Predicting Patient Outcomes in the EPOCH Clinical Trial

by

Nithin Parsan

S.B. Computer Science and Molecular Biology, Massachusetts Institute of Technology, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Nithin Parsan Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Kenney Ng, Ph.D. Principal Investigator MIT-IBM Watson AI Lab, Thesis Supervisor
Certified by:	Aude Oliva, Ph.D. Principal Investigator, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair Master of Engineering Thesis Committee

Predicting Patient Outcomes in the EPOCH Clinical Trial

by

Nithin Parsan

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

ABSTRACT

Metastatic colorectal cancer (mCRC) has a poor prognosis and high mortality rate, but innovative therapies such as transarterial radioembolization (TARE) can improve patient outcomes. The EPOCH clinical trial demonstrated that TARE improved hepatic progression-free survival (hPFS) in patients with colorectal liver metastases, and computational methods to analyze the multimodal data collected can identify patient subgroups and predict treatment response for personalized medicine. First, a comprehensive data preprocessing pipeline curated a high-quality dataset of liver-region Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans paired with patient biomarkers. Multi-Dimensional Subset Scanning (MDSS) identified a group of patients with shared biomarkers that exhibited poor response to TARE, and Cox Proportional Hazards (CoxPH) modeling revealed hazard ratios for biomarkers aligning with clinical expectations, albeit with a limited C-index. Augmenting CoxPH modeling with embeddings from a deep learning foundation model pre-trained on liver CT and MRI scans and fine-tuned to predict treatment response resulted in a substantially higher C-index. Interestingly, models fine-tuned to predict one clinical feature had improved predictive accuracy for other features they were not specifically trained on, and Class Activation Mapping (CAM) visualizations showed that salient embedding dimensions focus on the liver region, providing interpretability. The ensemble of computational techniques applied to multimodal clinical trial data successfully identified patient subgroups, extracted predictive biomarkers, and enhanced the accuracy of treatment response predictions, contributing to the development of more effective, personalized treatment strategies for mCRC patients undergoing TARE.

Thesis supervisor: Kenney Ng, Ph.D.
Title: Principal Investigator MIT-IBM Watson AI Lab

Thesis supervisor: Aude Oliva, Ph.D.
Title: Principal Investigator

Acknowledgments

With heartfelt sincerity, I wish to convey my profound gratitude to my advisor, Kenney Ng for his guidance throughout the past two semesters of my MEng program. I am immensely appreciative of the MIT-IBM Watson AI Lab, the team at Boston Scientific, and Professor Aude Oliva for granting me the privilege to contribute to such a meaningful project. I am truly grateful for the opportunity to conduct research that directly impacts clinical practice and resonates deeply with me on a personal level. It has been an honor to call MIT my home for the past four years, and my work this year has provided a gratifying conclusion to an extraordinary journey. Before proceeding, I would like to briefly acknowledge those who supported me most throughout my time at MIT.

Reflecting back on my years at MIT, my greatest memories have been with the people around me, who have motivated me to grow and change in unbelievable ways. I am proud I was able to call PKT my home and I will forever cherish the lifelong friendships community that I found there. The highs, lows, and countless stories will last not four years, but a lifetime. Every person in PKT has impacted my life, but I'd like to give special thanks to my longest roommate Matt McManus for the hilarious moments and mentorship and my class for the numerous adventures and joyful memories we shared.

Finally, I would like to thank my parents, grandmother, and my sister, Anisha. Their lifetime of encouragement and support is the reason I am here today. I'm motivated daily by their unconditional belief in me and for that I am eternally grateful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The global burden of cancer continues to escalate, representing one of the most significant health challenges of the 21st century. Colorectal cancer (CRC), in particular, ranks as the third most diagnosed cancer worldwide, with metastatic CRC (mCRC) often presenting a formidable prognosis due to its advanced stage at diagnosis and limited treatment success rates. The urgent need for improved diagnostic and therapeutic strategies is clear, as early detection and tailored treatment approaches can substantially enhance patient outcomes and survival rates [1].

In this context, the integration of computer science, particularly artificial intelligence (AI), into medical research offers transformative potential for cancer diagnosis and treatment. AI methodologies, especially those harnessing deep learning, are increasingly recognized for their ability to unearth complex patterns in data that are imperceptible to human analysis. In the realm of oncology, AI-driven approaches are adept at identifying novel clinical biomarkers, which are critical for early disease detection, prognosis, and the customization of treatment strategies [2]. These biomarkers, particularly when derived from imaging data, can significantly inform clinical decisions and influence the development of personalized medicine, ultimately leading to better patient care outcomes [3].

Recent advancements in computer vision technology have significantly enhanced the utility of AI in medical imaging. The evolution of deep learning models, which employ sophisticated neural network architectures with increasingly deep layers, has revolutionized the analysis of medical images. Such models are particularly adept at learning complex features across billions of captured images and are similarly well adapted for medical imaging modalities like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Recent efforts to collate large-scale datasets such as the UK Biobank [4], which includes extensive imaging data, have become invaluable to train robust AI models. The development of pretrained foundation models in medical imaging, utilizing these extensive datasets, exemplifies this technological progress, providing a versatile base from which customized diagnostic tools can be efficiently developed to address specific clinical needs, such as the diagnosis of tuberculosis from chest X-ray images [5].

We seek to apply computer vision technologies to Boston Scientific's EPOCH clinical trial, which focuses on transarterial radioembolization (TARE) therapy for mCRC [6]. By deploying sophisticated, multi-modal tools to interpret and cluster imaging data from modalities such as CT and MRI scans, we aim to identify predictive biomarkers of treatment response. Our goal is to assist in the development of more effective, personalized treatment protocols to improve outcomes for patients with complex oncological conditions.

1.2 Problem Statement

Despite advancements in treatment modalities, patient outcomes in mCRC remain heterogeneous, underscoring the need for personalized approaches to optimize therapeutic efficacy. The EPOCH clinical trial, a multicenter, open-label phase III study sponsored by Boston Scientific, investigated the impact of transarterial Yttrium-90 radioembolization (TARE) in combination with second-line systemic chemotherapy for colorectal liver metastases (CLM) [6]. While the trial demonstrated improved progression-free survival (PFS) and hepatic PFS (hPFS) in patients receiving TARE [6], the heterogeneity in patient responses necessitates further investigation to improve treatment regimes.

The heterogeneous response to TARE in mCRC patients suggests that there may be underlying factors influencing treatment efficacy that are not yet fully understood. While these factors could include patient-specific characteristics, such as genetic profiles, tumor biology, and microenvironmental factors, as well as treatment-related variables, such as dosing, timing, and combination with other therapies [7], directly measuring these variables can be challenging and resource-intensive. However, medical imaging data, such as computed tomography (CT) and magnetic resonance imaging (MRI), provide a high-dimensional modality that is well-suited for feature extraction and may serve as a proxy for some of these hidden variables.

Medical images contain a wealth of information about tumor characteristics, including size, shape, texture, and vascularization, which can reflect underlying biological processes and treatment response. By applying advanced computational techniques, such as radiomics and deep learning, to these images, we can extract a wide range of quantitative features that may capture the complex interplay between patient-specific factors and treatment efficacy

[8]. These imaging-derived features could potentially serve as non-invasive biomarkers for predicting patient outcomes and guiding treatment decisions. The EPOCH trial collected a rich dataset, including patient demographics, clinical characteristics, and high-dimensional imaging data from CT and MRI modalities, which can be used jointly to uncover complex patterns and biomarkers predictive of treatment response [6]. However, the heterogeneous, multi-modal nature of the data poses significant challenges for traditional analytical approaches, necessitating significant preprocessing and feature extraction to detect differences in medical images that correlate with patient outcomes following treatment.

