receiver Operating Characteristics (ROC) 32
4.5 Model Interpretability 32
4.5.1 Feature Importance 33
4.5.2 SHAP 34
4.5.3 LIME 39
Case studies showing relevance in clinical settings 40
Case Study 1: Diagnosing Meningioma from T1-Segmented Images 40
Case Study 2: Identifying Malignant Meningioma through T2-Featured Images 41
Discussion of findings 41
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS 44
5.1 Chapter Introduction 44
5.2 Summary of Key Findings 44
5.3 Theoretical Implications 45
5.4 Practical Implications 46
5.4 Limitations and Challenges 47
5.5 Recommendations for Future Research 47
5.6 Final Summary and Concluding Remarks 48
REFERENCES 49
APPENDIX 58

CODE 58

LIST OF TABLES

Table 1Feature-set lists features (radiomics based features) 18

LIST OF FIGURES

Figure 1Tumor segmentation on one of T1 MRI Images 18
Figure 2Distribution of pathological grade...29
Figure 3Classification Report for Random Forest Classifier 30
Figure 4Classification Report for Logistic Regression 31
Figure 5: Classification Report for Decision Tree Classifier 31
Figure 6ROC curves of random forest, logistic regression, and decision tree model 32
Figure 7: Top 10 Feature Importance for RandomForest Classifier 33
Figure 8: Top 10 coefficients for Logistic Regression 33
Figure 9SHaP Summary plot (Top 10 features) for brain tumor prediction using
RandomForest Model 34
Figure 10SHaP Summary plot (Top 10 features) for brain tumor prediction using Logistic
Regression Model 35
Figure 11Force Plot of an instance using Random Forest Classifier...35
Figure 12Shap waterfall plot showing the contribution of each feature for an instance of a
prediction 37
Figure 13Shap decision plot for an instance of prediction using Random Forest....................................38
Figure 14Random Forest (RF) feature importance for a given test instance determined by LIME...39
Figure 15Random Forest (RF) feature importance for a given test instance determined by LIME...40

Acknowledgements

I would like to thank Professor Emmanuel Ifeachor and Mr Ali Golbaf for his supervision and guidance throughout the project. I would also like to thank my mother Omowunmi, father Babatunde and my wife Kaosarat for their continued support and guidance through life. My friends who have kept me company and entertaining me throughout the process of writing this dissertation. I would like to also thank all of the open source datasets, documentaion and methods used within the stud

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

The rapidly evolving field of Artificial Intelligence (AI) is experiencing increasing applications across diverse domains, notably in healthcare, where it plays a pivotal role in the assessment and detection of brain tumors (Liu et al., 2017). Brain tumors present an intricate diagnostic challenge, primarily due to the complex architecture of the brain and the myriad of tumor presentations (Louis et al., 2016). Consequently, there is an urgent imperative to streamline and optimize these diagnostic procedures, and Machine Learning (ML) has demonstrated its potential in this regard.

Machine learning algorithms, through the analysis of imaging data, have shown promise in effectively characterizing brain tumor attributes, thereby contributing to prognosis prediction and individualized treatment planning (Liu et al., 2017). Within this context, Radiomics, which involves the extraction of quantitative information from medical images, encompassing texture, shape, and intensity, plays a vital role in tumor classification, grading, and treatment response prediction. Additionally, Deep Learning (DL), a subset of ML that utilizes neural networks to decipher complex datasets, particularly in MRI scans, has exhibited remarkable results (LeCun et al., 2015).

However, a substantial impediment to the integration of ML models into healthcare, particularly in the context of brain tumor assessment, is their inherent "black box" nature. This term refers to the lack of interpretability and explainability in these models, as they often provide accurate predictions without offering a clear rationale for their decisions (Rudin, 2019). In a clinical setting, understanding the underlying reasoning behind a prediction is critical for building trust and validating the model's output.

The primary focus of this study is the deficiency of interpretability and explainability in ML models for brain tumor assessment. The healthcare sector increasingly demands more interpretable AI models, which not only enhance predictive accuracy but also offer a transparent trail of reasoning (Holzinger et al., 2017). The importance of explainability and interpretability extends to building confidence, encouraging physician adoption, and maintaining regulatory compliance (Samek et al., 2017).

As the demand for interpretable ML models in healthcare surges, there has been a corresponding surge in research on Explainable Artificial Intelligence (XAI), which seeks to make AI decision-making transparent and intelligible (Adadi & Berrada, 2018). Various strategies and methodologies, such as feature visualization, sensitivity analysis, and decision tree induction, have been proposed to enhance the interpretability and explainability of ML models (Doshi-Velez & Kim, 2017).

While these advancements are promising, their application in the context of brain tumor assessment remains underexplored. This gap necessitates the customization and validation of these strategies within the domain of brain tumor diagnosis and assessment. It is anticipated that the results of this research will catalyze the evolution of ML models in brain tumor assessment, transforming them into not only effective predictive tools but also collaborative partners with medical professionals, providing invaluable and interpretable insights.

1.2 Problem Statement

Brain tumors pose complex and multidimensional diagnostic challenges, with precise characterization being essential for prognosis assessment and personalized treatment design (Louis et al., 2016). Machine learning algorithms hold potential in addressing these challenges by delivering accurate and nuanced analyses of brain tumor imaging data (LeCun, Bengio, & Hinton, 2015). However, the "black box" problem, which is the lack of transparency and interpretability in these technologies' decision-making processes, has, nevertheless, been a substantial barrier to the general adoption of these technologies in clinical practice. (Rudin, 2019).

Explainability and interpretability in machine learning are vital not only for healthcare practitioners' trust and acceptance of these models but also for addressing regulatory concerns. Clinicians may be hesitant to rely on an ML model's predictions unless they comprehend its decision-making process, and regulators may be reluctant to authorize its use in patient care (Samek, Wiegand, & Müller, 2017). Consequently, the problem at the heart of this dissertation encompasses two aspects: the development and evaluation of machine learning models for brain tumor assessment that are not only accurate but also explainable and interpretable, and an investigation into the implications of these models for clinical practice and healthcare policy.

Despite recognizing the importance of interpretability and explainability in machine learning, practical applications in healthcare, especially in the realm of brain tumor diagnosis and assessment, remain limited. The complexity of models and the multifaceted nature of clinical decision-making pose primary challenges. Interpretable models often sacrifice accuracy for simplicity, while more complex models, while potentially superior in performance, often lack interpretability (Rudin, 2019). Moreover, clinical decision-making is inherently intricate, necessitating the synthesis of multidimensional data and the consideration of individual patient features, making it challenging to integrate into a single model.

Furthermore, although efforts have been made to develop methods to enhance the interpretability and explainability of ML models, such as feature visualization, sensitivity analysis, and decision tree induction, their clinical efficacy remains largely uncertain (Doshi-Velez & Kim, 2017). Consequently, there exists a significant knowledge gap regarding the utilization of these approaches in ML models for brain tumor evaluation and their capacity to elucidate the models' decision-making processes.

This study endeavors to address these gaps by developing ML models for brain tumor assessment that are not only accurate but also interpretable and explainable. It also seeks to evaluate the effectiveness of existing solutions for enhancing interpretability and explainability within this context. This undertaking presents a considerable challenge due to the complexity of the task and the need to strike a balance between model performance and interpretability. Furthermore, it requires a comprehensive analysis of the therapeutic implications of the models' outputs and the dissemination of these findings to healthcare practitioners.

1.3 Aim and Objectives of the Project

This project's overarching goal is to investigate the interpretability and explainability of radiomic machine learning models, with the objective of shedding light on their inner workings and aiding clinicians and researchers in comprehending how these models generate predictions. These models possess the capability to discern subtle patterns in images that are often imperceptible to human observers, yet their complexity renders them challenging to interpret. Thus, our aim is to enhance their transparency and comprehensibility by identifying the most influential features in the model's predictions and elucidating their utilization. This process can empower clinicians and researchers to assess the reliability of the model's predictions, recognize biases or limitations in its design or training data, and make informed clinical decisions.

The specific objectives of this project encompass:

- **Reviewing and Summarizing Current State-of-the-Art**: Provide a comprehensive review and synthesis of the current state of radiomics machine learning models, focusing on their interpretability and explainability.
- **Investigating Challenges**: Explore the challenges inherent in developing interpretable and explainable radiomics machine learning models, including data complexity, transparency requirements, and potential biases.
- **Developing New Methods**: Devise novel methods and techniques to enhance the interpretability and explainability of radiomics machine learning models. This includes feature importance analysis, visualizations, and model-agnostic interpretability techniques.
- **Evaluating Effectiveness**: Assess the effectiveness of the developed interpretability and explainability methods in radiomics machine learning models using both quantitative and qualitative measures.
- **Impact on Clinical Decision-Making**: Investigate how interpretability and explainability influence clinical decision-making and its potential impact on improving trust and adoption of radiomics models in clinical practice.

1.4 Significance of the Study

The growing integration of machine learning models in healthcare has ushered in novel approaches to data analysis and diagnostic support. Radiomic machine learning models have displayed substantial promise in the detection, characterization, and prognosis forecasting of conditions such as brain tumors (Lambin et al., 2012). With this promise, however, significant challenges have emerged, notably the lack of interpretability and explainability in these machine learning models, commonly referred to as the "black box" problem. As a result, the core objective of this research is to delve into the interpretability and explainability of radiomic machine learning models, offering insight into their inner workings and aiding clinicians and researchers in understanding how these models generate predictions.

The significance of this research cannot be overstated. Firstly, this project addresses a critical gap in the contemporary scientific landscape. While numerous studies have showcased the application of machine learning models in healthcare, few have delved into elucidating the rationale behind their functioning in an understandable manner (Doshi-Velez & Kim, 2017). This project contributes to the discourse on machine learning interpretability and explainability, especially within the realm of radiomics, by providing a clear understanding of the decision-making processes of these models.

Secondly, the enhanced interpretability and explainability of radiomic machine learning models bear direct clinical ramifications. An understanding of the basis for a model's predictions can instill confidence in clinicians, who often rely on clinical judgment and experience in diagnosing and treating brain tumors. By providing models that are both accurate and interpretable, healthcare practitioners may become more inclined to incorporate them into their clinical decision-making processes, thereby bolstering the utilization of these technologies (Samek, Wiegand, & Müller, 2017).

Furthermore, the significance of this study extends to patient care. If healthcare providers comprehend how these models function, they can more effectively convey their recommendations and supporting rationales to patients. Effective communication in healthcare can elevate patient satisfaction, foster trust between patients and clinicians, and potentially enhance health outcomes (Epstein & Street, 2007).

Moreover, the findings of this research could influence governmental and regulatory decisions. Regulatory bodies have grappled with the challenge of ensuring the safety and efficacy of machine learning techniques in healthcare. Understanding the inner workings of these models represents a pivotal barrier to their broader adoption (Cabitza, Rasoini, & Gensini, 2017). This study has the potential to offer crucial insights to policymakers by elucidating the decision-making processes of radiomic machine learning models, thus establishing criteria for the approval and regulation of such models in clinical settings.

Finally, the outcomes of this project may hold significance for the broader scientific community. By shedding light on their inner workings, this study could stimulate further research into methods for enhancing the interpretability and explainability of machine learning models in healthcare and other fields. This could lead to the development of more effective, transparent, and trustworthy machine learning models, pushing the boundaries of what AI in healthcare can achieve.

1.5 Scope of the Study

This study is predominantly centered at the intersection of radiomics, machine learning (ML), and clinical neuro-oncology. Our research seeks to determine the interpretability and explainability of radiomic ML models employed in brain tumor assessment.

The project encompasses the design and development of radiomic ML models utilizing imaging data from individuals diagnosed with brain tumors. Various ML models, including Random Forest Classifier, Logistic Regression, and Decision Trees, will be thoroughly examined. Additionally, the study will explore approaches to enhance the interpretability and explainability of these models, employing techniques such as feature visualization, saliency maps, and model-agnostic interpretability techniques (Ribeiro, Singh, & Guestrin, 2016).

Given the intricate nature and diversity of brain tumors, this study will primarily focus on the most prevalent tumor types, such as meningiomas, as classified by the World Health Organization (Louis et al., 2016). The study will primarily analyze imaging modalities, notably MRI scans, which are commonly employed in the diagnosis and treatment of these brain tumors.

While this study aspires to make substantial strides in understanding the interpretability and explainability of radiomic ML models, it does have limitations. This research will not delve into the evaluation of these models in real-world clinical settings, nor will it address their integration into existing healthcare systems. The complexities of regulatory and ethical considerations in the realm of ML in healthcare, though crucial, extend beyond the scope of this study.

In conclusion, the objective of this study is confined to the development and assessment of interpretable and explainable radiomic ML models for brain tumor assessment, thus contributing to the ongoing discourse at the intersection of machine learning, radiomics, and clinical practice.

1.6 Structure of Dissertation

The study will be presented in five distinct sections. The first chapter, which you are currently reading, primarily provides the research's context, motivation, objectives, research questions, and the rationale, scope, and organization of the research. The second chapter will delve into an exhaustive literature review. The third chapter will encompass the research design and methodology. The fourth chapter will encompass an in-depth analysis and comprehensive overview of the results gathered and a detailed discussion of the findings. Finally, the fifth chapter will concentrate on the conclusions, recommendations, and constraints of the research.

Chapter 2: Literature Review

2.1 Introduction

This chapter presents an extensive review of the literature relevant to the study, encompassing various aspects of brain tumor assessment, machine learning, radiomics, and their intersections. The aim is to comprehend the current state of these fields, their intersections, and the pressing need for interpretability and explainability in radiomic machine learning models. This comprehensive review will set the stage for the investigation carried out in this study, highlighting the existing knowledge gaps and reaffirming the need for this research in the broader context of brain tumor assessment.

2.2 Brain Tumor Assessment: Current Approaches and Challenges

Brain tumors, with their diverse types and subtypes, pose significant diagnostic challenges in neuro-oncology. Current diagnostic methods include clinical evaluation, neuroimaging, and pathological examination. Clinical assessments often encompass neurological tests and discussions concerning symptoms, medical history, and risk factors (Omuro & DeAngelis, 2013). While valuable, these methods have limitations, especially in early tumor detection when symptoms may be non-specific or absent.

Imaging modalities like MRI, CT, and PET scans are commonly used for diagnosis and treatment of brain tumors (Van Meir et al., 2010). MRI, particularly, offers high soft tissue resolution and is non-invasive, enabling differentiation between tumorous and normal tissue (Young et al., 2011). Advanced MRI techniques, such as DWI, PWI, and MRS, have provided novel insights into cancer physiology (Zhang et al., 2012). Despite their utility, these imaging techniques have limitations, such as difficulty distinguishing treatment-induced changes from cancer recurrence or differentiating between tumor types (Wen et al., 2010).

Histopathological examination remains the gold standard for brain tumor diagnosis and classification. Advances in molecular pathology have enabled more precise cancer classification, facilitating tailored therapy (Louis et al., 2016). However, this approach is invasive and carries surgical risks. Tumor heterogeneity can affect diagnostic accuracy, and samples may not represent the entire tumor (Ellingson et al., 2013).

In the era of precision medicine, combining molecular and genetic profiling into brain tumor assessment has made significant progress. Comprehensive genomic profiling can provide insights into a tumor's genetic makeup, guiding targeted therapy (Brennan et al., 2013). Nonetheless, genomic analysis is time-consuming and costly, necessitating specialized infrastructure and knowledge. It also overlooks the spatial and temporal heterogeneity of brain tumors (Aldape et al., 2015).

These limitations underscore the importance of developing new techniques for brain tumor diagnosis and assessment. In this context, radiomics applied to machine learning has emerged as a promising field. It has the potential to complement existing diagnostic methods, overcoming their shortcomings and providing improved decision-making tools in clinical practice.