Identifying unique patient subgroups based on their imaging and clinical characteristics is essential for personalized treatment strategies in mCRC. However, this task is challenging due to the complexity and high dimensionality of medical imaging data, the need to integrate it with other clinical variables, and the requirement for advanced computational methods to process and analyze large datasets effectively. Extracting meaningful features from medical images that capture tumor biology and treatment response, and developing multimodal machine learning models that can combine and interpret disparate data types, are critical steps in identifying these patient subgroups. By stratifying patients into distinct groups based on their imaging and clinical characteristics, we can better understand the factors contributing to heterogeneous treatment responses and develop targeted interventions that optimize patient outcomes, ultimately improving the effectiveness of personalized treatment strategies for mCRC patients.

1.3 Research Question and Objectives

The primary research question that this study aims to address is: Can advanced computational methods, specifically deep learning and clustering techniques, be applied to the multi-modal data collected in the EPOCH clinical trial to identify unique patient subgroups and predictive biomarkers that correlate with treatment response to TARE in mCRC patients?

To answer this question, the study will pursue the following objectives:

1. Generate comprehensive statistics on the available categorical and numerical biomarkers and data points in the EPOCH trial dataset. Preprocess the image data by applying restrictions based on extracted metadata, image quality, and liver region identification to create a high-quality dataset suitable for further analysis.

2. Develop baseline models to group and predict patient outcomes using the preprocessed data. Employ clustering algorithms to identify unique patient subgroups and utilize

hazard modeling techniques, such as Cox proportional hazards (CoxPH) regression, to predict treatment response and survival outcomes.

3. Fine-tune pre-trained medical foundation models to augment the structured data with embedding features. Evaluate the performance of these models in classifying clinically relevant characteristics and identify limitations and issues with classification accuracy.

4. Extract image features and generate embeddings to improve classification performance and identify clinically relevant groupings. Investigate the impact of incorporating image embeddings and structured data on the performance of CoxPH modeling and classification tasks. Assess the effects of fine-tuning the models for different downstream tasks on classification tasks and evaluate model interpretability using techniques such as hierarchical clustering of correlated image embeddings and class activation mapping (CAM) for CoxPH modeling.

In this thesis, we aim to demonstrate the potential of deep learning and clustering techniques in identifying unique patient subgroups and predictive biomarkers that correlate with treatment response to TARE in mCRC patients.

1.4 Thesis Structure Overview

This thesis is organized into several chapters and the structure is as follows:

- **Chapter 1: Introduction** - This chapter provides an overview of the thesis, including the motivation behind the research, the problem statement regarding heterogeneous patient responses to TARE in mCRC, the research questions, the proposed approach utilizing deep learning and clustering techniques, and the significance of the research in the context of personalized medicine.

- **Chapter 2: Background on Metastatic Colorectal Cancer and the EPOCH Clinical Trial** - This chapter presents a comprehensive overview of mCRC, including the specific subtype used in the study, common treatments, and an introduction to TARE therapy, specifically focusing on the TheraSphere system. It also provides an overview of the EPOCH clinical trial, its relevant results, and highlights the heterogeneous patient response to treatment.

- **Chapter 3: Leveraging Biomarkers from the EPOCH trial** - This chapter discusses common relevant biomarkers for cancer found in the literature, the biomarkers available in the EPOCH trial, their biological relevance, and potential biomarkers that can be derived from CT and MRI images.

- **Chapter 4: Technical Background** - This chapter provides a technical background on supervised machine learning, multi-modal models, embedding spaces from deep convolutional models, dimensionality reduction and visualization techniques (e.g., t-SNE), clustering methods, Cox proportional hazards modeling, and classifiers used in the study (e.g., XGBoost, Random Forest, Multi-Layer Perceptron).

- **Chapter 5: Model Architectures and Algorithmic Tools** - This chapter describes the preprocessing pipeline and relevant tools (e.g., TotalSegmentator), clustering methodology, and sets of features examined. It also presents the pre-trained model architectures as well as the embedding extraction process, feature combination for multimodal groupings, and multi-dimensional subset scanning.

- **Chapter 6: Dataset Overview and Preprocessing** - This chapter provides data statistics for structured and unstructured data, biomarkers used in identification, and the preprocessing pipeline. A comprehensive table quantifying different categories such as sex, lesion class, MR, and CT is presented.

- **Chapter 7: Preliminary Patient Subgroup Identification and Outcome Prediction** - This chapter focuses on baseline clustering of patient responses using structured data and multi-dimensional subset scanning (MDSS), as well as Cox proportional hazards prediction modeling.

- **Chapter 8: Augmenting Structured Data with Embedding Features from Fine-Tuned Pre-trained Medical Foundation Models** - This chapter explores the fine-tuning of pre-trained medical foundation models to augment structured data with embedding features. It evaluates the performance of these models in classifying clinically relevant characteristics, such as sex, and discusses the limitations and issues with classification accuracy.

- **Chapter 9: Enhancing Classification Performance and Clinically Relevant Groupings with Image Feature Extraction and Embeddings** - This chapter presents the embedding generation pipeline and compares the results from clustering image embeddings with the baseline models. It investigates the impact of incorporating image embeddings and structured data on the performance of Cox proportional hazards modeling and classification tasks, and assesses the effects of fine-tuning the models for different downstream tasks. Model interpretability is evaluated using class activation mapping on salient embedding dimensions identified by Cox proportional hazards modeling.

- **Chapter 10: Conclusion** - This chapter summarizes the key findings of the research, discusses the implications for personalized medicine in mCRC treated with TARE, and provides recommendations for future research directions.

Chapter 2

Background on Metastatic Colorectal Cancer and the EPOCH Clinical Trial

2.1 Overview of Metastatic Colorectal Cancer (mCRC)

Colorectal cancer (CRC) is a significant global health burden, ranking as the third most commonly diagnosed cancer and the second leading cause of cancer-related deaths worldwide. In 2020, approximately 1.9 million new cases of CRC were diagnosed, and around 935,000 deaths were attributed to this disease. The progression from benign adenomatous polyps to malignant carcinoma involves a series of genetic and epigenetic alterations, commonly known as the adenoma-carcinoma sequence, with key mutations in genes such as APC, KRAS, and TP53 [9].

Metastatic colorectal cancer (mCRC) occurs when cancer cells spread from the primary tumor in the colon or rectum to distant organs, with the liver being the most common site of metastasis. Approximately 20-25% of patients have metastatic disease at the time of diagnosis, and nearly 50% of CRC patients will develop metastases during their illness [10]. The prognosis for mCRC is generally poor, with a five-year survival rate of about 14% [10]. However, advances in systemic therapies, targeted treatments, and surgical techniques have improved outcomes, with the median overall survival for mCRC patients now exceeding 30 months in recent clinical trials [11].

The treatment of mCRC typically involves a combination of systemic chemotherapy, targeted therapy, and locoregional treatments such as surgery and radioembolization. Common chemotherapy regimens include combinations of fluoropyrimidines (5-FU or capecitabine) with oxaliplatin (FOLFOX) or irinotecan (FOLFIRI). Targeted therapies, such as bevacizumab (anti-VEGF) and cetuximab or panitumumab (anti-EGFR), are used based on the molecular profile of the tumor, including KRAS, NRAS, and BRAF mutation status [11].

2.2 Metastatic Colorectal Cancer in the Liver

The liver is the most common site for metastasis in colorectal cancer (CRC) patients, with approximately 20-30% of individuals with diagnosed with CRC developing liver metastases [12]. This high incidence is primarily due to the liver's unique blood supply, receiving blood directly from the gastrointestinal tract via the portal vein, which facilitates the dissemination of cancer cells from the colon and rectum to the liver [13]. Liver metastases significantly impact patient prognosis and are a major determinant of morbidity and mortality in mCRC [12].

Management of colorectal liver metastases (CLM) involves a multidisciplinary approach, integrating systemic chemotherapy, surgical resection, and locoregional therapies [13]. Surgical resection of liver metastases, when feasible, offers the best chance for long-term survival and potential cure. However, only 20-30% of patients with CLM are candidates for surgery due to factors such as the number, size, and location of metastases, as well as the patient's overall health and liver function [14]. For patients who are not surgical candidates, systemic chemotherapy regimes, referenced earlier, remain a cornerstone of treatment [11].

2.3 Overview of TARE Therapy and TheraSphere

Transarterial radioembolization (TARE) is a locoregional therapy that has gained prominence in the management of colorectal liver metastases (CLM), especially for patients who are not candidates for surgical resection. TARE involves the targeted delivery of radioactive microspheres directly to liver tumors via the hepatic artery, enabling high-dose radiation to be administered to the tumor while sparing the surrounding healthy liver tissue. This precise approach helps maximize tumor control and minimize systemic side effects [11].

TheraSphere, a specific type of TARE, uses Yttrium-90 (Y-90) glass microspheres. These microspheres are approximately 20-30 micrometers in diameter and are embedded with the radioactive isotope Y-90, which emits high-energy beta radiation. Once administered, these microspheres become lodged in the microvasculature of liver tumors, delivering targeted radiation over a period of several days to weeks. The high-dose radiation induces DNA damage and subsequent cell death in the tumor cells, leading to tumor shrinkage and potential necrosis [15]. The procedure for TARE with TheraSphere involves a multidisciplinary team, including interventional radiologists, oncologists, and nuclear medicine specialists. The process typically begins with a planning angiogram to map the hepatic vasculature and assess the tumor's blood supply. This is followed by a dosimetry calculation to determine the appropriate dose of Y-90 microspheres needed for effective treatment. The microspheres are then delivered via a catheter inserted into the hepatic artery, directly targeting the liver tumors [15].

TheraSphere has shown promising results in terms of tumor response and disease control in patients with liver-dominant mCRC. Clinical studies have demonstrated improvements in progression-free survival (PFS) and hepatic progression-free survival (hPFS) with the use of TARE. The ability to deliver high-dose radiation precisely to the tumor while sparing healthy tissue makes TheraSphere a valuable option for patients with inoperable liver metastases [6]. Despite the advancements and benefits of TARE, managing CLM remains complex due to the heterogeneous nature of the disease and the variability in patient responses to therapy.

2.4 Overview of the EPOCH Clinical Trial

The EPOCH (Evaluating TheraSphere in Patients with metastatic colorectal carcinoma Of the liver who have progressed on first-line Chemotherapy) clinical trial is a randomized, open-label, international, multicenter, phase III study that investigated the impact of adding transarterial Yttrium-90 radioembolization (TARE) to standard second-line chemotherapy for patients with colorectal liver metastases (CLM). The trial was designed to address the heterogeneous patient responses to treatment observed in previous studies and to evaluate the potential of TARE in improving outcomes for patients with limited treatment options after progression on first-line therapy [6].

The EPOCH trial enrolled 428 patients from 95 centers in North America, Europe, and Asia, who were randomly assigned 1:1 to receive either second-line chemotherapy alone or in combination with TARE using TheraSphere glass microspheres as seen in Figure 2.1 [6]. The study population included patients with unresectable unilobar or bilobar CLM, who had progressed on first-line oxaliplatin- or irinotecan-based chemotherapy. Key eligibility criteria included age 18 years, measurable disease by RECIST 1.1, performance status 0 or 1, and adequate liver function. Patients with prior arterial or radiotherapy to the liver, clinically evident ascites, or confirmed extrahepatic metastases were excluded [6].

The two primary endpoints of the EPOCH trial were progression-free survival (PFS) and hepatic PFS (hPFS), assessed by blinded independent central review using RECIST 1.1 criteria. Secondary endpoints included overall survival (OS), objective response rate (ORR), and disease control rate (DCR). The study was designed to have 80% power to detect a hazard ratio (HR) of 0.71 for PFS and 0.65 for hPFS, favoring TARE plus chemotherapy over chemotherapy alone [16].

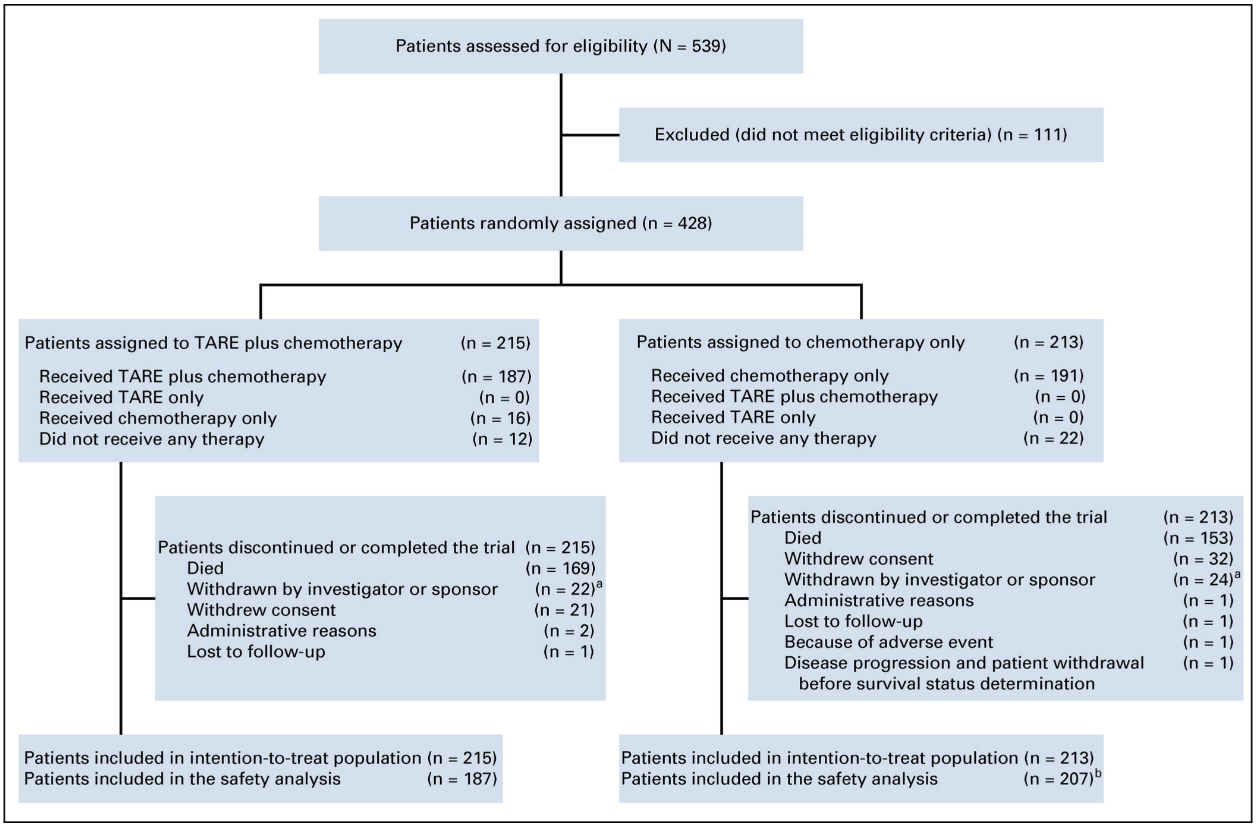
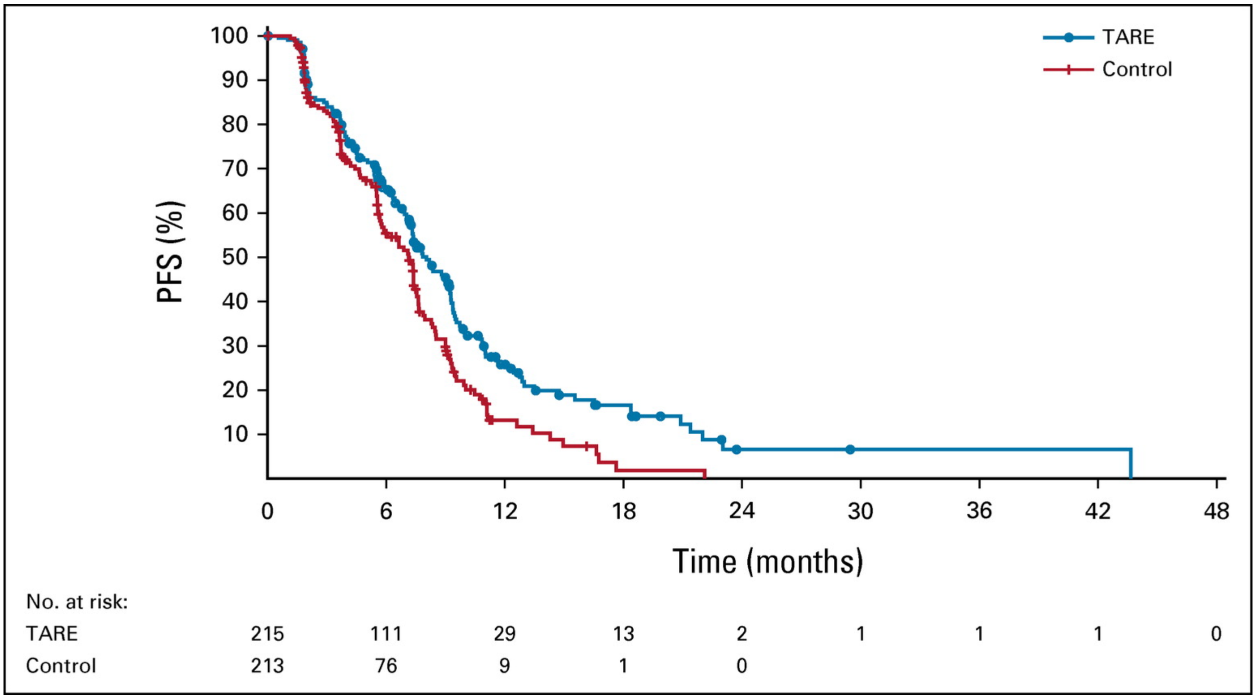