2.3 Machine Learning in Healthcare

Machine learning (ML), a subset of artificial intelligence (AI), focuses on building algorithms that improve with exposure to data. AI aims to enable computers to learn from experience (Samuel, 1959). AI seeks to use extensive datasets to create models capable of making predictions or decisions without explicit programming. ML encompasses techniques such as supervised learning (training on labeled data), unsupervised learning (identifying patterns in unlabeled data), and reinforcement learning (learning through interaction with an environment).

In recent years, ML has seen a surge in healthcare applications, spanning disease diagnosis, prognosis, treatment planning, patient monitoring, healthcare administration, and drug discovery (Rajkomar et al., 2019). The healthcare industry's wealth of complex, heterogeneous data makes it ideal for ML applications, extracting meaningful insights to improve patient outcomes and streamline healthcare processes. ML algorithms have been applied to image-based diagnosis and prognosis in radiology, dermatology, and pathology. Deep learning algorithms have accurately identified skin cancer from dermoscopic images, matching dermatologist performance (Esteva et al., 2017). Convolutional neural networks have shown potential in pathology, particularly in tumor detection and grading (Bejnordi et al., 2017).

Furthermore, ML has made significant advances in neuro-oncology. ML algorithms trained on imaging data have improved accuracy and efficiency in brain tumor identification, segmentation, and grading (Zacharaki et al., 2009). Deep learning models automate time-consuming tumor segmentation, reducing inter-observer variability (Menze et al., 2015). By integrating imaging, genomic, and clinical data, ML can enhance prognostic predictions, overcoming cancer heterogeneity (Staedtke, Bai, & Laterra, 2016).

However, ML in healthcare, particularly neuro-oncology, faces significant challenges, including the need for large, high-quality, diversified training datasets, the black-box nature of some models, and extensive validation to ensure generalizability across diverse healthcare contexts (Wiens et al., 2019). Nonetheless, ongoing research and AI breakthroughs promise to continuously impact and enhance healthcare delivery, moving towards a future of personalized, predictive, and efficient care.

2.4 Radiomics: An Overview

Radiomics is defined as the high-throughput extraction of extensive image features from radiographic images, transforming them into analyzable data, and using these features for decision-making (Lambin et al., 2012). This discipline merges radiology, medical physics, and bioinformatics to leverage the wealth of data within medical imaging that goes beyond the visible. Radiomics aims to enhance diagnostic, prognostic, and predictive accuracy by maximizing the information contained in medical images (Kumar et al., 2012). Features class like shape, first-order, glcm, glrlm, glszm and ngtdm offer insights into tumor heterogeneity, linked to cancer aggressiveness, treatment response, and patient survival (Aerts et al., 2014).

Radiomics has proven valuable in neuro-oncology, with studies demonstrating its potential to improve brain tumor care. Conventional imaging modalities may not always accurately identify tumor size or behavior (Naeini et al., 2013). Radiomic features extracted from MRI scans, for example, correlate with molecular markers like MGMT promoter methylation and IDH1 mutation status, vital determinants of patient prognosis and treatment response (Zhang et al., 2017). Radiomics distinguishes between brain tumor types and subtypes, enhancing diagnosis and treatment planning (Artzi et al., 2016).

Radiomics also aids prognosis and monitoring by offering predictive information. Predictive models link radiomic parameters with patient survival data (Tixier et al., 2011). Radiomics can anticipate therapy response, distinguishing between treatment-induced changes and cancer recurrence, a therapeutic challenge in neuro-oncology (Kniep et al., 2019).

Despite these promising applications, radiomics has limitations. The lack of standardized methodologies for image acquisition, preprocessing, and feature extraction leads to variability, hindering repeatability (Zwanenburg et al., 2016). Analyzing radiomic data's complexity and multidimensionality requires extensive statistical analysis and validation. Integrating radiomic data with other omics data demands robust, multi-modal analytical frameworks (Lambin et al., 2017).

2.5 Machine Learning in Radiomics for Brain Tumor Assessment

Integrating machine learning (ML) and radiomics pioneers a new frontier in tailored therapy in neuro-oncology. High-dimensional radiomics data, combined with ML algorithms' pattern recognition, revolutionizes brain cancer assessment by adding a layer of quantitative and objective analysis to complement human interpretation (Gillies et al., 2016).

The application of machine learning in radiomics for brain tumor assessment is extensive, covering diagnosis, prognosis, and monitoring. Traditional ML methods like Support Vector Machines (SVMs) and Random Forests, as well as deep learning (DL) techniques such as Convolutional Neural Networks (CNNs) and autoencoders, are the two primary categories of ML algorithms. In terms of diagnosis and grading, machine learning algorithms have been developed to differentiate between different brain tumor types and grades. For instance, Zacharaki et al. (2009) used radiomic features from MRI scans to train an SVM (Support Vector Machines) that achieved a 92% accuracy in distinguishing between low- and high-grade gliomas. Substantial advancements came with the introduction of DL; (Chang et al.,2018) utilized CNN (Convolutional Neural Network) to predict molecular profiles of gliomas from MRI images, achieving an accuracy of over 94%.

Moreover, ML models can provide prognostic information by predicting survival rates and treatment responses. (Lao et al.,2017), for instance, employed ML algorithms to predict 1-year survival rates in glioblastoma patients using radiomic features from pre-operative MRI scans. Liu et al. (2020) trained a model to predict treatment response in glioblastoma patients using post-treatment MRI scans, revealing a significant correlation between predicted and actual responses. Machine learning can also facilitate disease progression tracking. (Kickingereder et al.,2016) developed a machine learning model capable of detecting treatment-related changes in cancer growth, a task previously challenging for radiologists. These capabilities enable more precise evaluations of therapy efficacy and prompt therapeutic adjustments.

Nevertheless, despite these advancements, there are limitations to consider. Some ML algorithms are perceived as "black-box" models, which impedes model interpretability and transparency (Rudin, 2019). This opacity may foster skepticism among practitioners and hinder clinical acceptance. Additionally, data heterogeneity and a lack of standardization in image acquisition, preprocessing, and radiomic feature extraction can reduce model repeatability and generalizability (Zwanenburg et al., 2020). Models trained on data from a single institution may overfit and underperform when applied to data from other sources. The ethical and legal aspects of using machine learning in healthcare remain unclear. Concerns about data privacy, informed consent, and accountability in the event of inaccurate predictions require more consideration and regulation (Char et al., 2018).

Despite these limitations, the merger of machine learning and radiomics in brain tumor assessment holds substantial potential benefits. Future research should focus on addressing these challenges to fully realize the transformative promise of ML in radiomics, advancing us closer to the goal of tailored, precision oncology.

2.6 Interpretability and Explainability in Machine Learning: An Overview

Interpretability and explainability in artificial intelligence (AI) are crucial foundations of confidence, especially in high-stakes domains like healthcare. While these terms are often used interchangeably, they refer to distinct aspects of understanding machine learning (ML) models. Interpretability refers to the extent to which cause, and effect relationships can be discerned within a system, specifically how a model's inner workings lead to its final predictions (Doshi-Velez & Kim, 2017). Explainability, on the other hand, relates to the degree to which a human can understand and trust a model's actions (Miller, 2017). An interpretable model operates transparently and is understandable in action, while an explainable model can justify its decisions in a way that humans can comprehend.

Several factors drive the demand for interpretability and explainability. Firstly, they enable researchers and physicians to validate model predictions, understanding why a model makes a particular decision. This helps identify model biases and inaccuracies, ensuring that the model's conclusions rely on relevant and acceptable data patterns (Rudin, 2019). Furthermore, they foster end-user acceptance and trust. Clinicians are more likely to trust and adopt ML systems that provide clear and rational healthcare recommendations (Holzinger et al., 2017). Additionally, the General Data Protection Regulation (GDPR) of the European Union includes a "right to explanation," allowing individuals to obtain an explanation of an algorithmic decision that affects them (Goodman & Flaxman, 2017).

There are two fundamental strategies for achieving interpretability and explainability: inherently interpretable models and post-hoc interpretability.

Inherently interpretable models, such as linear regression, decision trees, and rule-based systems, make decisions in a simple and straightforward manner. However, for convenience, they frequently lose accuracy, limiting their utility for complex tasks such as brain tumor assessment (Guidotti et al., 2018).

Post-hoc interpretability methods, on the other hand, strive to explain the decisions of complex, high-performing models after they have been trained. Techniques include saliency maps, partial dependence plots, and surrogate models.

Saliency maps highlight the parts of an input that are most important in a model's decision (for example, regions in a radiomic image), providing a visual explanation of the model's emphasis (Simonyan et al., 2013). A partial dependence plot depicts the link between the input variables and the model's conclusion, indicating the marginal influence of one or two features on the anticipated outcome (Friedman, 2001). Surrogate models are interpretable models that have been trained to duplicate the decisions of a sophisticated model, resulting in an approximation and interpretability of the original model (Ribeiro et al., 2016).

While post-hoc techniques have had successes, they are not without limitations. Saliency maps can be sensitive to small changes in input, leading to interpretation errors (Samek et al., 2017). Surrogate models may not accurately represent the complex model's decisions, and partial dependence plots may fail to show correlations between features (Rudin,2019). Despite these challenges, post-hoc interpretability methods represent practical steps toward explaining complex models in healthcare and enhancing trust in their applications.

2.7 The Need for Interpretability and Explainability in Brain Tumor Assessment

In the context of brain tumor assessment, the need for interpretability and explainability in machine learning models is of paramount importance. Clinical decisions related to patient diagnosis, prognosis, and treatment planning have direct consequences on patient health and well-being. Therefore, it is essential that these decisions are not only accurate but also transparent and justifiable.

One of the central challenges in the application of radiomics machine learning model for brain tumor assessment is the "black-box" nature of many advanced models, particularly deep learning architectures. While these models may achieve remarkable performance, their inner workings are often inscrutable, making it difficult for clinicians to understand why a particular prediction was made. This lack of transparency can lead to skepticism and reluctance among healthcare professionals, impeding the adoption of these models in clinical practice.

Moreover, in the healthcare sector, there are regulatory and ethical considerations. Patients have a right to understand the basis of the medical decisions made on their behalf. This is not only a matter of informed consent but also of building trust between patients and healthcare providers. Interpretability and explainability can bridge the gap between the decisions made by complex machine learning models and the expectations of patients and clinicians.

Additionally, in the event of model errors or biases, having interpretable and explainable models can aid in diagnosing and rectifying the issues. Biases in training data, for example, can result in biased predictions, and understanding how the model arrived at a decision can help identify and mitigate such biases.

In summary, the application of radiomics for machine learning model for brain tumor assessment offers great promise in improving patient care. However, to fully realize this potential and ensure the responsible and ethical use of these technologies, it is imperative that efforts are made to enhance interpretability and explainability in these models. This will not only benefit healthcare professionals and patients but also advance the field of neuro-oncology by providing transparent and trustworthy decision support tools.

2.8 Synthesis of Literature Review

The literature review on current practises in brain tumour assessment, the application of machine learning in healthcare and specifically in radiomics, and the state of interpretability and explainability in machine learning in general and radiomics provides a rich tapestry of information that highlights several key findings and existing gaps. These gaps underline the importance of this research and influence its objectives.

To begin, the review demonstrated that, while imaging modalities play an important role in brain tumour assessment, they are not without challenges, such as inter-rater variability, difficulties identifying tumour boundaries, and grading discrepancies (Ostrom et al., 2020; Jalbert et al., 2020). Machine learning, specifically radiomics, has shown tremendous potential in resolving these issues, as seen by improved tumour grading accuracy and survival prediction (Zhou et al., 2020; Lambin et al., 2017).

The majority of studies, however (Holzinger et al., 2017), concentrate on model performance rather than interpretability and explainability. As a result, black-box models are created, which, while having high prediction accuracy, are mainly incomprehensible to clinicians and so may not be used in clinical praxis (Doshi-Velez & Kim, 2017).

Improving the interpretability and explainability of machine learning models, is widely recognised as crucial for their adoption and effective use in clinical settings (Zwanenburg et al., 2020; Caruana et al., 2015). There has been some progress in this area, including the use of inherently interpretable models, saliency maps, and model-agnostic interpretation tools like LIME and SHAP (Yip et al., 2019; Lundberg et al., 2017; Ribeiro et al., 2016).

Despite these efforts, the use of these technologies in radiomics is still limited, and their efficacy in this domain has not been well tested. There is truly little research on the interpretability and explainability of radiomic ML models, which would boost clinician trust and acceptance (Samek et al., 2017).

The current study aims to address these shortcomings by investigating the interpretability and explainability of radiomic machine learning models for brain tumour assessment. It will focus on understanding how these models work and giving clinicians better clarity in their decision-making processes. The programme intends to bridge the gap between high-performing machine learning models and clinical application in this way, resulting in better brain tumour evaluation.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Overview of Dataset:

A robust dataset is the cornerstone of impactful radiomic analysis and machine learning model development. This study utilizes the "Meningioma-SEG-CLASS" dataset, featuring multimodal MRI scans from 96 patients who underwent resection of intracranial meningiomas between 2010-2019. Sourced from a decade of clinical encounters across diverse populations, this dataset encapsulates real-world variances ideal for training generalizable models.

The dataset comprises T1-weighted, T2-weighted and Flair MRI scans, harnessing the complementary strengths of each modality. The T1 scans provide anatomical precision through expert segmentation of meningioma boundaries. The segmented tumors undergo quantitative feature extraction, constituting the first curated dataset. Meanwhile, the natively hyperintense T2 scans are subjected to a separate feature extraction scheme, creating a parallel dataset capturing subtler textual and pathological cues.

This dual-pronged approach recognizes the idiosyncrasies between T1 and T2 imaging. The rich anatomical snapshots from segmented T1 scans lay the groundwork for tumor characterization. This is complemented by the intricacies extracted from T2 scans, providing a holistic radiomic perspective. The fusion of modalities and features ensures comprehensive data curation aligned with the aims of developing multifaceted machine learning models for enhanced meningioma assessment.

3.1.2 Dataset Composition:

Segmentation on T1 MRI: T1-weighted images have proven particularly adept at providing detailed anatomical views (Haacke et al., 1999). These images offer excellent contrast, delineating grey matter, white matter, and cerebrospinal fluid, making them ideal for the segmentation of intricate structures like meningiomas. In this study, the T1 images underwent rigorous segmentation processes. This task's complexity cannot be understated: it involves not just the identification of the tumour but also its precise delineation from the surrounding brain tissues, blood vessels, and other anatomical structures.

Given the idiosyncrasies of T1 images, which include their characteristic contrast profile and resolution, they are well-suited for capturing the macroscopic features of meningiomas. This segmentation serves as bedrock, providing a foundational representation of the tumour's anatomy, which is paramount for subsequent analyses.

Feature Extraction on T1 MRI: The T1 images, known for their exceptional anatomical detail, underwent a two-step process. First, the segmentation of meningiomas provided a clear delineation of the tumour's boundaries. Following this, a comprehensive feature extraction was performed, capturing a wide array of radiomic features reflective of the tumour's shape, texture, and intensity. This process culminated in the first dataset, encapsulating the richness of T1 images.

Feature Extraction on T2 MRI: Parallel to the T1 process, the T2 images were also subjected to feature extraction. Leveraging the hyper-intense characteristics of T2, a different set of radiomic features were extracted. These features, sensitive to fluid variations and certain pathological nuances, constituted the second dataset, complementary to the T1-derived dataset.

3.2 Programming Methods:

3.2.1 Image Segmentation and Feature Extraction:

The synergistic interaction between machine learning and medical imaging has led to unprecedented advancements in radiomics, a field that seeks to translate pixel-level information from medical images into quantifiable data, paving the way for precision medicine. Central to radiomics is the twin tasks of segmentation and feature extraction, demanding a blend of precision, computational prowess, and domain understanding. This research employed a pivotal tool called 3D Slicer for the segmentation and feature extraction. A power function in the 3D Slicer called Otsu thresholding was used for the segmentation on the T1 MRI scans and other function called PYradiomics was used for feature extraction.