Patients assessed for eligibility (N = 539)
Excluded (did not meet eligibility criteria) (n = 111)
Patients randomly assigned (n = 428)
Patients assigned to TARE plus chemotherapy (n = 215)
Received TARE plus chemotherapy (n = 187)
Received TARE only (n = 0)
Received chemotherapy only (n = 16)
Did not receive any therapy (n = 12)
Patients assigned to chemotherapy only (n = 213)
Received chemotherapy only (n = 191)
Received TARE plus chemotherapy (n = 0)
Received TARE only (n = 0)
Did not receive any therapy (n = 22)
Patients discontinued or completed the trial (n = 215)
Died (n = 169)
Withdrawn by investigator or sponsor (n = 22)[a]
Withdrew consent (n = 21)
Administrative reasons (n = 2)
Lost to follow-up (n = 1)
Patients discontinued or completed the trial (n = 213)
Died (n = 153)
Withdrew consent (n = 32)
Withdrawn by investigator or sponsor (n = 24)[a]
Administrative reasons (n = 1)
Lost to follow-up (n = 1)
Because of adverse event (n = 1)
Disease progression and patient withdrawal before survival status determination (n = 1)
Patients included in intention-to-treat population (n = 215)
Patients included in the safety analysis (n = 187)
Patients included in intention-to-treat population (n = 213)
Patients included in the safety analysis (n = 207)[b]

Figure 2.1: CONSORT diagram showing subject enrollment, treatment allocation, patient disposition, and data analysis in the EPOCH trial [6].

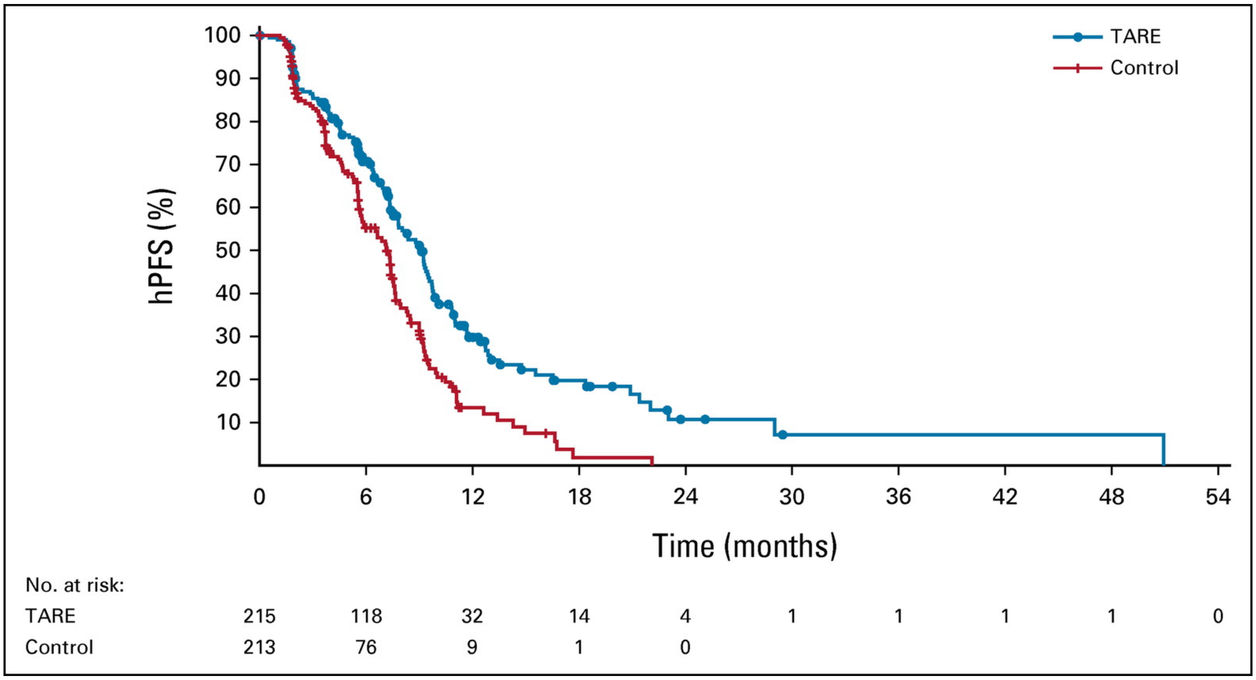
The addition of TARE to second-line chemotherapy resulted in significantly longer PFS and hPFS compared to chemotherapy alone (Figure 2.2, Figure 2.3) [6]. The HR for PFS was 0.69 (95% CI, 0.54 to 0.88; 1-sided $P = .0013$), with a median PFS of 8.0 months in the TARE plus chemotherapy group versus 7.2 months in the chemotherapy alone group. The HR for hPFS was 0.59 (95% CI, 0.46 to 0.77; 1-sided $P < .0001$), with a median hPFS of 9.1 months and 7.2 months, respectively.

Despite the significant improvements in PFS and hPFS, the EPOCH trial revealed heterogeneity in patient responses to treatment, with some subgroups deriving greater benefit from the addition of TARE to second-line chemotherapy than others. Furthermore, there was no significant difference in overall survival between the TARE plus chemotherapy and chemotherapy alone groups. Further subgroup analyses are needed to identify patient populations who may benefit most from TARE to guide personalized treatment decisions.



No. at risk:									
TARE	215	111	29	13	2	1	1	1	0
Control	213	76	9	1	0				

Figure 2.2: Kaplan-Meier analysis of overall PFS for TARE plus chemotherapy versus chemotherapy in the intention-to-treat population. [6]



No. at risk:										
TARE	215	118	32	14	4	1	1	1	1	0
Control	213	76	9	1	0					

Figure 2.3: Kaplan-Meier analysis of hPFS for TARE plus chemotherapy versus chemotherapy in the intention-to-treat population.

2.5 Heterogeneous Patient Responses and the Need for Subgroup Analyses

The EPOCH trial results revealed significant heterogeneity in patient responses to treatment, underscoring the importance of conducting subgroup analyses to identify patient populations that may derive greater benefit from TARE. While the overall response rate (ORR) was significantly higher in the TARE plus chemotherapy group compared to chemotherapy alone (34.0% vs. 21.1%; 1-sided P = .0019), there was no significant difference in overall survival (OS) between the two groups (median OS 14.0 vs. 14.4 months; HR 1.07; 95% CI, 0.86 to 1.32; 1-sided P = .7229) [6]. This discrepancy between ORR and OS highlights the complex nature of treatment responses in mCRC and the need for further investigation into factors influencing patient outcomes.

Subgroup analyses in the EPOCH trial revealed that the progression-free survival (PFS) benefit with TARE was more pronounced in specific patient subgroups, including those with KRAS mutant tumors (HR 0.57; 95% CI, 0.40 to 0.80), left-side primary tumors (HR 0.65; 95% CI, 0.48 to 0.88), hepatic tumor burden of 10%-25% (HR 0.43; 95% CI, 0.26 to 0.72), 3 lesions (HR 0.33; 95% CI, 0.14 to 0.76), and resected primary tumors (HR 0.63; 95% CI, 0.46 to 0.85) [6]. These findings suggest that certain patient and tumor characteristics may influence the efficacy of TARE in combination with second-line chemotherapy, and that a one-size-fits-all approach may not be optimal for managing mCRC.

The heterogeneity in treatment responses observed in the EPOCH trial is consistent with the complex biology and molecular landscape of mCRC. Colorectal tumors exhibit significant genetic and epigenetic alterations, such as mutations in KRAS, NRAS, BRAF, and microsatellite instability, which can influence prognosis and response to targeted therapies [16]. Additionally, the primary tumor location (left-sided vs. right-sided) has been shown to impact clinical outcomes and treatment efficacy in mCRC [16]. These biological factors, along with patient-specific characteristics such as tumor burden and number of lesions, likely contribute to the heterogeneous responses to TARE and chemotherapy observed in the EPOCH trial. Further research is necessary to validate these subgroups and to identify additional biomarkers and clinical factors that can guide personalized treatment strategies in mCRC.

Chapter 3

Leveraging biomarkers from the EPOCH trial

3.1 Explanation of Biomarkers and Response Measurements from the EPOCH Trial

The EPOCH trial collected an extensive array of biomarkers to evaluate the treatment response and progression in patients with metastatic colorectal cancer (mCRC). We categorized these biomarkers into those directly related to tumor characteristics and those associated with treatment response. Here, we provide examples of biomarkers' biological relevance and their potential as predictive or prognostic indicators.

Carcinoembryonic antigen (CEA) is a glycoprotein involved in cell adhesion, commonly elevated in colorectal cancer patients. It is used as a tumor marker for monitoring disease progression and treatment response. Elevated CEA levels are associated with tumor burden and metastatic potential. Monitoring CEA levels can help assess the effectiveness of treatments like TARE and chemotherapy [17].

KRAS mutation status is another crucial biomarker. KRAS is a gene encoding a protein involved in cell signaling pathways that regulate cell growth and apoptosis. Mutations in KRAS are common in colorectal cancer and can influence treatment outcomes. KRAS mutations are associated with resistance to anti-EGFR therapies and can affect prognosis and response to treatments. Patients with KRAS mutations may respond differently to TARE [18].

Tumor size and burden, as indicated by the maximum liver lesion size and liver tumor burden percentage, are also important biomarkers collected in the EPOCH trial. Larger tumor sizes and higher tumor burdens are often associated with worse prognosis and can significantly impact treatment planning and outcomes [9]. In addition to these, the EPOCH trial collected data on other relevant biomarkers such as NRAS and BRAF mutations, which are known to influence the behavior of colorectal cancers and their response to various treatments. While NRAS mutations occur in a smaller percentage of colorectal cancer cases, they are critical for understanding resistance mechanisms to targeted therapies. BRAF mutations, particularly the V600E variant, are associated with poor prognosis and aggressive disease [16].

The trial also included response-related biomarkers such as overall survival (OS), progression-free survival (PFS), and hepatic progression-free survival (hPFS). These metrics are essential for evaluating the effectiveness of treatments and understanding patient responses. Overall survival (OS) measures the time from the start of treatment until death from any cause, providing a comprehensive outcome measure. Progression-free survival (PFS) is the length of time during and after treatment that a patient lives with the disease without it worsening, reflecting the efficacy of the treatment in controlling the disease. Hepatic progression-free survival (hPFS) specifically focuses on the liver metastases, and is a primary study outcome-of-interest given the liver-dominant nature of mCRC [6].

3.2 Integrating Biomarkers into Machine Learning Models

Biomarkers from the EPOCH trial, such as carcinoembryonic antigen (CEA) levels, KRAS mutation status, tumor size, and progression-free survival (PFS), can serve as critical input features for supervised ML algorithms, which can discern patterns and associations between the biomarkers and clinical outcomes.

To construct predictive models, feature vectors are created from the collected biomarker data. These vectors enable the training of ML models to predict outcomes such as overall survival (OS) and treatment efficacy. Advanced feature selection and engineering techniques are employed to identify the most relevant biomarkers to enhance model accuracy. For example, principal component analysis (PCA) can be used to reduce the dimensionality of the data, highlighting the most salient features and improving predictive power. Additionally, extracted features from image data, such as those obtained from CT or MRI scans, can be used to augment the existing feature vector and improve the accuracy and robustness of the predictive models, offering a more comprehensive understanding of the tumor characteristics and treatment response.

Chapter 4

Technical Background

We rely on an array of computational techniques to uncover patterns and relationships within the complex, multi-modal dataset. This chapter provides an overview of the key methods employed in our approach, including supervised machine learning algorithms, multi-modal models, embedding spaces from deep convolutional models, dimensionality reduction and visualization techniques, clustering methods, and survival analysis using Cox proportional hazards modeling.

4.1 Supervised Machine Learning

Supervised machine learning is a powerful tool for predicting outcomes based on labeled input data. At its core, supervised learning involves training a model to map input features to corresponding output labels, enabling the model to make predictions on new, unseen data. To illustrate the fundamental concepts of supervised learning, we begin with a simple example: linear regression.

In linear regression, the goal is to find a linear relationship between input features $\mathbf{x} = (x_1, \ldots, x_n)$ and a continuous output variable y. The model is defined by a set of weights $\mathbf{w} = (w_1, \ldots, w_n)$ and a bias term b, such that:

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b \tag{4.1}$$

where $\hat{y}$ is the predicted output. The objective is to find the optimal weights and bias that minimize the difference between the predicted and actual outputs, typically measured using mean squared error (MSE):

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{4.2}$$

where m is the number of training examples, and y_i and $\hat{y}_i$ are the actual and predicted outputs for the i-th example, respectively.

While linear regression is effective for simple problems, many real-world applications, such as those in the EPOCH trial, require more powerful, non-linear models. Artificial neural networks (ANNs), also known as deep learning models, are a class of supervised learning algorithms inspired by the structure and function of biological neural networks.

ANNs consist of interconnected layers of artificial neurons, or nodes, that process and transmit information. Each neuron in a layer receives weighted inputs from the previous layer, applies a non-linear activation function, and passes the output to the next layer. The basic computation performed by a single neuron can be expressed as:

$$a_j = \sigma\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) \tag{4.3}$$

where a_j is the activation of the j-th neuron, σ is the activation function (e.g., sigmoid, ReLU), w_{ij} is the weight connecting the i-th input to the j-th neuron, x_i is the i-th input, and b_j is the bias term for the j-th neuron.