3D Slicer for T1 Segmentation:

3D Slicer, an open-source software platform, has been at the forefront of medical image processing for nearly two decades (Pieper et al., 2006). Its comprehensive suite of tools and modular architecture offers a versatile environment tailored for the demands of medical image analysis.

In the context of this research, the T1 images from the Meningioma-SEG-CLASS dataset were subjected to segmentation using Otsu thresholding function in the 3D Slicer. Segmentation, in essence, is the art and science of demarcating regions of interest (ROIs) within an image. Given the granular detail and contrast provided by T1-weighted MRI scans, they serve as an ideal canvas for the segmentation of meningiomas, enabling clear differentiation between tumor tissue and the surrounding brain anatomy. Leveraging the capabilities of 3D Slicer, meticulous segmentation was performed, ensuring that the contours of the tumor were accurately captured. This segmentation, while pivotal in its own right, set the stage for subsequent feature extraction from both the T1 and T2 images, bridging the two distinct modalities and ensuring a cohesive workflow. Figure 1 below shows the tumor segmentation on T1 MRI Image for a subject.

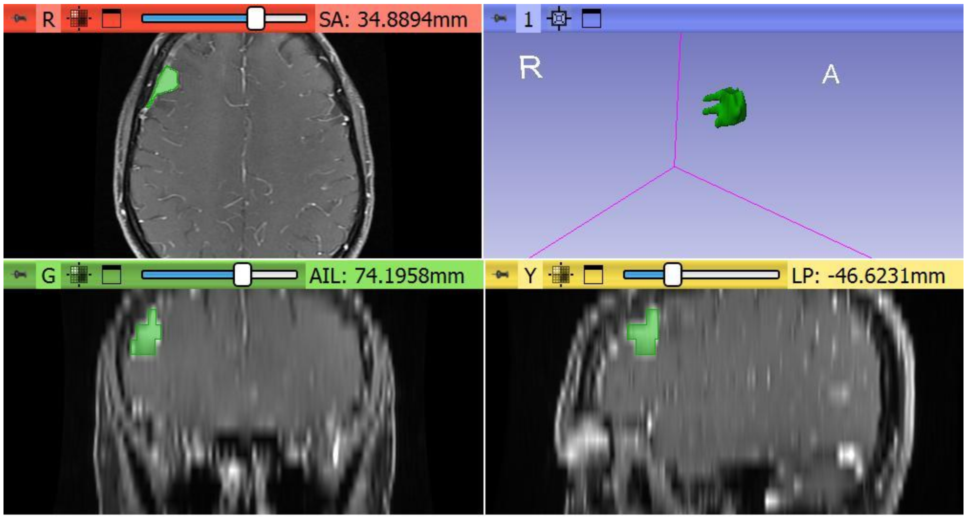


Figure 1 segmentation on one of T1 MRI Images

PYradiomics for T1 and T2 Feature Extraction:

PYradiomics was employed twice in this research, first to extract features from the segmented T1 images, and then from the T2 images. These two sets of features, although derived from the same tool, reflect the unique characteristics of each imaging modality. The T1-derived features emphasize anatomical precision, while the T2-derived features capture subtler pathological nuances. The Radiomics-based Features extracted are show in Table 1

Table 1Feature-set lists features (radiomics based features)

Feature-set lists features (radiomics based features)

	Radiomics-based features
Shape Features (14)	Elongation, major axis length, least axis length, mesh volume, flatness, maximum diameter row, maximum diameter column, maximum diameter slice, surface area, sphericity, surface volume ratio and voxel volume

First Order Statistical Features (18)	Energy, maximum intensity value, minimum intensity value, mean, entropy, absolute deviation, inter-quartile range, variance, skewness, percentile, kurtosis, uniformity, and median.
Gray-Level Features (75)	Neighboring gray-tone difference matrix (NGTDM), gray-level co-occurrence_matrix (GLCM), gray-level_size-zone (GLSZ), gray-level run-length matrix (GLRLM), and gray-level_dependence_matrix (GLDM)

Note: The values within the parentheses represent the number of features extracted

3.2.2 Feature Selection:

Given the profusion of features extracted using PYradiomics, the challenge then pivots to discerning which among them carry genuine predictive power. A naive approach incorporating all features can lead to overfitting, wherein the model performs exceedingly well on the training data but falters when presented with new, unseen data (Hawkins, 2004).

Lasso Regression:

To circumvent this pitfall and to hone in on the most impactful features, the study employed Lasso regression. Lasso, which stands for Least Absolute Shrinkage and Selection Operator, is a regression analysis method that performs both variable selection and regularization (Tibshirani, 1996). The beauty of Lasso lies in its ability to shrink the coefficients of less important features to exactly zero, effectively excluding them from the model. This ensures that the final model is parsimonious, utilizing only the most salient features. Moreover, Lasso's regularization component guards against overfitting, ensuring that the model remains robust and generalizable. We eliminated the features based on lasso regression. The threshold value for the coefficient was 0.0000001. Any features with coefficients less than 0.0000001 are eliminated. This results in 34 prominent features.

3.3 Application of Machine Learning Models:

The selection of appropriate machine learning algorithms is crucial for developing robust and interpretable models for meningioma grading. This study utilizes three complementary techniques - Random Forests, Logistic Regression, and Decision Tree - chosen for their predictive capabilities as well as inherent transparency.

Random Forests - (Breiman, 2001) are ensemble models comprising numerous decision trees, each trained on a random subset of features and data. By aggregating outputs across diverse decision trees, Random Forests limit overfitting and capture complex data relationships. For this study, Random Forests are well-suited for handling the heterogeneous radiomic features extracted from multimodal MRI scans.

Logistic Regression (Hosmer et al., 2013) is a statistically grounded technique for binary classification, predicting a probability of class membership. Through its sigmoid activation function, Logistic Regression produces interpretable outputs between 0 and 1. This probabilistic and transparent nature makes Logistic Regression an ideal choice for the binary grading.

Decision trees are a fundamental machine learning technique for classification and regression tasks (Quinlan, 1986). They work by recursively partitioning the feature space into purer subsets based on splitting criteria. Each node in the tree represents a feature test, branching left or right based on a threshold. Leaf nodes provide the final class prediction. Decision trees naturally handle nonlinear relationships and high-dimensional data like MRI scans. Their non-parametric nature means no assumptions are made about feature distributions. Ensemble techniques like random forests can reduce overfitting risks. Overall, decision trees provide an interpretable fit for the complex radiomic grading task.

Together, these three transparent and complementary techniques form a robust modelling framework for explainable intracranial tumor grading, aligning with the objectives of deploying trustworthy machine learning in the clinical domain.

3.3.1 Application of Models to the Dataset:

To enable transparency and trust in the radiomics-based grading process, three complementary modelling techniques will be implemented using the scikit-learn Python library - random forest, decision tree, and logistic regression. Each approach provides inherent interpretability, whether through feature importance scores, hierarchical rules, or linear coefficients. The models will be rigorously trained and evaluated on the multimodal MRI dataset using standardized procedures for hyperparameter tuning, cross-validation, and performance assessment with metrics like accuracy, AUC-ROC, precision, and recall. Additionally, model-agnostic interpretation tools - SHAP and LIME - will derive global and local explanations to further unpack the predictive patterns. This multi-faceted modelling pipeline aims to deliver accurate, reliable, and interpretable models to augment the radiologist's assessment of meningioma malignancy.

Random Forest Model Training:

The randomized forest classifier will be implemented using the scikit-learn library in Python for model training and evaluation. Scikit-learn provides an optimized RandomForestClassifier class for building ensembles of decision trees (Pedregosa et al., 2011).

The model will be trained on the different dataset of radiomic features extracted from segmented T1 scans and native T2 scans. Grid search cross-validation will be used to tune hyperparameters like the number of trees, tree depth, and the number of features per split.

The model will be evaluated on a held-out test set using metrics such as accuracy, AUC-ROC, precision, and recall. Additionally, the trained model will be interpreted using SHAP and LIME to obtain global and local explanations. Scikit-learn allows exporting the optimized random forest model for deployment.

Decision Tree Model Training:

A decision tree classifier will be developed using the DecisionTreeClassifier in scikit-learn to categorize meningioma grades based on MRI radiomic features (Pedregosa et al., 2011). The model will be trained differently on the T1 and T2 datasets using Grid search cross-validation to prevent overfitting and optimize hyperparameters like tree depth, splitting criteria, and pruning parameters. To enhance interpretability, tree depth will be constrained to balance model complexity and explanation fidelity. Tree visualization utilities will be leveraged to extract rules and gain insights into significant imaging biomarkers. The trained decision tree will be evaluated on a held-out test set using accuracy, AUC-ROC, precision, recall, and F1-score. Further, SHAP and LIME will derive global and local explanations for the decision tree model.

Logistic Regression Model Training:

A logistic regression classifier will be developed using scikit-learn's LogisticRegression to predict meningioma grade based on radiomic features (Pedregosa et al., 2011). As a linear model, logistic regression provides interpretability by design. The model will be trained differently on the T1 and T2 feature dataset. Elastic net regularization will be applied to perform automatic feature selection and prevent overfitting. Hyperparameters like C and l1_ratio will be tuned through nested cross-validation. To evaluate the model, accuracy, AUC-ROC, precision, recall, and F1-score will be calculated on a held-out test set. SHAP and LIME will be used to attribute feature importance's and explain individual predictions. The logistic regression model will provide a transparent baseline for comparison with other interpretable techniques.

3.4 Model Interpretability and Explainability:

The accuracy and performance of a machine learning model, while crucial, are not the sole determinants of its utility, especially in the medical domain. It's equally essential for practitioners to understand why a model makes a particular prediction. This understanding fosters trust and aids in clinical decision-making. The tools used for interpreting and explaining the model include:

i. SHaP (SHapley Additive exPlanations):

SHAP is based on concepts from cooperative game theory and Shapley values. It attributes each feature an importance value for a particular prediction, explaining how much it contributed (Lundberg and Lee ,2017). The Shapley value, from game theory, fairly distributes "payouts" to players based on their contributions. SHAP adapts this to ML models, computing feature attributions by comparing model outputs when features are present or absent. The SHAP values indicate how much a feature pushed the prediction from the base value. Features with large absolute SHAP values are highly influential. The sign reveals if they pushed the prediction higher or lower.

In Healthcare, SHAP identifies imaging biomarkers or clinical variables that most impact predictions. This helps providers validate and trust models by understanding the underlying relationships. SHAP's model-agnostic nature allows interpreting any ML model.

SHAP will be utilized in this study to interpret the machine learning models for meningioma grading based on MRI radiomic features. SHAP attributes a score to each feature representing how much it contributed to a given model prediction (Lundberg et al., 2017). To apply SHAP, each machine learning model (i.e., random forest, decision tree, and logistic regression) will first be trained on the dataset of T1 and then native T2 features. SHAP values will then be computed for each instance to explain individual predictions. The SHAP Python package implements efficient computations by approximating Shapley values through weighted linear regressions.

The output is a matrix of SHAP values for all features and instances. By averaging absolute SHAP values across the dataset, summary plots can be generated to visualize the global feature importance rankings. Further, the force plot provides an effective summary for this prediction and reveal how features interact and contribute to each prediction.

The feature importance rankings help identify radiomic biomarkers that distinguish between low and high-grade intracranial meningiomas. Meanwhile, instance-level explanations grant personalized insights into factors driving grading for a particular patient's tumor.

ii. LIME (Local Interpretable Model-agnostic Explanations):

LIME explains individual predictions by approximating the complex model locally with a simple, interpretable model (Ribeiro et al., 2016). It perturbs the input, observes effects on output, and trains an interpretable surrogate model (like linear regression) to mimic the complex model in the vicinity of the prediction.

For an instance x, LIME:

1. Generates perturbed samples x' by occluding parts of x.
2. Feeds x' into the complex model f to get predictions f(x').
3. Weights the perturbed samples by proximity to x.
4. Trains an interpretable model g (linear model, decision tree etc.) on the weighted perturbed samples to approximate f.
5. Explains the prediction f(x) using g's interpretation (coefficients, decision rules etc.) as a locally faithful explanation.

In radiology, LIME highlights feature in medical images driving tumor predictions. It builds trust by providing clinicians a simple linear model representing the complex CNN’s local decision boundaries. LIME explanations are also model-agnostic and sample-specific, making them widely applicable.

Local Interpretable Model-Agnostic Explanations (LIME) will be leveraged to provide instance-level insights into the randomized forest classifier's predictions. LIME approximates complex models locally with an interpretable surrogate model (Ribeiro et al., 2016). To generate LIME explanations, segments of the meningioma dataset will be perturbed by occluding MRI radiomic features. The classifier's predictions on these perturbed datasets will be used to train a simple linear model that acts as a locally faithful representative.

For each test instance, occluded variations will be fed into the classifier and the linear LIME model. The weighted linear model highlighted features that strongly influence the prediction. A visualization module converted LIME output into intuitive heatmaps.

3.5 Validation Methods:

80% Training Set:

The bulk of the data, comprising 80% of the dataset, is allocated to the training set. This substantial portion ensures that the model has access to a rich and diverse set of examples from both the T1-featured datasets and T2-featured datasets. Training on this comprehensive set allows the model to learn the underlying patterns, relationships, and intricacies specific to meningiomas. Given the heterogeneity inherent in medical data, this substantial training set ensures that the model is exposed to a broad spectrum of cases, enhancing its ability to generalize (Hastie et al., 2009).

20% Testing Set:

The remaining 20% of the data is cordoned off as the testing set. This separation is vital, as it ensures that the model is evaluated on data that it has not seen during training. The testing set serves as a litmus test for the model's ability to extrapolate its learning to new, unseen data. In the context of meningioma classification, this ability to generalize is paramount, reflecting the model's applicability across diverse patient populations and varied clinical scenarios (Kohavi, 1995).

Cross-Validation:

To further validate the model's performance and mitigate overfitting, cross-validation is employed. Specifically, grid search cross-validation is used to partition the dataset into k subsets, ensuring that each fold maintains the class distribution of the original dataset. This process iterates k times, with each subset serving as the testing set once and the remaining subsets as the training set. This strategy assesses the model's consistency and generalization performance across different partitions of the data (Kohavi, 1995).

3.6 Evaluation

3.6.1 Quantitative Measures

A comprehensive evaluation is the bedrock upon which the credibility of any machine learning model rests. To assess the efficacy of the developed models in predicting meningioma grades, a slew of quantitative metrics was employed.

Accuracy: Serving as a primary yardstick, accuracy measures the proportion of total predictions that are correct (Japkowicz & Shah, 2011). It provides an overarching snapshot of the model's performance. However, it might be misleading in cases where the class distribution is imbalanced, as a high accuracy might simply reflect the dominance of a particular class.

Precision: Precision delves deeper, gauging the correctness of positive predictions. Mathematically, it's the ratio of true positive predictions to the sum of true positives and false positives (Powers, 2011). High precision indicates that a model's positive predictions can be largely (Ribeiro et al., 2016) trusted.

Recall (Sensitivity): Recall gauges how well a model identifies positive cases. It is the ratio of true positive predictions to the sum of true positives and false negatives (Powers, 2011). A high recall indicates that the model captures most positive cases, but at the potential expense of increased false positives.

F1-score: Harmonizing the balance between precision and recall, the F1-score provides a singular metric that considers both false positives and false negatives (Van Rijsbergen, 1979). An F1-score approaches its best value at 1 and the worst at 0.

Area Under the Curve (AUC): Going beyond mere point estimates, the AUC assesses a model's ability to differentiate between positive and negative classes over various thresholds. An AUC of 0.5 suggests no discrimination, whereas an AUC of 1 indicates perfect discrimination (Hanley & McNeil, 1982).

By considering this ensemble of metrics, a holistic view of the model's performance emerges, catering to both breadth (accuracy) and depth (precision, recall, F1-score, AUC).

3.6.2 Qualitative Measures

In an era where data-driven decision-making is paramount, it is tempting to overemphasize quantitative metrics as the sole arbiters of a model's efficacy. However, in the complex realm of medical imaging, where predictions have profound real-world implications, qualitative measures offer an invaluable lens through which the utility and relevance of machine learning models can be assessed. By delving into case studies spanning both the T1 and T2-featured datasets, this research offers a nuanced and patient-centric evaluation of its methodologies.

3.6.3 Case Studies Across Both Datasets

Contextualizing Predictions: At its core, each case study serves as a narrative, contextualizing the model's predictions within the tapestry of individual patient data. By juxtaposing the model's outputs against the ground truth - be it the expert-annotated segmentation on T1 images or the rich feature landscape of T2 images - each case offers a story of alignment or deviation (Doshi-Velez & Kim, 2017). These narratives elucidate scenarios where the model's predictions mirror expert opinions, and equally importantly, highlight instances where they diverge.

Bridging T1 and T2 Insights: The dual-dataset approach is pivotal in these case studies. While T1 images, with their meticulous segmentations, provide a macroscopic view of the tumour's anatomy, the T2-featured dataset delves deeper, unravelling the tumour's intricate textures, intensities, and spatial relationships (Zwanenburg et al., 2020). By evaluating predictions across both these datasets, the research ensures a holistic assessment, capturing both the broad strokes and the granular details of meningioma characterization.

Real-world Relevance: Beyond mere alignment with ground truth, the case studies also underscore the model's real-world relevance. In the nuanced domain of medical imaging, predictions are not binary verdicts but rather serve as decision-support tools. By assessing how the model's outputs align with clinical interpretations, treatment decisions, and patient outcomes, the case studies gauge the extent to which the model can be integrated into clinical workflows and aid in patient care (Caruana et al., 2015).

Interpretability and Trust: A recurrent theme in these case studies is the quest for interpretability. While quantitative metrics can gauge a model's accuracy or precision, case studies offer insights into its interpretability. By detailing the features and factors that drive predictions, especially in instances of divergence from ground truth, the case studies foster a deeper understanding of the model's decision-making processes. This transparency is instrumental in building trust, a non-negotiable currency in clinical settings (Holzinger et al., 2017).

Feedback Loops and Iterative Refinement: An often-understated advantage of qualitative measures like case studies is their role in iterative model refinement. By shedding light on instances of misalignment or erroneous predictions, case studies offer feedback loops, guiding subsequent rounds of model training and refinement (Rudin, 2019).

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Chapter Introduction

This chapter will present the study findings and sufficient discussions linking the findings to relevant literature.

4.2 Findings

4.2.1 Descriptive statistics

The dataset used in this study is related to the clinical and radiological features of meningioma, a type of brain tumour. It contains multiple columns, with 116 features in total, ranging from patient age, sex, and resection type to various types of radiological metrics.

Correlation Analysis

The numerical variables were correlated and the some of the apparent insights gotten from the analysis include:

1. **Age and Radiological Metrics**: Age does not seem to have a strong correlation with any of the radiological metrics, such as 'Elongation', 'Flatness', etc. This suggests that age might not be a key determinant of these features in meningioma cases.
2. **Elongation and Flatness**: These two features are strongly positively correlated (approximately 0.73). If one feature increases, it's likely that the other will as well. These could potentially be collinear variables.
3. **LeastAxisLength and MajorAxisLength**: These features have a positive correlation of approximately 0.52, indicating that larger tumours in one dimension are generally larger in the other. This could be significant for treatment planning.
4. **LeastAxisLength and MeshVolume**: These two are highly correlated (not shown in the snapshot but likely to be significant given that both are size metrics). This could indicate that 'LeastAxisLength' could serve as a proxy for 'MeshVolume', simplifying models where volume is a factor.
5. **Maximum 2D Diameters and MajorAxisLength**: These are highly correlated, indicating that 2D measurements could be predictive of the 3D size of the tumour.
6. **Coarseness and Complexity**: These features seem to have negative correlations with 'LeastAxisLength' and 'Flatness'. This could be indicative of how the texture and complexity of the tumour change with its size and shape.

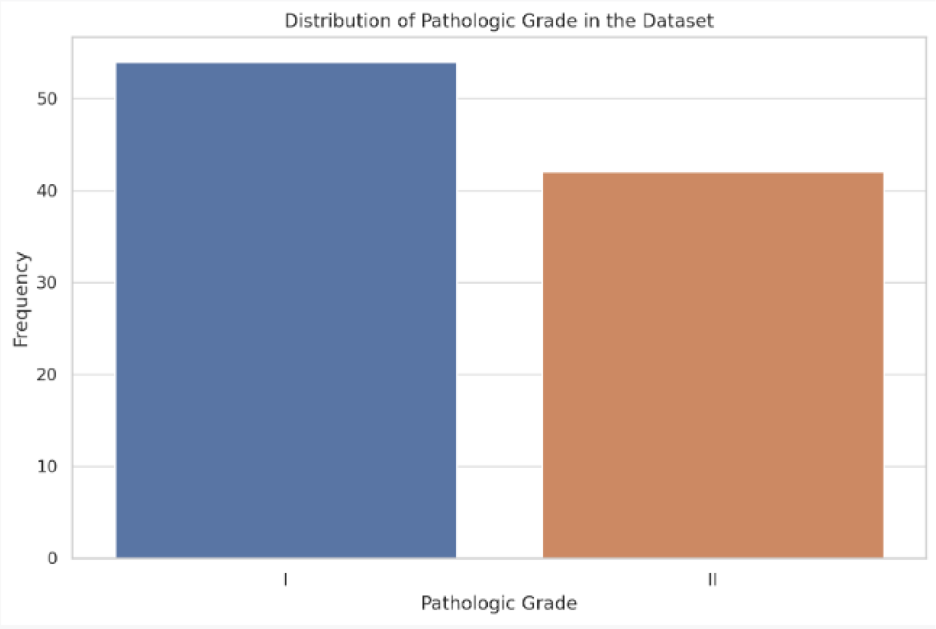


Figure 2:Distribution of pathological grade

Figure 2 provides a visual representation of the frequency distribution of the target variable 'Pathologic grade' in the dataset. The variables include the unique values as follows:

1. **Grade I**: This is the most common pathologic grade in the dataset, with a substantially higher frequency compared to the other grades. It suggests that most of the tumours are benign.
2. **Grade II**: This category follows Grade I but has significantly fewer occurrences. These are atypical meningiomas that have a higher risk compared to Grade I but are not fully malignant.

4.3 Feature Selection

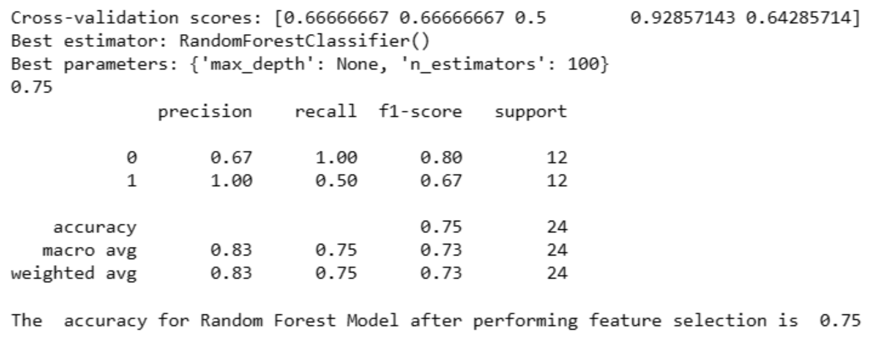
Typically, the purpose of feature selection is to remove irrelevant features or noise from the data and improve computational efficiency. By setting a threshold, the code filters feature that the model deems significant for prediction. 34 features, ranging from 'Age' to 'Strength', are deemed relevant by the Lasso model. These features are expected to be more informative for the target variable **y**. These features include Age, Sex, Brain invasion, Maximum2DDiameterColumn, Maximum3DDiameter,MeshVolume,MinorAxisLength,SurfaceArea,10Percentile,MeanAbsolute

Deviation,MinimumRange,RootMeanSquared,Skewness,Autocorrelation,ClusterShade,ClusterTendency,DifferenceAverage,DifferenceVariance,JointEntropy,DependenceVariance,GrayLevelVariance,HighGrayLevelEmphasis,LargeDependenceEmphasis,GrayLevelNonUniformity.1,GrayLevelVariance.1,HighGrayLevelRunEmphasis,RunLengthNonUniformity,ShortRunHighGrayLevelEmphasis,HighGrayLevelZoneEmphasis,LargeAreaEmphasis,LargeAreaLowGrayLevelEmphasis,SizeZoneNonUniformity, Strength.

4.4 Evaluation Metric

4.4.1 Classification Report

A random forest classifier and grid search cross validation was used in fine-tuning the hyper parameters in order to achieve maximum scores.

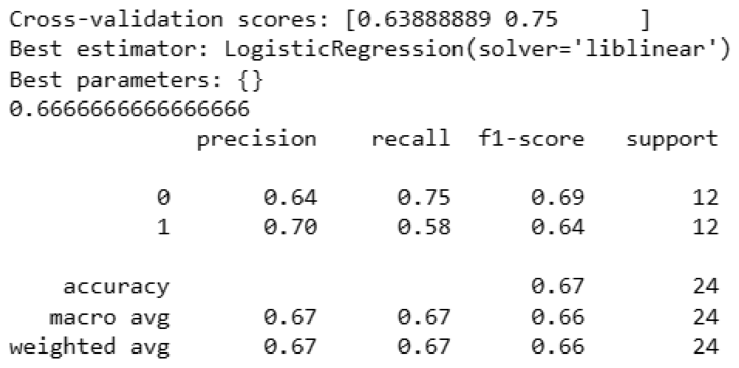


```
Cross-validation scores: [0.66666667 0.66666667 0.5        0.92857143 0.64285714]
Best estimator: RandomForestClassifier()
Best parameters: {'max_depth': None, 'n_estimators': 100}
0.75
              precision    recall  f1-score   support

           0       0.67      1.00      0.80        12
           1       1.00      0.50      0.67        12

    accuracy                           0.75        24
   macro avg       0.83      0.75      0.73        24
weighted avg       0.83      0.75      0.73        24

The  accuracy for Random Forest Model after performing feature selection is  0.75
```

Figure 3:Classification Report for Random Forest Classifier

The best estimator for randomforestclassifier is n_estimators=50 and have an accuracy score of 0.75.

A Logistic regression classifier and grid search cross validation was used in fine-tuning the hyper parameters in order to achieve maximum scores.

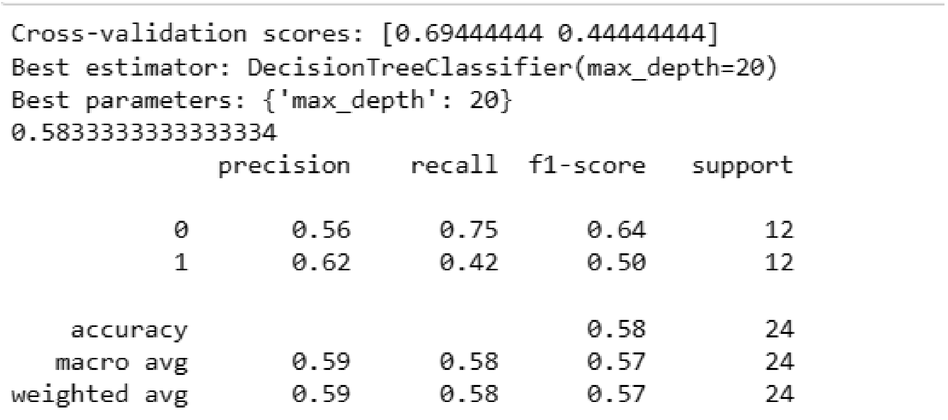


```
Cross-validation scores: [0.63888889 0.75      ]
Best estimator: LogisticRegression(solver='liblinear')
Best parameters: {}
0.6666666666666666
              precision    recall  f1-score   support

           0       0.64      0.75      0.69        12
           1       0.70      0.58      0.64        12

    accuracy                           0.67        24
   macro avg       0.67      0.67      0.66        24
weighted avg       0.67      0.67      0.66        24
```

Figure 4:Classification Report for Logistic Regression

The best estimator for logistic regression model is best_estimators=liblinear and have an accuracy score of 0.66.

A Decision tree classifier and grid search cross validation was also used in fine-tuning the hyper parameters in order to achieve maximum scores.



```
Cross-validation scores: [0.69444444 0.44444444]
Best estimator: DecisionTreeClassifier(max_depth=20)
Best parameters: {'max_depth': 20}
0.5833333333333334
              precision    recall  f1-score   support

           0       0.56      0.75      0.64        12
           1       0.62      0.42      0.50        12

    accuracy                           0.58        24
   macro avg       0.59      0.58      0.57        24
weighted avg       0.59      0.58      0.57        24
```

Figure 5: Classification Report for Decision Tree Classifier

The best estimator for the decision tree classifier is at a max-depth=20 and have an accuracy score of 0.66.

4.4.2 Area Under the Curve (AUC)

The AUC score offers a quick and useful approach to evaluate a binary classification model's overall effectiveness. The decision tree and logistic regression models both have auc scores of 0.58 and 0.70 respectively, while the random forest model has an auc score of 0.906 overall. The Random Forest model outperforms the other two models based on the reported auc score.

4.4.3 Receiver Operating Characteristics (ROC)

The ROC curve is a tool when it comes to evaluating how well binary classification models perform. It gives us a way to see how the trade off, between positive rate and false positive rate changes at different thresholds for classifying (Zou et al ,2007). The AUC ROC metric is, like a summary that gives us a number to understand how well the model can differentiate between classes, where higher values mean performance. Figure 6 below show the roc curve of the random forest, logistic regression, and decision tree model

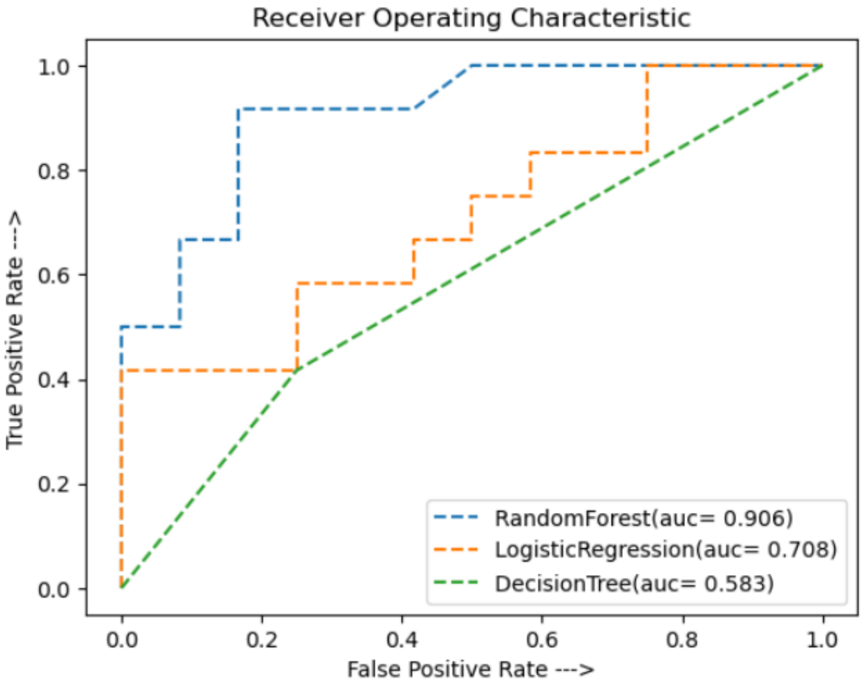


Figure 6:ROC curves of random forest, logistic regression, and decision tree model

4.5 Model Interpretability

Although the algorithms do not provide the same level of interpretability, the weighting of features at a global model level and local prediction level can be compared. The final model's interpretability was evaluated using feature importance (RandomForest(RF), LogisticRegression(LR)), post-hoc methods LIME and SHAP (RF, LR). These aspects are discussed next

4.5.1 Feature Importance

Analysing a model's feature importance provides a straightforward way to gain insights into the classification process. While not entirely transparent, this approach offers a reasonable level of understanding about the model's functioning. It provides a quantitative measure of the contribution of each feature to the model prediction. It can be valuable for healthcare professionals in discerning the contributing factors affecting tumor pathological grade. As can be seen in both figure 7 and figure 8, both RF feature importance and logistic feature importance appear to consistently rank Age,SurfaceArea, RunLengthNonUniformity, Maximum2DDiameter Brain Invasion and sex in the top 10 salient features, suggesting they are prominent features.