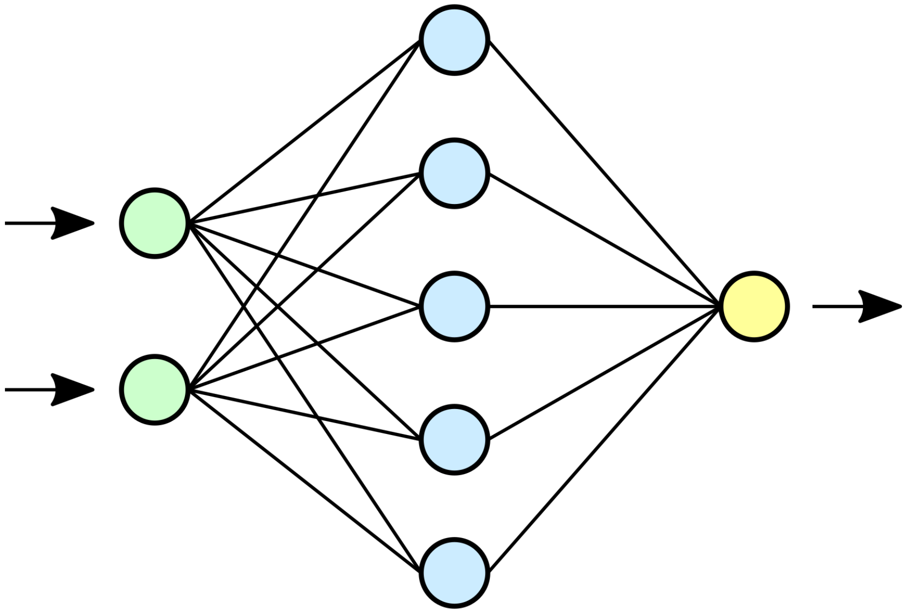


Figure 4.1: A simple artificial neural network.

Figure 4.1 depicts a simple ANN with an input layer, one hidden layers, and an output layer. The goal of training an ANN is to find the optimal weights and biases that minimize the difference between the predicted and actual outputs, similar to linear regression. However, due to the non-linear nature of ANNs, the optimization process is more complex and typically involves gradient-based methods, such as backpropagation.

The training process for ANNs in supervised learning involves the following steps:

1. Forward propagation: Input data is fed through the network, and the output is computed using the current weights and biases.

2. Loss computation: The difference between the predicted and actual outputs is measured using a loss function, such as cross-entropy for classification tasks or mean squared error for regression tasks.

3. Backpropagation: The gradients of the loss function with respect to the weights and biases are computed using the chain rule of differentiation. These gradients indicate the direction in which the parameters should be updated to minimize the loss.

4. Parameter update: The weights and biases are updated using an optimization algorithm, such as stochastic gradient descent (SGD), which takes a step in the direction of the negative gradient to minimize the loss.

These steps are repeated iteratively until the model converges or a predefined stopping criterion is met.

For categorical classification tasks, such as those encountered for clinical features in the EPOCH trial, the output layer of the ANN typically consists of a softmax activation function, which produces a probability distribution over the possible classes. The cross-entropy loss function is then used to measure the difference between the predicted and actual class probabilities:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{m} \sum_{j=1}^{c} y_{ij} \log(\hat{y}_{ij}) \tag{4.4}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the actual and predicted class probabilities, respectively, m is the number of training examples, and c is the number of classes.

By minimizing the cross-entropy loss during training, the ANN learns to map input features to the correct class probabilities, enabling accurate predictions on new, unseen data. Supervised learning techniques can be employed to predict patient outcomes, such as overall survival or treatment response, based on input features derived from biomarkers, clinical data, and imaging data.

4.2 Deep Convolutional Neural Networks

Deep convolutional neural networks (CNNs) are a specialized type of ANN designed for processing grid-like data, such as images and videos. CNNs have achieved state-of-the-art performance in various computer vision tasks, including image classification, object detection, and semantic segmentation. The key characteristic of CNNs is their ability to learn hierarchical features from raw input data through a series of convolutional layers, pooling layers, and fully connected layers.

Convolutional layers are the core building blocks of CNNs. They consist of a set of learnable filters, or kernels, that slide over the input data, performing element-wise multiplications and summing the results to produce feature maps. The operation performed by a single convolutional layer can be expressed as:

$$\mathbf{Y}_{i,j,k} = \sum_{m} \sum_{n} \sum_{c} \mathbf{W}_{m,n,c,k} \cdot \mathbf{X}_{i+m,j+n,c} + b_k \tag{4.5}$$

where $\mathbf{Y}_{i,j,k}$ is the output value at position (i, j) in the k-th feature map, $\mathbf{W}_{m,n,c,k}$ is the weight at position (m, n) in the c-th input channel and k-th output channel, $\mathbf{X}_{i+m,j+n,c}$ is the input value at position $(i + m, j + n)$ in the c-th input channel, and b_k is the bias term for the k-th output channel.

Pooling layers are used to downsample the feature maps produced by convolutional layers, reducing the spatial dimensions of the data while retaining the most important information. The most common types of pooling are max pooling and average pooling, which compute the maximum and average values, respectively, within a specified window size.

ResNet (Residual Network) is a popular CNN architecture that introduced the concept of residual connections to address the vanishing gradient problem in deep networks. ResNets consist of a series of residual blocks, each containing convolutional layers and a skip connection that allows the input to bypass the layers and be added directly to the output. This skip connection enables the network to learn residual functions, which are easier to optimize than the original mapping.

4.2.1 2D ResNet Architecture

The 2D ResNet architecture is designed for processing 2D input data, such as images. A typical 2D ResNet consists of an initial convolutional layer followed by a series of residual blocks [19], each containing two or three convolutional layers and a skip connection (4.2 [20]). The output of the last residual block is then passed through a global average pooling layer and a fully connected layer for classification. The key components of a 2D ResNet architecture are:

1. Initial convolutional layer: This layer applies a set of learnable filters to the input image, producing a set of feature maps.

2. Residual blocks: Each residual block consists of two or three convolutional layers, followed by batch normalization and activation functions (e.g., ReLU). The input to the block is added to the output of the last convolutional layer via a skip connection, which allows the network to learn residual functions.

3. Global average pooling layer: This layer reduces the spatial dimensions of the feature maps produced by the last residual block, yielding a fixed-size vector representation of the input image.

4. Fully connected layer: This layer takes the output of the global average pooling layer and performs classification or regression tasks by learning a set of weights that map the feature vector to the desired output classes or values.

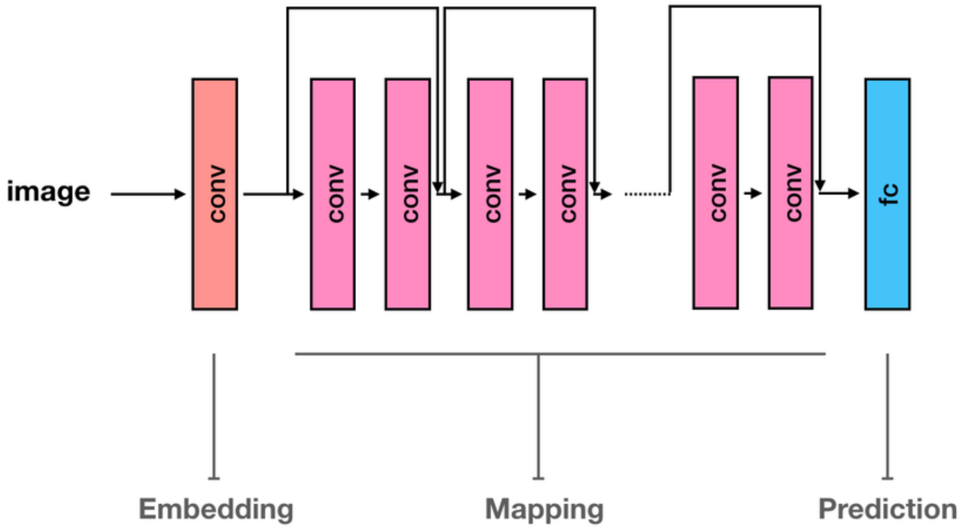


Figure 4.2: Illustration of the 2D ResNet architecture for image classification.[20].