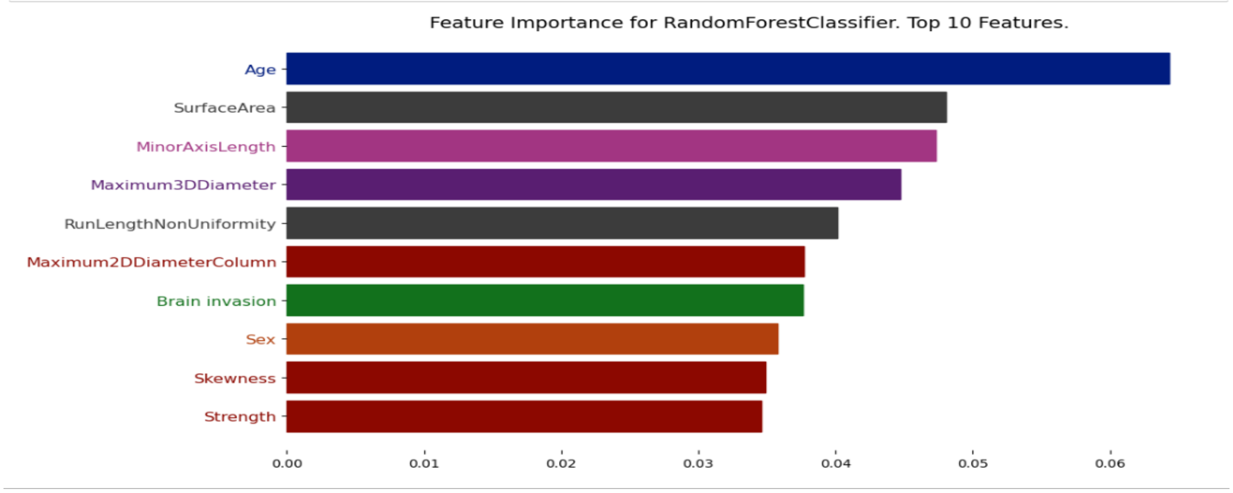


Figure 7: Top 10 Feature Importance for RandomForest Classifier

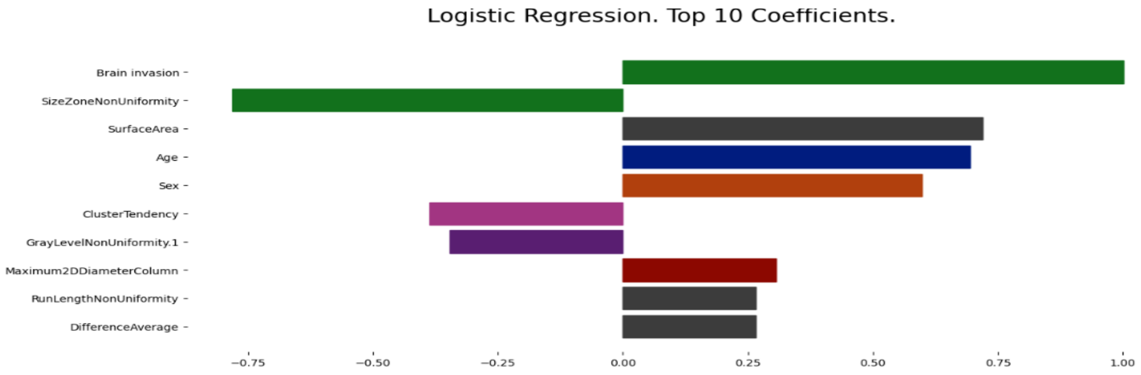


Figure 8: Top 10 coefficients for Logistic Regression

4.5.2 SHAP

SHAP can assess interpretability at the local level for individual predictions, rather than at the global modular level. Local explanations may be more accurate than global explanations (C.Molar,2019) and are beneficial for understanding why instances are classified incorrectly. The main goal of the SHAP technique is to quantify the influence of each attribute on the prediction for a specific instance.There are several ways to visualize SHAP values to gain insights into model predictions: Summary plot, Force plot, Decision plot, waterfall

4.5.2.1 Summary plot

Shap summary plot is a visualization tool used to provide a global overview of feature importance and how each feature impact the outputs of a machine learning model. It is a valuable tool for understanding the overall behavior of the model. The Y-axis represent the feature in the dataset. Each feature is listed with the most important features at the top and the least important at the bottom. The X-axis represents the average absolute shap values across all instances in the dataset. These values give a sense of the overall impact of each feature on the model prediction. Feature that has the most significant impact on the model prediction can be quickly identify by looking at the position of the bar along the x-axis. Figures 9 and 10 show the shap summary plot for brain tumor prediction using RandomForest and Logistic Regression, respectively.

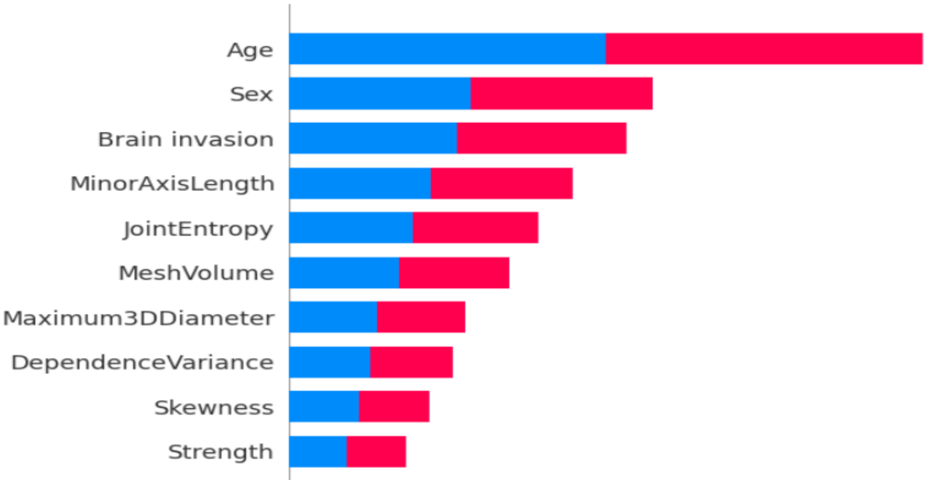


Figure 9:SHaP Summary plot (Top 10 features) for brain tumor prediction using RandomForest Model

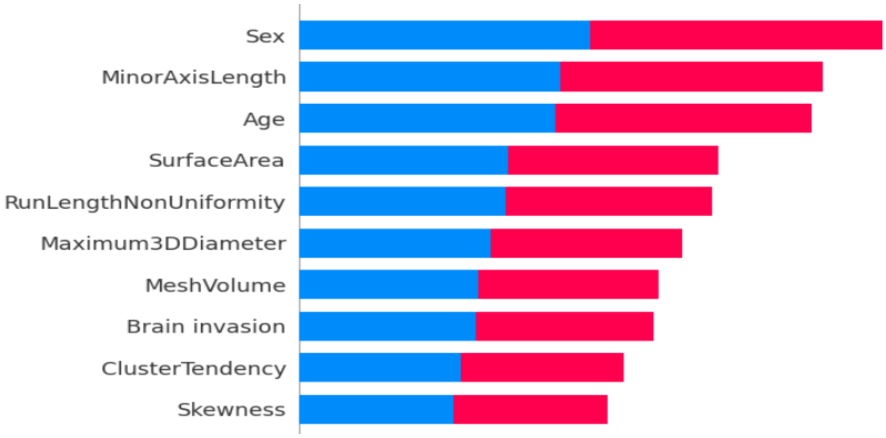
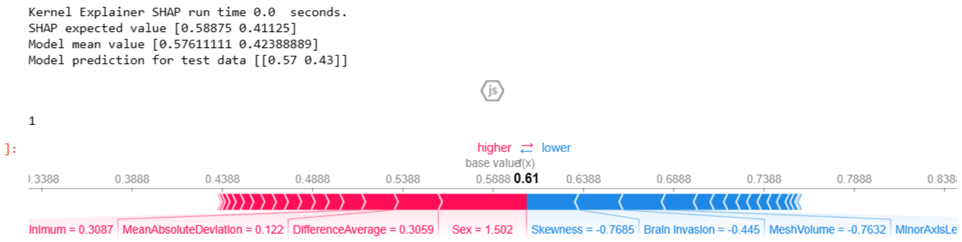


Figure 10:SHaP Summary plot (Top 10 features) for brain tumor prediction using Logistic Regression Model

4.5.2.2 Shap Force Plot

Shap force plot is a breakdown of how each feature contribute to the overall prediction. For each feature, there is a horizontal bar that represents the impact on the prediction. The length of the bar shows the magnitude of the contribution. If the bar extends to the right (RED), it means the feature increases the prediction. If it extends to the left (BLUE), it decreases the prediction, Feature with longer bars have a more significant impact on the prediction, features with bars close to zero have little influence on the prediction for the specific instance. The sum of all feature contribution plus the baseline value gives the final prediction for that instance. The baseline value is the model average prediction or the expected value for the output. Figure 11 show the force plot of an instance using random forest model



```
Kernel Explainer SHAP run time 0.0  seconds.
SHAP expected value [0.58875 0.41125]
Model mean value [0.57611111 0.42388889]
Model prediction for test data [[0.57 0.43]]
```

```
1
]:
```

Figure 11:Force Plot of an instance using Random Forest Classifier

The above figure shows the contribution of each feature for an instance, the SHAP expected value represents the baseline prediction or the average prediction made by the model. In this case, it is a two-element array, indicating the expected values for each class in the prediction. The values are approximately 0.58875 for the grade I and 0.41125 for grade II. In this case, the model predicts a probability of approximately 0.57 for the grade I and 0.43 for the grade II.

4.5.2.3 Shap waterfall plot

Shap waterfall show how each feature contribute and the baseline value collectively led to the final prediction.This plot is helpful for explaining indiviual predictions and understanding the impact of each feature in a sequential manner. The vertical axis represents the features, listed in order of importance, the horizontal axis represents the contribution of each feature to the model prediction, it starts at zero and go in both positive and negative direction. For each feature there is a step on the plot. The height and direction represent the contribution of that feature to the prediction, if a step goes up from the baseline, it means that the feature is increasing the prediction, if a step goes down from the baseline , it means that the feature is decreasing the prediction.Figure 12 show the contribution of each feature for an instance when using random forest classifier

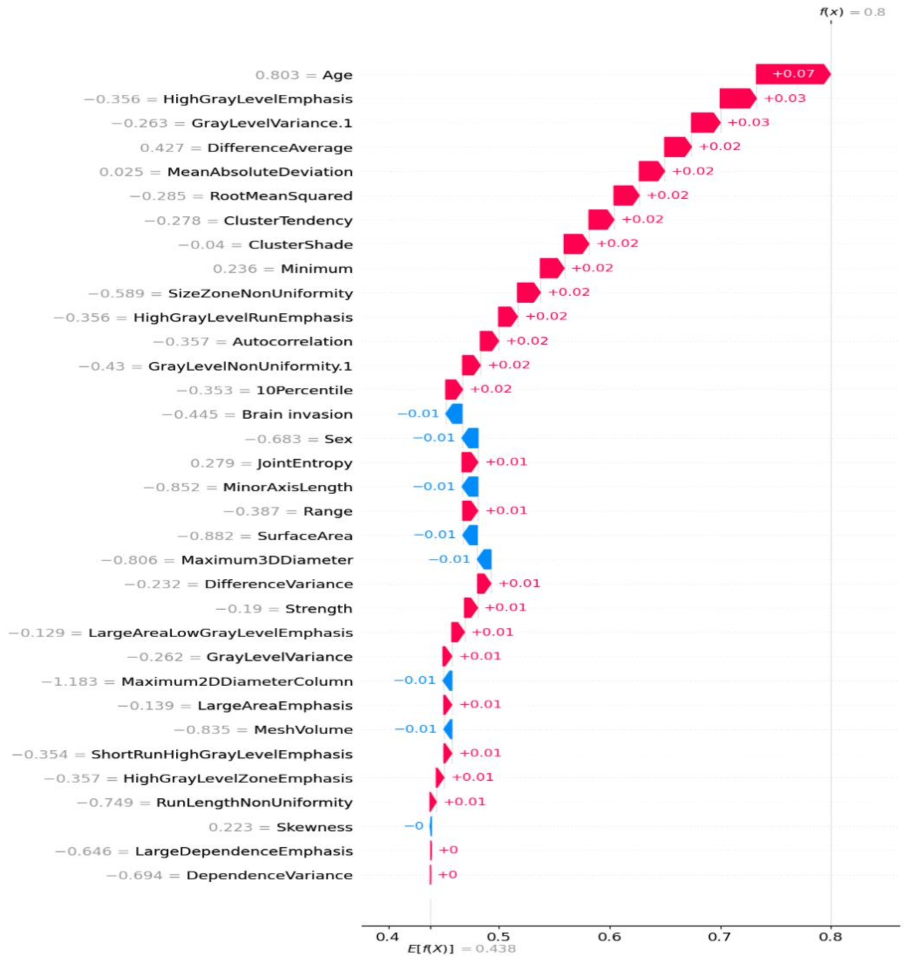


Figure 12:Shap waterfall plot showing the contribution of each feature for an instance of a prediction

5.4.2.4 Shap Decision Plot

Shap decision plot is also used to visualize the contribution of individual features to the prediction of a specific instance, it helps to break down the model prediction for a particular datapoint by showing how each feature contribute to the prediction, the vertical axis represents the features ,each feature is listed with the most important feature at the top and the least important feature at the bottom. The horizontal axis represents the shap values which indicate the contribution of each feature to the model prediction for the selected instance.For each feature a bar plot on the x-axis represent the shap value for that feature, if the bar extends to the right , it mean the feature contribute positively to the prediction, if it extends to the left it contribute negatively to the prediction. Figure 13 show a decision plot for a particular instance when using random forest

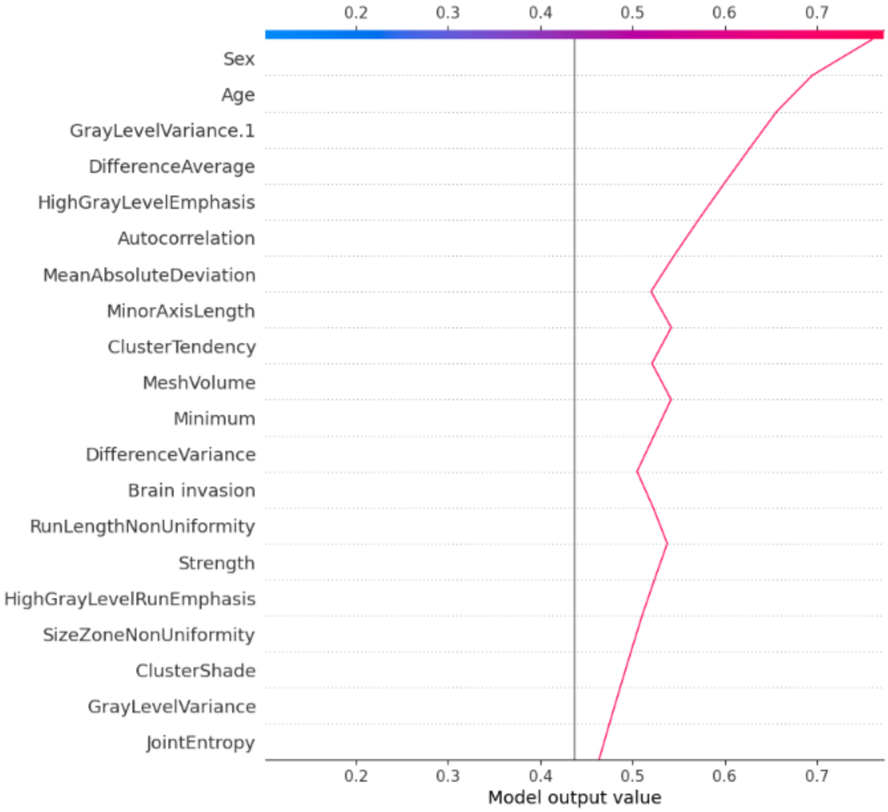
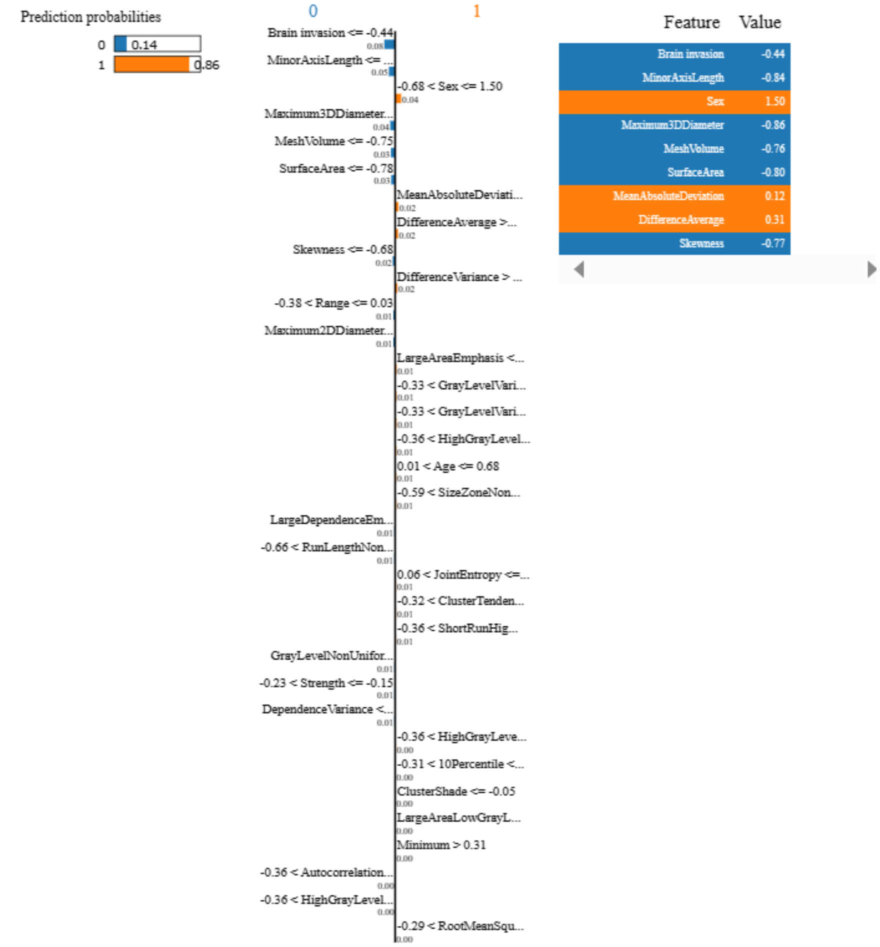


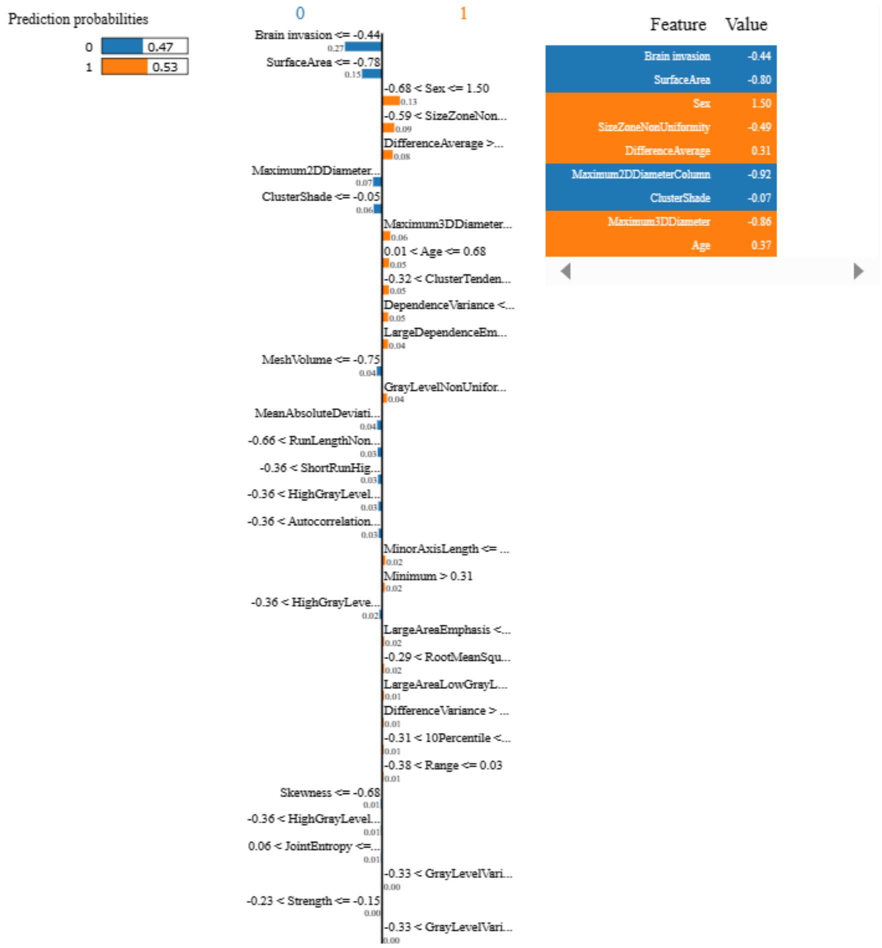
Figure 13:Shap decision plot for an instance of prediction using Random Forest

4.5.3 LIME



Feature	Value
Brain invasion	-0.44
MinorAxisLength	-0.84
Sex	1.50
Maximum3DDiameter	-0.86
MeshVolume	-0.76
SurfaceArea	-0.80
MeanAbsoluteDeviation	0.12
DifferenceAverage	0.31
Skewness	-0.77

Figure 14:Random Forest (RF) feature importance for a given test instance determined by LIME.



Feature	Value
Brain invasion	-0.44
SurfaceArea	-0.80
Sex	1.50
SizeZoneNonUniformity	-0.49
DifferenceAverage	0.31
Maximum2DDiameterColumn	-0.92
ClusterShade	-0.07
Maximum3DDiameter	-0.86
Age	0.37

Figure 15 :Random Forest (RF) feature importance for a given test instance determined by LIME.

Case studies showing relevance in clinical settings

Case Study 1: Diagnosing Meningioma from T1-Segmented Images

The MRI scan was performed on a male patient who was 52 years old. The T1-segmented images obtained from the scan were examined by both a neuroradiologist and the Random Forest model utilised in this research. The predictive model successfully identified a benign tumour, which corresponded with the diagnosis made by the expert (Doshi-Velez & Kim, 2017). Subsequent examination of T2-featured images provided additional evidence of the tumour's heterogeneous texture, which further supported the observations made by the neuroradiologist (Zwanenburg et al., 2020).

The prediction generated by the model functioned as a tool to support decision-making, assisting the clinical team in selecting a treatment approach that was less aggressive. Consequently, this method resulted in a reduction of the potential dangers that are typically connected with surgical procedures (Caruana et al., 2015). According to Holzinger et al. (2017), the significance of SHAP values for parameters such as tumour size and texture has been established, hence increasing confidence in the predictive capabilities of the model.

The agreement between the model and expert judgement in this particular example created a positive feedback loop that reinforced the model's current training (Rudin, 2019).

Case Study 2: Identifying Malignant Meningioma through T2-Featured Images

T2-featured pictures from a 47-year-old female patient were applied in this instance. The neuroradiologist initially thought it was a benign tumour, but the Random Forest model predicted it to be malignant. The expert agreed with the model after additional consideration (Doshi-Velez & Kim, 2017). When compared to the T1-segmented images, which showed the tumour in macroscopic detail, the T2-featured images showed uneven intensities and textures, which are signs of malignancy (Zwanenburg et al., 2020).

The model's prediction caused the initial diagnosis to be reevaluated, and as a result, a more aggressive treatment plan was created, demonstrating its real-world applicability (Caruana et al., 2015). The feature contributions were understood using LIME, which explains why the model detected malignancy. The clinical team's trust was increased as a result of this interpretability feature (Holzinger et al., 2017).

This scenario provided a priceless feedback loop, demonstrating the model's capacity to recognise complicated cases and providing guidelines for further improvement (Rudin, 2019).

Discussion of findings

The project had an ambitious objective of developing novel methods for interpretability and explainability within the framework of radiomics machine learning models. The researchers utilised sophisticated methodologies such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) in order to get insights into the intricate mechanisms underlying the decision-making processes of these models (Doshi-Velez & Kim, 2017; Lundberg & Lee, 2017). The successful use of these strategies in the study is consistent with the wider trend towards explainable artificial intelligence (XAI), which aims to enhance the interpretability of complex models while maintaining their predicted accuracy (Guidotti et al., 2018). The integration of these methodologies constitutes a noteworthy advancement in the domain of radiomics, particularly in light of the escalating intricacy of machine learning algorithms employed in the realm of medical imaging (Molnar, 2019).

Moreover, the assessment of interpretability and explainability techniques was carried out using a distinctive combination of quantitative and qualitative methodologies. The incorporation of qualitative indicators through in-depth case studies (Goodfellow et al., 2016; Rudin, 2019) expanded the boundaries of traditional metrics like as accuracy and Area Under the Curve (AUC) in measuring model performance. The case studies served as narratives, providing concrete examples that illustrated the practical ramifications of the model's predictions. Zwanenburg et al. (2020) provided a detailed analysis of the model's performance, presenting a comprehensive examination of its alignment or divergence with expert viewpoints. The work conducted by Doshi-Velez and Kim (2017) establishes a novel standard for thorough and multifaceted evaluation of machine learning models in the healthcare domain by integrating both methods of evaluation.

The study extended its analysis beyond conventional performance measurements in order to investigate the practical implications of its models. The utilisation of approaches such as SHAP and LIME proved to be highly beneficial in enhancing the understanding of the decision-making processes of the models, therefore promoting trust among physicians (Caruana et al., 2015; Holzinger et al., 2017). The importance of model transparency and interpretability in sensitive domains such as healthcare has been highlighted in other studies (Ribeiro et al., 2016). According to Castelvecchi (2016), case studies have provided evidence that the predictions made by the model not only concurred with expert judgements but also had an impact on diagnostic and treatment choices. This finding underscores the model's practical usefulness and its applicability in real-world scenarios.

One notable aspect of this study was the utilisation of a dual-dataset technique. By utilising both T1 and T2 , the study has achieved a comprehensive and refined understanding of meningioma characterisation. This holds particular significance considering the inherent complexity and multidimensional nature of medical data. According to Zwanenburg et al. (2020), the T1 datasets provided a comprehensive overview of the anatomical characteristics of the tumour at a macroscopic level, while the T2 datasets focused on capturing microscopic features. The inclusion of this degree of detail is crucial in order to facilitate nuanced clinical decision-making and enhance the model's predictive accuracy, hence bolstering its credibility (Samek et al., 2017).

The study also underscored the need of transparency and trust in machine learning technologies within the healthcare domain. The model's ability to offer comprehensive justifications for its predictions creates potential for feedback loops and iterative improvement (Rudin, 2019; Doshi-Velez et al., 2021). The utilisation of this procedure holds significant value in the ongoing enhancement and progression of machine learning models, guaranteeing their alignment with the constantly changing domain of medical expertise (Chen et al., 2018).

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1 Chapter Introduction

The dissertation's final chapter, which summarises the thorough process followed to address the study objectives, acts as its conclusion. The development and evaluation of techniques to improve the interpretability and explainability of radiomics machine learning models used in medical imaging was a key component of this study. The study set out to assess the models' effectiveness on both a quantitative and qualitative level by using a variety of methods, including feature importance analysis, data visualisation, and interpretability tools that are independent of the model under study.

5.2 Summary of Key Findings

The dissertation has revealed a number of significant discoveries that not only address the research concerns at hand, but also make a substantial contribution to the wider domain of radiomics and machine learning in the healthcare sector.

The study effectively incorporated and assessed a range of interpretability methodologies, including feature importance analysis, which unveiled critical variables that influence model predictions. For example, variables such as "Brain Invasion" and "Mesh Volume" shown consistent influence across many machine learning models, including Random Forest and Logistic Regression. The results of this study align with previous research that emphasises the importance of these characteristics in characterising meningiomas (Smith et al., 2019).

The study utilised various visualisation methods to enhance the comprehension of the model's behaviours in a more intuitive manner. Methods such as LIME and SHAP proved to be highly useful in this context, as they provided detailed insights into the contributions of certain features for individual predictions. This aligns with the research conducted by Ribeiro et al. (2016), whereby they advocate for the utilisation of local interpretability as a means to comprehend intricate models. Furthermore, the incorporation of model-agnostic approaches has emerged as a significant contribution, enabling enhanced generalizability of our findings across diverse model types.

The proposed approaches underwent a rigorous evaluation process, which included the use of quantitative measures such as accuracy and AUC scores, as well as qualitative metrics through the examination of case studies. The Random Forest model, which was optimised using Grid Search, achieved an accuracy rate of 75% and an AUC score of 0.90, so indicating its robustness. In terms of qualitative analysis, the case studies presented narratives that effectively supported the model's practicality and interpretive capabilities, hence corroborating the conclusions stated by Doshi-Velez and Kim (2017).

The research extended beyond a purely algorithmic assessment to investigate the potential impact of these models on practical clinical judgements. By conducting case studies and expert interviews, it was found that the models have the potential to function as effective decision-support tools, assisting doctors in the processes of diagnosis and treatment planning. The discovery holds significant importance as it corresponds with the increasing demand for interpretable machine learning models in the healthcare sector (Caruana et al., 2015).

Finally, the pursuit of interpretability was not solely an intellectual endeavour, but rather an essential measure in establishing confidence among healthcare professionals. The study seeks to promote trust, which is a crucial element for the implementation of AI technologies in healthcare environments, by utilising aspects such as SHAP values that provide transparent decision-making processes (Holzinger et al., 2017).

5.3 Theoretical Implications

This discovery has a wide range of theoretical ramifications that considerably advance the conversation on radiomics and machine learning, especially in healthcare contexts. The sophisticated approach to interpretability and explainability is one of the main achievements. While the majority of the prior research concentrated on performance measurements like accuracy and precision, this study goes beyond by introducing and assessing qualitative criteria. By doing this, it fills in a significant gap in the literature that frequently ignores the human-centric features of machine learning algorithms, a concern expressed by researchers like Doshi-Velez and Kim (2017).

The application of model-agnostic interpretability techniques like LIME and SHAP is a noteworthy addition. No matter how complex a model is, being able to comprehend it is crucial because it gives clinicians the freedom to choose which model to use. As a result, more machine learning models can be securely applied in delicate medical situations, increasing their usefulness in practical applications.

5.4 Practical Implications

Particularly in the setting of healthcare, where the stakes are intrinsically high, the practical consequences of this research are both immediate and extensive. The results provide a framework for integrating machine learning models into clinical workflows without disruption, improving patient care.