4.2.2 3D ResNet Architecture

The 3D ResNet architecture extends the 2D ResNet to process 3D input data, such as volumetric medical images or video sequences. In a 3D ResNet, the convolutional layers and residual blocks are replaced by their 3D counterparts, which operate on 3D input data using 3D convolutions and 3D pooling [21].

The main components of a 3D ResNet architecture are similar to those of a 2D ResNet, with the following modifications:

1. **3D convolutional layers:** These layers apply 3D filters to the input data, capturing spatial and temporal information simultaneously.

2. **3D residual blocks:** Each 3D residual block consists of two or three 3D convolutional layers, followed by batch normalization and activation functions. The input to the block is added to the output of the last 3D convolutional layer via a skip connection.

3. **3D pooling layers:** These layers downsample the 3D feature maps produced by the 3D convolutional layers, reducing the spatial and temporal dimensions of the data.

4. **Global average pooling layer and fully connected layer:** Similar to the 2D ResNet, these layers are used for classification or regression tasks, but they operate on the 3D feature maps produced by the last 3D residual block.

The 3D ResNet architecture is particularly well-suited for processing volumetric medical images, such as CT or MRI scans, as it can capture the spatial relationships between adjacent slices and learn 3D features that are relevant for predicting patient outcomes or identifying distinct patient subgroups.

4.3 Embedding Space and Feature Extraction

In addition to their use for classification and regression tasks, deep convolutional networks can be employed for feature extraction and representation learning. The idea is to use the activations of the intermediate layers of a pre-trained CNN as a compact, high-level representation of the input data, known as an embedding.

To extract embeddings from a CNN, the network is first trained on a large dataset for a specific task, such as image classification. Once trained, the fully connected layers used for classification are removed, and the activations of the last convolutional layer or pooling layer are used as the embedding vectors.

The extracted embeddings can be used for various downstream tasks, such as clustering, similarity search, or as input features for other machine learning models. In the context of the EPOCH trial, CNN embeddings can be used to represent the imaging data (e.g., CT or MRI scans) in a compact, high-level format that captures the most relevant features for predicting patient outcomes or identifying distinct patient subgroups. To leverage the embedding space for prediction tasks, the embeddings can be combined with other clinical and biomarker data and multimodal models can then be trained to map the combined input features to the desired output labels, such as overall survival or treatment response.

4.4 Dimensionality Reduction and Visualization

High-dimensional imaging data and extracted embeddings from machine learning models can pose challenges for traditional methods of data analysis, visualization, and interpretation. Dimensionality reduction techniques, such as t-Distributed Stochastic Neighbor Embedding (t-SNE), can be employed to address these challenges by projecting the high-dimensional data into a lower-dimensional space while preserving the important structural information and relationships between data points.

Dimensionality reduction refers to a class of techniques that transform high-dimensional data into a lower-dimensional representation while retaining the most important features and relationships between data points. The main objectives of dimensionality reduction are:

1. **Data compression:** Reducing the number of dimensions can help compress the data, making it more efficient to store, process, and transmit.
2. **Visualization:** Projecting high-dimensional data into a 2D or 3D space enables the visualization of complex relationships and patterns that would otherwise be difficult to perceive.
3. **Feature extraction:** Dimensionality reduction can help identify the most informative features or combinations of features that contribute to the underlying structure of the data.

4.4.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a popular non-linear dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data in a lower-dimensional space, typically 2D or 3D [22]. The main idea behind t-SNE is to preserve the local structure of the high-dimensional data in the low-dimensional representation, such that similar data points in the original space are mapped to nearby points in the low-dimensional space, while dissimilar data points are mapped to distant points. The t-SNE algorithm consists of two main steps:

1. **Similarity computation:** t-SNE computes the pairwise similarities between data points in the high-dimensional space using a Gaussian probability distribution. The similarity between two data points is proportional to the probability of observing one data point as a neighbor of the other, given a certain level of Gaussian noise.

2. **Low-dimensional embedding:** t-SNE maps the data points to a low-dimensional space, typically 2D or 3D, by minimizing the Kullback-Leibler (KL) divergence between the probability distributions in the high-dimensional and low-dimensional spaces. The KL divergence measures the difference between the two probability distributions, and by minimizing it, t-SNE ensures that the local structure of the data is preserved in the low-dimensional representation.

One of the key advantages of t-SNE is its ability to reduce the dimensionality of embeddings extracted from the imaging data to visualize patient subgroups and treatment response patterns in a 2D or 3D space.

4.5 Cox Proportional Hazards (CoxPH) Modeling

The Cox Proportional Hazards (CoxPH) model is a fundamental statistical technique used in survival analysis to investigate the association between the survival time of patients and one or more predictor variables [23]. Unlike other regression models, CoxPH does not assume any specific baseline hazard function, making it a semi-parametric model. Instead, it focuses on the effect of covariates on the hazard rate.

The CoxPH model can be expressed mathematically as follows:

$$h(t|X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{4.6}$$

where $h(t|X)$ is the hazard function at time t given the covariates $X = (X_1, X_2, \ldots, X_p)$, $h_0(t)$ is the baseline hazard function, which represents the hazard when all covariates are zero, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients representing the effect of each covariate on the hazard function.

The hazard ratio (HR) is a key concept in CoxPH modeling, quantifying the effect of a covariate on the hazard rate. It is defined as the ratio of the hazard rates corresponding to different values of the covariate. For a given covariate X_j, the hazard ratio can be expressed as:

$$HR = \exp(\beta_j) \tag{4.7}$$

A hazard ratio greater than 1 indicates that an increase in the covariate is associated with an increased hazard rate (i.e., decreased survival time), while a hazard ratio less than 1 indicates a decreased hazard rate (i.e., increased survival time).

The concordance index (C-index) is a measure used to evaluate the predictive accuracy of the CoxPH model. It quantifies the degree of concordance between the predicted and actual survival times. The C-index ranges from 0.5 to 1, where a C-index of 0.5 indicates no predictive ability (random prediction) and a C-index of 1 indicates perfect prediction. The C-index is computed by considering all pairs of patients and determining the proportion of concordant pairs, where a pair is concordant if the patient with the higher predicted risk (hazard) actually experiences the event (e.g., death) before the patient with the lower predicted risk. Mathematically, the C-index is calculated as:

$$C = \frac{\text{Number of concordant pairs}}{\text{Number of comparable pairs}} \tag{4.8}$$

CoxPH modeling can be utilized to analyze the impact of various biomarkers and clinical covariates on patient survival outcomes [23]. By fitting a CoxPH model to the trial data, we can identify significant predictors of survival and estimate their effects using hazard ratios to better understand the heterogeneous patient responses to treatment. Importantly, we gain additional understanding on the effect of each covariate on the hazard rate.

4.6 Classification Methods Used for Feature Prediction

Several classifiers were employed to predict patient outcomes using biomarker feature vectors and image embeddings. These include XGBoost, Linear Discriminant Analysis (LDA), Logistic Regression, and Multilayer Perceptron (MLP).

- **XGBoost**: Ensemble learning algorithm that combines multiple decision trees. Iteratively trains trees to correct errors, minimizing a regularized objective function. Handles high-dimensional data and captures complex feature interactions [24].

- **Linear Discriminant Analysis (LDA)**: Supervised learning algorithm for classification and dimensionality reduction. Finds a linear combination of features to maximize between-class variance and minimize within-class variance, assuming Gaussian distribution and equal class covariance matrices.

- **Logistic Regression**: Statistical model for binary classification. Models the probability of an instance belonging to a class using a logistic function of a linear combination of input features. Coefficients are estimated using maximum likelihood estimation.

- **Multilayer Perceptron (MLP)**: Feedforward artificial neural network with input, hidden, and output layers. Nodes compute weighted sums of inputs and apply non-linear activation functions. Trained using backpropagation to learn complex, non-linear relationships between input features and output variables.