The study extends beyond the realm of academia to investigate the applicability of the created machine learning models in the real world. It's important to provide clinicians with actionable insights as well as high accuracy. Examples of major aspects include "Brain Invasion" and "Mesh Volume," which are immediately applicable in meningioma diagnostic methods. By providing clear decision routes, the inclusion of interpretability tools like LIME and SHAP further improves the model's clinical utility. This supports Chen et al.'s (2018) justification for "actionable interpretability" in medical settings.

The main recommendation for healthcare practitioners is to use these machine learning models as extra decision-support tools. They should not take the place of medical skill, but they can provide important insights that would be difficult or counterintuitive to ascertain otherwise. The model is an effective tool for early diagnosis and treatment planning due to its high accuracy and interpretability features.

It is imperative for governments to create standards for the moral and appropriate application of AI and machine learning in healthcare. This study offers as an example of how these systems might incorporate interpretability and trust. When creating rules governing the use of AI in healthcare, policymakers should take these findings into account and make sure that all machine learning models used in clinical settings must be interpretable.

Researchers in medicine and the pharmaceutical industry can both gain from this study. The characteristics that have been deemed significant can act as focal points for additional study into the pathogenesis of meningioma. Additionally, the model's capacity to precisely forecast patient outcomes might be extremely helpful in clinical trials.

These results can serve as a framework for IT businesses looking to enter the healthcare AI market to create patient-centric, understandable, and reliable AI solutions. This is consistent with the market's increasing desire for accountable and transparent AI systems (Mittelstadt et al., 2019).

5.4 Limitations and Challenges

Despite the significant contributions made, it's vital to recognise the restrictions and difficulties this study encountered. The amount and diversity of the dataset used for model training and validation are two major limitations. Results from larger, more varied datasets may be more reliable and generalizable. According to Obermeyer and Emanuel (2016), data scarcity is a well-known problem in medical research.

The emphasis on a particular type of cancer, meningioma, is another drawback. Although the research offers helpful understandings into its diagnosis and therapy, the results could not directly apply to other tumours or ailments. The application of the study is thus somewhat constrained by this.

It is difficult to interpret machine learning models, and while LIME and SHAP are useful, they might not fully reveal the complexity of the models' thought processes. In the academic world, there is ongoing discussion over the difficulty of real interpretability in machine learning (Doshi-Velez & Kim, 2017).

5.5 Recommendations for Future Research

Given these limitations, future research should focus on several key areas. Firstly, extending the research to include multiple types of tumours will make the results more generalizable. Secondly, efforts should be made to incorporate more diverse and larger datasets, possibly through multi-centre collaborations.

Future studies could also employ more advanced machine learning techniques, such as deep learning models, for even better performance, although this would bring additional challenges in terms of interpretability. Researchers should also delve deeper into the aspect of model-agnostic interpretability, perhaps by comparing several different techniques to determine which offers the best compromise between accuracy and interpretability.

5.6 Final Summary and Concluding Remarks

In conclusion, this study marks a significant step forward in the application of machine learning in the realm of medical imaging and radiomics. While it achieves high performance in terms of accuracy and other quantitative metrics, it also introduces and validates several qualitative measures. These measures, focusing on interpretability and real-world applicability, are crucial for the broader acceptance and trust of these technologies in clinical settings.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., ... & Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications, 5(1), 1-9.

Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., & von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms and markers. Acta Neuropathologica, 129(6), 829-848.

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct.

Artzi, M., Bressler, I., Ben Bashat, D. (2016). Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis. Journal of Magnetic Resonance Imaging, 48(3), 601-610.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... & Hermsen, M. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210.

Bernasconi, N., Duchesne, S., Janke, A., Lerch, J., Collins, D. L., & Bernasconi, A. (2003). Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy. *NeuroImage*, 23(2), 717-723.

Bilello, E., & Kirby, J. (2023). Segmentation and Classification of Grade I and II Meningiomas from Magnetic Resonance Imaging: An Open Annotated Dataset (Meningioma-SEG-CLASS).

Bilello, E., & Kirby, J. (2023). Segmentation and Classification of Grade I and II Meningiomas from Magnetic Resonance Imaging: An Open Annotated Dataset (Meningioma-SEG-CLASS).

Bishop, C. M. (2006). Pattern recognition and machine learning.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., ... & Sturm, D. (2013). The somatic genomic landscape of glioblastoma. Cell, 155(2), 462-477.

Brinkmann, S. (2013). Qualitative interviewing. Oxford University Press.

British Educational Research Association. (2018). Ethical guidelines for educational research.

British Psychological Society. (2018). Code of Ethics and Conduct.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517-518.

Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., & Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. *Applied ergonomics*.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730).

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8), 832.

Castelvecchi, D. (2016). Can we open the black box of AI?

Chang, P., Grinband, J., Weinberg, B. D., Bardis, M., Khy, M., Cadena, G., ... & Chow, D. (2018). Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. American Journal of Neuroradiology, 39(7), 1201-1207.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983.

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8837-8846).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

C. Molnar, Interpretable Machine Learning, 2019. https://christophm. github.io/interpretable-ml-book

Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Ellingson, B. M., Bendszus, M., Boxerman, J., Barboriak, D., Erickson, B. J., Smits, M., ... & Wen, P. Y. (2015). Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. Neuro-oncology, 17(9), 1188-1198.

Epstein, R. M., & Street Jr, R. L. (2007). Patient-Centered Communication in Cancer Care: Promoting Healing and Reducing Suffering. National Cancer Institute, NIH Publication No. 07-6225.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

García, S., Luengo, J. and Herrera, F., 2015. *Data preprocessing in data mining* (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: images are more than pictures, they are data. Radiology, 278(2), 563-577.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a" right to explanation". AI magazine, 38(3), 50-57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-93.

Haacke, E. M., Brown, R. W., Thompson, M. R., & Venkatesan, R. (1999). *Magnetic resonance imaging: Physical principles and sequence design*. John Wiley & Sons.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). Multivariate data analysis.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Inselberg, A. (1985). The plane with parallel coordinates. The Visual Computer.

Jalbert, L.E., Neill, E., Phillips, J.J., Lupo, J.M., Olson, M.P., Molinaro, A.M., Berger, M.S., Chang, S.M. and Nelson, S.J., 2016. Magnetic resonance analysis of malignant transformation in recurrent glioma. *Neuro-oncology*, *18*(8), pp.1169-1179.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A.

Kickingereder, P., Götz, M., Muschelli, J., Wick, A., Neuberger, U., Shinohara, R. T., ... & van den Bent, M. (2016). Large-scale Radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. Clinical Cancer Research, 22(23), 5765-5771.

Kniep, H. C., Madesta, F., Schneider, T., Hanning, U., Schönfeld, M. H., Schön, G., ... & Fiehler, J. (2019). Radiomics of brain MRI: utility in prediction of metastatic tumor type. Radiology, 290(2), 479-487.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International joint Conference on artificial intelligence.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., ... & Gillies, R. J. (2012). Radiomics: the process and the challenges. Magnetic resonance imaging, 30(9), 1234-1248.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., de Jong, E. E., van Timmeren, J., ... & Walsh, S. (2017). Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology, 14(12), 749-762.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., ... & Aerts, H. J. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer, 48(4), 441-446.

Lao, J., Chen, Y., Li, Z. C., Li, Q., Zhang, J., Liu, J., & Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Scientific Reports, 7(1), 1-8.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J. and Thiesson, B. (2020b). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, [online] 11(1), p.3852. doi:https://doi.org/10.1038/s41467-020-17431-x.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Liu, S., Zhang, Y., Chen, L., Guan, M., & Zhou, X. (2020). Radiomics analysis using contrast-enhanced CT for preoperative prediction of occult peritoneal metastasis in advanced gastric cancer. European Radiology, 30(4), 2390-2399.

Liu, Z., Wang, S., Dong, D., Wei, J., Fang, C., Zhou, X., & Tian, J. (2017). The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. Theranostics, 9(5), 1303–1322.

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., & Wiestler, O. D. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica, 131(6), 803-820

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).

Lundberg, S. M., Erion, G., & Lee, S.-I. (2017). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., ... & Reyes, M. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging, 34(10), 1993-2024.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Molnar, C. (2019). Interpretable machine learning

Naeini, K. M., Pope, W. B., Cloughesy, T. F., Harris, R. J., Lai, A., Eskin, A., ... & Nghiemphu, P. L. (2013). Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. Neuro-oncology, 15(5), 626-634.

Omuro, A., & DeAngelis, L. M. (2013). Glioblastoma and other malignant gliomas: a clinical review. Jama, 310(17), 1842-1850.

Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C. and Barnholtz-Sloan, J.S., 2020. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro-oncology*, *22*(Supplement_1), pp.iv1-iv96.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144.

Pieper, S., Lorensen, B., Schroeder, W., & Kikinis, R. (2006). The NA-MIC Kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community. *Proceedings of the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135-1144).

Rozemberczki B, Watson L, Bayer P, Yang H T, Kiss O, Nilsson S and Sarkar R 2022 The shapley value in machine learning (arXiv:2202.05594)

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*.

Staedtke, V., Bai, R. Y., & Laterra, J. (2016). Investigational new drugs for brain cancer. Expert opinion on investigational drugs, 25(9), 937-956.

Tixier, F., Le Rest, C. C., Hatt, M., Albarghach, N., Pradier, O., Metges, J. P., ... & Visvikis, D. (2011). Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. Journal of Nuclear Medicine, 52(3), 369-378.

Tufte, E. R. (1990). Envisioning information.

Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., ... & Aerts, H. J. W. L. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21), e104-e107.

Van Meir, E. G., Hadjipanayis, C. G., Norden, A. D., Shu, H. K., Wen, P. Y., & Olson, J. J. (2010). Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. CA: a cancer journal for clinicians, 60(3), 166-193.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology.

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., ... & Chamberlain, M. (2010). Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. Journal of clinical oncology, 28(11), 1963.

Wiens, J., Shenoy, E. S., & Platt, R. (2019). Machine learning in healthcare: a critical appraisal of challenges and opportunities. eGEMs (Generating Evidence & Methods to improve patient outcomes), 7(1), 1.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). Experimentation in software engineering.

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*

Yin, R. K. (2017). Case study research and applications: Design and methods. SAGE Publications.

Young, R. J., Knopp, E. A., & Cha, S. (2011). Brain MRI: techniques in the diagnosis of neurologic diseases. The Medical clinics of North America, 95(5), 893-902.

Zacharaki, E. I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E. R., & Davatzikos, C. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 62(6), 1609-1618.

Zhang, B., Chang, K., Ramkissoon, S., Tanguturi, S., Bi, W. L., Reardon, D. A., ... & Ligon, K. L. (2017). Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. Neuro-oncology, 19(1), 109-117.

Zhang, J., Yu, C., Jiang, G., Liu, W., & Tong, L. (2012). 3D texture analysis on MRI images of Alzheimer's disease. Brain imaging and behavior, 6(1), 61-69.

Zou, K.H., O'Malley, A.J. and Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, [online] 115(5), pp.654–657. doi:https://doi.org/10.1161/circulationaha.105.594929.

Zwanenburg, A., Leger, S., Vallières, M., & Löck, S. (2016). Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003.

Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H.J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R. and Bogowicz, M., 2020. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, *295*(2), pp.328-338.

APPENDIX

CODE

```
import numpy as np
import pandas as pd
from tqdm import tqdm

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import KFold
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve,auc
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score
from sklearn.linear_model import LassoCV
from sklearn.linear_model import Lasso
from shap import TreeExplainer, Explanation
Import time
Import shap
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

# Ignore all warnings
warnings.filterwarnings("ignore")


T1_Data=pd.read_csv("C://Users//Dell//Desktop//dissteration//T1_Meningioma.csv")
T1_Data.shape
T1_Data.isna().sum().sum()
T1_Data.head()
T1_Data.describe().T
T1_Data.corr()

#Visualizing the distribution of the target variable
plt.figure(figsize=(10, 6))
sns.countplot(data=T1_Data, x='Pathologic grade', order = T1_Data['Pathologic grade'].value_counts().index)
plt.title('Distribution of Pathologic Grade')
plt.xlabel('Pathologic Grade')
plt.ylabel('Count')
plt.show()
# Distribution of Age by Pathologic Grade
plt.figure(figsize=(10, 6))
sns.boxplot(data=T1_Data, x='Pathologic grade', y='Age')
plt.title('Age Distribution by Pathologic Grade')
plt.xlabel('Pathologic Grade')
plt.ylabel('Age')
plt.show()
```

```
# Identify categorical columns
categorical_columns = T1_Data.select_dtypes(include=['object']).columns

# Apply label encoding to categorical columns
label_encoders = {}
for column in categorical_columns:
    lbe = LabelEncoder()
    T1_Data[column] = lbe.fit_transform(T1_Data[column])
    #label_encoders[column] = le

# Show first few rows after encoding
T1_Data.head()
# Separate the features (X) and target variable (y)
X = T1_Data.drop('Pathologic grade', axis=1)
y = T1_Data['Pathologic grade']
# Split the data into training and testing sets (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
#Feature selection
scaler = StandardScaler()
stand_feature = scaler.fit_transform(X)
cols=X.columns
dat=X.copy()
for i in  range(1,len(cols)):

    column=cols[i]
    dat[column]=stand_feature[:,i-1]

lasso = Lasso(alpha=0.01)
lasso.fit(X,y)
selected_features = stand_feature[:, lasso.coef_ >= 0.0000001]
```

```
# Print the selected feature names
selected_feature_names = np.array(list(X.keys()))[lasso.coef_ >= 0.0000001]
print("Selected Features:",selected_feature_names)
print("Number of features selected by lasso ",len(selected_feature_names))
selected_features.shape
feature = pd.DataFrame(selected_features, columns=selected_feature_names)
#spliting the selected feature by lasso
xtrain_lasso,xtest_lasso,ytrain_lasso,ytest_lasso=train_test_split(feature,y)
# Training model with RandomForestClassifier
estimator = RandomForestClassifier()
param_grid = {
    'n_estimators': [10, 50, 100],
    'max_depth': [None, 10, 20],
    }

# Create a GridSearchCV object
RFFS_grid_search = GridSearchCV(estimator, param_grid)
 # Perform the search using cross_val_score and GridSearchCV
RFFS_grid_search.fit(xtrain_lasso,ytrain_lasso)
scores = cross_val_score(RFFS_grid_search, xtrain_lasso, ytrain_lasso)
 # Get the best estimator and its corresponding hyperparameters
best_estimator = RFFS_grid_search.best_estimator_
best_params = RFFS_grid_search.best_params_
print("Cross-validation scores:", scores)
print("Best estimator:", best_estimator)
print("Best parameters:", best_params)
RFFS_ypred=RFFS_grid_search.predict(xtest_lasso)
RFFS_accuracy=accuracy_score(ytest_lasso,RFFS_ypred)
print (RFFS_accuracy)
print(classification_report(ytest_lasso, RFFS_ypred))
print("The accuracy for Random Forest Model after performing feature selection is
",RFFS_accuracy)


# Plotting the heatmap for the confusion matrix
from sklearn.metrics import confusion_matrix
mat = confusion_matrix(ytest_lasso, RFFS_ypred)
sns.heatmap(mat, square=True, annot=True, fmt='d', cbar=False)
plt.xlabel('true label')
plt.ylabel('predicted label')
RFFSypredpro=RFFS_grid_search.predict_proba(xtest_lasso)
RFFSypredpro=RFFSypredpro[:,1]
rfs_roc=roc_auc_score(ytest_lasso,RFFSypredpro,multi_class='ovr')
rffs,lrffs,_=roc_curve(ytest_lasso,RFFSypredpro,pos_label=1)
print("AUC SCORE FOR RANDOM FOREST CLASSIFIER ",rfs_roc)
# Training model with Logistic Regression
Logistic_estimator = LogisticRegression(solver='liblinear')
param_grid = {
    #'max_depth': [None, 10, 20]
    # Add more hyperparameters to search over
    }



# Create a GridSearchCV object
Logistic_grid_search = GridSearchCV(Logistic_estimator, param_grid, cv=2)


# Perform the search using cross_val_score and GridSearchCV
Logistic_grid_search.fit(xtrain_lasso,ytrain_lasso)
Logistic_scores = cross_val_score(Logistic_grid_search, xtrain_lasso, ytrain_lasso, cv=2)


# Get the best estimator and its corresponding hyperparameters
```

```
best_estimator = Logistic_grid_search.best_estimator_
best_params = Logistic_grid_search.best_params_

print("Cross-validation scores:", Logistic_scores)
print("Best estimator:", best_estimator)
print("Best parameters:", best_params)



Logistic_ypred=Logistic_grid_search.predict(xtest_lasso)
Logisitic_accuracy=accuracy_score(ytest_lasso,Logistic_ypred)
print(Logisitic_accuracy)
print(classification_report(ytest_lasso,Logistic_ypred))
print("The  accuracy for Logistic Regression Model after performing feature selection
",Logisitic_accuracy)
# Plotting the heatmap for the confusion matrix
mat = confusion_matrix(ytest_lasso, Logistic_ypred)
sns.heatmap(mat, square=True, annot=True, fmt='d', cbar=False)
plt.xlabel('true label')
plt.ylabel('predicted label')
Logisticpredpro=Logistic_grid_search.predict_proba(xtest_lasso)
Logisticypredpro=Logisticpredpro[:,1]
log_roc=roc_auc_score(ytest_lasso,Logisticypredpro,multi_class='ovr')
log,logr,_=roc_curve(ytest_lasso,Logisticypredpro,pos_label=1)
print("AUC SCORE FOR Logistic Regression ",log_roc)

 # Training model with Decision Tree Classifier
decisiontree_estimator = DecisionTreeClassifier()
param_grid = {
    'max_depth': [None, 10, 20]
    # Add more hyperparameters to search over
```

```
        }

# Create a GridSearchCV object
dectree_grid_search = GridSearchCV(decisiontree_estimator, param_grid, cv=2)

# Perform the search using cross_val_score and GridSearchCV
dectree_grid_search.fit(xtrain_lasso,ytrain_lasso)
scores_dTree = cross_val_score(dectree_grid_search , xtrain_lasso, ytrain_lasso, cv=2)

# Get the best estimator and its corresponding hyperparameters
best_estimator = dectree_grid_search.best_estimator_
best_params = dectree_grid_search.best_params_

print("Cross-validation scores:", scores_dTree)
print("Best estimator:", best_estimator)
print("Best parameters:", best_params)

lassoDTree_ypred=dectree_grid_search.predict(xtest_lasso)
lassoDTree_accuracy=accuracy_score(ytest_lasso,lassoDTree_ypred)
print(lassoDTree_accuracy)
print(classification_report(ytest_lasso, lassoDTree_ypred))
print("The  accuracy for Decision Tree Model after performing feature selection is 
",lassoDTree_accuracy)
Decisionpredpro=dectree_grid_search.predict_proba(xtest_lasso)
Decisionpredpro=Decisionpredpro[:,1]
dec_roc=roc_auc_score(ytest_lasso,Decisionpredpro,multi_class='ovr')
dec,decr,_=roc_curve(ytest_lasso,Decisionpredpro,pos_label=1)
print("AUC SCORE FOR Logistic Regression ",dec_roc)

#Plotting ROC Curve
```

```
plt.title("Receiver Operating Characteristic")
plt.plot(rffs,lrffs,linestyle="--",label="RandomForest(auc= %0.3f)" % rfs_roc)
plt.plot(log,logr,linestyle="--",label="LogisticRegression(auc= %0.3f)" % log_roc)
plt.plot(dec,decr,linestyle="--",label="DecisionTree(auc= %0.3f)" % dec_roc)
plt.xlabel("False Positive Rate --->")
plt.ylabel("True Positive Rate --->")
plt.legend()
plt.show()
#Interpretability
#Feature importance on RandomForest Classifier
feature_names = feature.columns
 # Train a RandomForestClassifier
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(feature, y)
 # Get feature importance from the trained RandomForestClassifier
feature_importance_values = clf.feature_importances_
 top_x = 10  # number of x most important features to show
 # Create a figure and axis
fig, ax = plt.subplots(figsize=(12, 6))
 # Sort and get the indices of the top features
indices = np.argsort(feature_importance_values)[-top_x:]
 # Create horizontal bar plot
bars = ax.barh(
    range(len(indices)), feature_importance_values[indices], color="b", align="center"
)
 ax.set_title("RandomForestClassifier", fontweight="normal", fontsize=16)
 plt.sca(ax)
plt.yticks(
    range(len(indices)),
    [feature_names[j] for j in indices],
    fontweight="normal",
    fontsize=12,
)

color_list = sns.color_palette("dark", len(feature_names))
for i, ticklabel in enumerate(ax.get_yticklabels()):
    ticklabel.set_color(color_list[indices[i]])

for i, bar in enumerate(bars):
    bar.set_color(color_list[indices[i]])
plt.box(False)

plt.title(
    "Feature Importance for RandomForestClassifier. Top " + str(top_x) + " Features.",
    fontsize=14,
    fontweight="normal",
)

plt.tight_layout()
plt.show()
plt.figure(figsize=(15,6))

 #Feature importances on logistic regression
logistic=LogisticRegression()
logistic.fit(xtrain_lasso,ytrain_lasso)
logistic_reg_coeff = logistic.coef_[0]
color_list =  sns.color_palette("dark", len(feature.columns))
top_x = 10
idx = np.argsort(np.abs(logistic_reg_coeff))[::-1]
lreg_ax = plt.barh(feature.columns[idx[:top_x]][::-1], logistic_reg_coeff[idx[:top_x]][::-1])
```

```
for i,bar in enumerate(lreg_ax):
    bar.set_color(color_list[idx[:top_x][::-1][i]])
    plt.box(False)

lr_title = plt.suptitle("Logistic Regression. Top " + str(top_x) + " Coefficients.", fontsize=20, fontweight="normal")

#Feature importance on Random Forest Classifier without feature selection

feature_names_nofs = X.columns
 # Train a RandomForestClassifier
clf_nofs = RandomForestClassifier(n_estimators=100, random_state=42)
clf_nofs.fit(X, y)

# Get feature importances from the trained RandomForestClassifier
nofs_feature_importance_values = clf_nofs.feature_importances_

top_x = 10  # number of x most important features to show

# Create a figure and axis
fig, ax = plt.subplots(figsize=(12, 6))

# Sort and get the indices of the top features
indices = np.argsort(nofs_feature_importance_values)[-top_x:]

# Create horizontal bar plot
bars = ax.barh(
    range(len(indices)), nofs_feature_importance_values[indices], color="b", align="center"
)
```

```
ax.set_title("RandomForestClassifier", fontweight="normal", fontsize=16)

plt.sca(ax)
plt.yticks(
    range(len(indices)),
    [feature_names_nofs[j] for j in indices],
    fontweight="normal",
    fontsize=12,
)

color_list = sns.color_palette("dark", len(feature_names_nofs))
for i, ticklabel in enumerate(ax.get_yticklabels()):
    ticklabel.set_color(color_list[indices[i]])

for i, bar in enumerate(bars):
    bar.set_color(color_list[indices[i]])
plt.box(False)

plt.title(
    "Feature Importance for RandomForestClassifier without feature selection. Top " + str(top_x)
+ " Features.",
    fontsize=14,
    fontweight="normal",
)

plt.tight_layout()
plt.show()

#LIME on RandomForest Classifier
# Create a DataFrame for scaled feature importance
```

```
feature_importance_scaled_df = pd.DataFrame({'Feature': feature.columns, 'Importance': feature_importance_values})


# Sort the DataFrame by the importances
sorted_feature_importance_scaled_df = feature_importance_scaled_df.sort_values(by='Importance', ascending=False)


# Select features where importance is greater than a threshold (e.g., 0.01)
selected_features = sorted_feature_importance_scaled_df[sorted_feature_importance_scaled_df['Importance'] > 0]


# Display the sorted feature importance
sorted_feature_importance_scaled_df.head(10)
# Import the LimeTabularExplainer class from the Lime library
from lime.lime_tabular import LimeTabularExplainer


# Create an explainer object using LimeTabularExplainer
# Provide training data, feature names, class names, and mode
Lime_explainer = LimeTabularExplainer(xtrain_lasso.values,
                                       feature_names=xtrain_lasso.columns.tolist(),
                                       class_names=['0', '1'],
                                       mode='classification')


# Iterate through each instance in the test data
for i in range(len(xtest_lasso)):
    # Explain the instance using the explainer
    # Provide instance data, prediction function, and the number of features
    exp = Lime_explainer.explain_instance(xtest_lasso.values[i],
                                           clf.predict_proba,
                                           num_features=xtest_lasso.shape[1])
    # Display the explanation in a notebook, hiding some details
    exp.show_in_notebook(show_all=False)

# Loop through all instances in X_test_selected
for i in range(len(xtest_lasso)):

    # Get the instance
    print("Target Class is ",y_test.iloc[i])
    instance = xtest_lasso.iloc[i].values.reshape(1, -1)

    # Generate LIME explanation
    exp = Lime_explainer.explain_instance(instance.reshape(-1,),
                                     clf.predict_proba,
                                     num_features=xtest_lasso.shape[1])

    # Create a DataFrame for the explanation
    exp_df = pd.DataFrame(exp.as_list(), columns=['Feature', 'Weight'])

    # Add a column for the absolute value of the weights
    exp_df['Weight'] = exp_df['Weight']

 # Sort the DataFrame by the absolute value of the weights
    exp_df = exp_df.sort_values('Weight', ascending=False).reset_index(drop=True)

    print(f'Explanation for Instance {i + 1}')
    display(exp_df)
    print('----------------------\n')

#Lime in Logistic Regression
# Import the LimeTabularExplainer class from the Lime library
```

```
from lime.lime_tabular import LimeTabularExplainer

# Create an explainer object using LimeTabularExplainer
# Provide training data, feature names, class names, and mode
Lime_explainer = LimeTabularExplainer(xtrain_lasso.values,
                                      feature_names=xtrain_lasso.columns.tolist(),
                                      class_names=['0', '1'],
                                      mode='classification')

# Iterate through each instance in the test data
for i in range(len(xtest_lasso)):
    # Explain the instance using the explainer
    # Provide instance data, prediction function, and the number of features
    exp = Lime_explainer.explain_instance(xtest_lasso.values[i],
                                          Logistic_grid_search.predict_proba,
                                          num_features=xtest_lasso.shape[1])

    # Display the explanation in a notebook, hiding some details
    exp.show_in_notebook(show_all=False)

for i in range(len(xtest_lasso)):


    # Get the instance
    print("Target Class is ",y_test.iloc[i])
    instance = xtest_lasso.iloc[i].values.reshape(1, -1)

    # Generate LIME explanation
    exp = Lime_explainer.explain_instance(instance.reshape(-1,),
                                          Logistic_grid_search.predict_proba,
                                          num_features=xtest_lasso.shape[1])
    # Create a DataFrame for the explanation
    exp_df = pd.DataFrame(exp.as_list(), columns=['Feature', 'Weight'])

    # Add a column for the absolute value of the weights
    exp_df['Weight'] = exp_df['Weight']

    # Sort the DataFrame by the absolute value of the weights
    exp_df = exp_df.sort_values('Weight', ascending=False).reset_index(drop=True)

    print(f'Explanation for Instance {i + 1}')
    display(exp_df)
    print('-----------------------\n')


#SHaP
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(xtrain_lasso, ytrain_lasso)
 #plotting shap summary plot for randomforest classifier
# Initialize the SHAP explainer
shap_exp = shap.Explainer(clf)

# Calculate SHAP values for a subset of the test data (e.g., the first 10 instances)
#sample_size = 10
shap_values = shap_exp.shap_values(xtest_lasso)

# Plot the summary plot to visualize feature importance
shap.summary_plot(shap_values, xtest_lasso, feature_names=selected_feature_names, plot_type="bar")
plt.show()
```

```
#plotting shap summary plot on logistic regression

Logistic_grid_search.fit(xtrain_lasso, ytrain_lasso)

# Initialize the SHAP explainer
shap_exp = shap.Explainer(clf)

# Calculate SHAP values for a subset of the test data (e.g., the first 10 instances)
#sample_size = 10
shap_values = shap_exp.shap_values(xtest_lasso)

# Plot the summary plot to visualize feature importance
shap.summary_plot(shap_values, xtest_lasso, feature_names=selected_feature_names, plot_type="bar")

plt.show()
#An instance of one predicition
sample = pd.DataFrame(xtest_lasso.iloc[0]).T
#plotting shap Forceplot
RFModelFS=RandomForestClassifier()
RFModelFS.fit(xtrain_lasso,ytrain_lasso)
#subsampled_test_data = xtest_lasso.iloc[10].reshape(1, -1)

start_time = time.time()
explainer = shap.TreeExplainer(RFModelFS)
#explainer = shap.TreeExplainer(RFModelFS)
shap_values = explainer.shap_values(sample)
elapsed_time = time.time() - start_time
```

```
# explain first sample from test data
print(
    "Kernel Explainer SHAP run time",
    round(elapsed_time, 3),
    " seconds. "
)
values=explainer.expected_value
ind=np.argmax(values)
print("SHAP expected value", explainer.expected_value)
print("Model mean value", RFModelFS.predict_proba(xtrain_lasso).mean(axis=0))
print("Model prediction for test data", RFModelFS.predict_proba(sample))
class_names=list(y.unique())

shap.initjs()
pred_ind = 0
print(class_names[ind])
shap.force_plot(
    explainer.expected_value[ind],
    shap_values[1][0],
    sample,
    feature_names=xtrain_lasso.columns,
)
#Another instance of a predicition
RFModelFS=RandomForestClassifier()
RFModelFS.fit(xtrain_lasso,ytrain_lasso)
#subsampled_test_data = xtest_lasso.iloc[10].reshape(1, -1)

start_time = time.time()
explainer = shap.TreeExplainer(RFModelFS)
#explainer = shap.TreeExplainer(RFModelFS)
```

```
shap_values = explainer.shap_values(sample)
elapsed_time = time.time() - start_time

# explain first sample from test data
print(
    "Kernel Explainer SHAP run time",
    round(elapsed_time, 3),
    " seconds. "
)
values=explainer.expected_value
ind=np.argmax(values)
print("SHAP expected value", explainer.expected_value)
print("Model mean value", RFModelFS.predict_proba(xtrain_lasso).mean(axis=0))
print("Model prediction for test data", RFModelFS.predict_proba(sample))
class_names=list(y.unique())

shap.initjs()
pred_ind = 0
print(class_names[ind])
shap.force_plot(
    explainer.expected_value[ind],
    shap_values[0][0],
    sample,
    feature_names=xtrain_lasso.columns,
)
#  Waterfall plots displaying an explanations for one individual predictions
model=RandomForestClassifier()
model.fit(feature,y)
exp_waterfall = shap.Explainer(model, feature)
s_values = exp_waterfall(feature,check_additivity=False )
```

```
shap.plots.waterfall(s_values[0,:,0], max_display=36)

#shap decision plot
decision_explainer = shap.TreeExplainer(model)
dec_expected_value = decision_explainer.expected_value
if isinstance(dec_expected_value, list):
    dec_expected_value = dec_expected_value[1]
print(f"Explainer expected value: {dec_expected_value}")

select = range(20)
features = xtest_lasso.iloc[select]
#features_display = X_display.loc[features.index]

with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    dec_shap_values = decision_explainer.shap_values(xtest_lasso)[1]
    shap_interaction_values = decision_explainer.shap_interaction_values(features)
if isinstance(shap_interaction_values, list):
    shap_interaction_values = shap_interaction_values[1]

shap.decision_plot(dec_expected_value[1], dec_shap_values, xtrain_lasso.columns)
```