Chapter 5

Model Architectures and Algorithmic Tools

This chapter presents the computational tools and model architectures employed in our predictive modeling approach. In particular, we detail the pre-trained model architectures utilized to generate meaningful embeddings from imaging data and segmentation models used for pre-processing.

5.1 Liver Region Restriction with TotalSegmentator

The CT and MRI scans obtained from the EPOCH trial encompass various anatomical regions such as the chest, abdomen, and pelvis. However, TARE therapy specifically targets the liver, making it essential to restrict the analysis to liver-specific features, requiring a preprocessing step to isolate the liver region from the rest of the scan.

TotalSegmentator, a deep learning-based tool, was employed for this task due to its robust and comprehensive segmentation capabilities. TotalSegmentator was developed to automatically segment 104 anatomical structures, including 27 organs, 59 bones, 10 muscles, and 8 vessels, from 3D CT volumes [25].

TotalSegmentator was trained on a diverse dataset of 1204 CT examinations, which included a wide range of clinical data with significant abnormalities. This dataset represents real-world conditions, including different ages, abnormalities, scanners, body parts, sequences, and sites, making it well suited for heterogeneity in patient scans and ages in the EPOCH dataset. The model achieved a Dice similarity coefficient of 0.943 on the test set indicating sufficient robustness and accuracy in segmentation [25].

The underlying architecture of TotalSegmentator is based on nnU-Net, which automatically configures itself for a given dataset by optimizing hyperparameters, network architecture, and training strategies. The nnU-Net framework includes a 3D U-Net architecture that processes the volumetric CT data in three dimensions, capturing spatial relationships and features essential for accurate segmentation [26].

The heterogeneity of the CT scans poses a significant challenge because scans vary widely, covering different body parts and often extending beyond the region of interest. By using TotalSegmentator, we can accurately segment the liver and then ensure that only the relevant slices containing the liver are included downstream for model training and embedding generation. We applied TotalSegmentator on entire CT and MRI scan volumes, then used the output segmentation mask to extract the slices that contain the liver. This segmented liver region is then used as the input for further analysis, ensuring that the models are trained on and generate predictions based on the most relevant anatomical features.

5.2 Multi-Dimensional Subset Scanning

Multi-Dimensional Subset Scanning (MDSS) is a technique used for identifying statistically significant subsets within high-dimensional data [27]. The primary purpose of MDSS in our case is to facilitate automatic stratification and subgroup analysis, enabling the identification of patient subgroups that exhibit distinct characteristics or responses to treatment. To implement MDSS, we utilize the AI Fairness 360 (AIF360) toolkit, which includes robust tools for detecting and mitigating bias in machine learning models and an MDSS detector that can be adapted for our needs [28].

The MDSS approach is mathematically grounded in the optimization of a scoring function over all possible subsets of the data. The goal is to identify subsets where the observed outcomes deviate significantly from the expected outcomes under the null hypothesis. The scoring function is designed to detect these deviations while controlling for multiple hypothesis testing.

The scoring function $S(S)$ for a subset S of the data can be defined as:

$$S(S) = \frac{O(S) - E(S)}{\sqrt{V(S)}} \tag{5.1}$$

where:

- $O(S)$ is the observed count of the outcome of interest in the subset S.
- $E(S)$ is the expected count of the outcome under the null hypothesis in the subset S.
- $V(S)$ is the variance of the outcome count under the null hypothesis in the subset S.

Our MDSS analysis involves preprocessing the data by normalizing numerical features and encoding categorical features. We define the scoring function based on the specific outcomes and features of interest, aiming to identify subsets with significant differences in treatment response. The MDSS algorithm is applied to search over all possible subsets of the data, iteratively evaluating the scoring function for each subset and identifying the subset that maximizes the score. To determine the statistical significance of the identified subset, we perform a permutation test by randomly permuting the outcomes multiple times to generate a null distribution of the scoring function, recalculating the scoring function for each permutation, and comparing the score of the identified subset to the null distribution to compute a p-value. A low p-value suggests that the identified subset is statistically significant.

5.3 Pre-trained Model Architectures

Two pre-trained model architectures were leveraged to generate embeddings from imaging data in the EPOCH trial: a 2-Dimensional Combined-Modality Model with 3D integration and the 3D MedicalNet backbone.

The first model we utilized is a 2-Dimensional Combined-Modality Model, built using IBM's proprietary, extensive internal dataset to pre-train a robust neural network using a self-supervised learning framework known as DINO (Distillation with No Labels) [29]. The DINO framework uses a contrastive learning approach to train a student-teacher network without requiring labeled data. This framework comprises two neural networks: a student network and a teacher network. Both networks process input images to produce feature embeddings, but only the student network's weights are updated during training. The teacher network's weights are updated using an exponential moving average (EMA) of the student network's weights. The objective is to minimize the cross-entropy loss between the student and teacher network outputs, encouraging the student network to produce similar embeddings to the teacher network. This process allows the model to learn meaningful representations from the data, with the goal of understanding local and global structural features across a wide set of liver scans. Figure 5.1 illustrates the Combined-Modality Model architecture.

The backbone of our Combined-Modality Model is a 2D ResNet-18. The model was pre-trained on a substantial dataset comprising 3,081 CT liver subjects (16,385 scans) and 2,117 liver MRI subjects (31,051 scans). This extensive pre-training enables the model to capture diverse anatomical features relevant to liver imaging. To adapt the 2D ResNet-18 model for 3D volumetric data, we employed a Sliced3D architecture [30]. The Sliced3D approach processes each slice of the 3D volume independently through the 2D ResNet-18

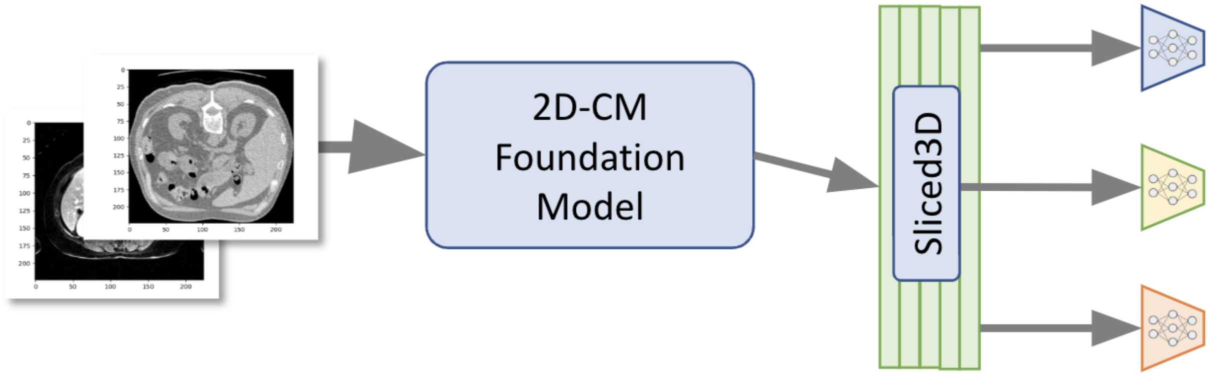


Figure 5.1: Combined-Modality Model leveraging 2D ResNet-18 with Sliced3D for volumetric data processing.

backbone, generating a tensor of dimension slice size x 512. Then, the model integrates these slice-level features over the entire 3D volume by stacking the features from each slice and applying a series of convolutional and pooling operations to aggregate information in the depth dimension, ensuring the preservation of spatial information across slices. The resulting 512-dimensional feature vector effectively represents the entire 3D volume, capturing both local and global anatomical structures. Finally, a multilayer perceptron (MLP) head is used for classification tasks, mapping the 512-dimensional feature vector to the desired output classes.

The training process for the 2-Dimensional Combined-Modality Model involved optimizing the model weights using the DINO framework. The cross-entropy loss was minimized between the student and teacher network outputs as shown in Figure 5.2. Training loss plateaued after 25k iterations, with a mirrored decrease in learning rate as shown in Appendix Figure A.1.

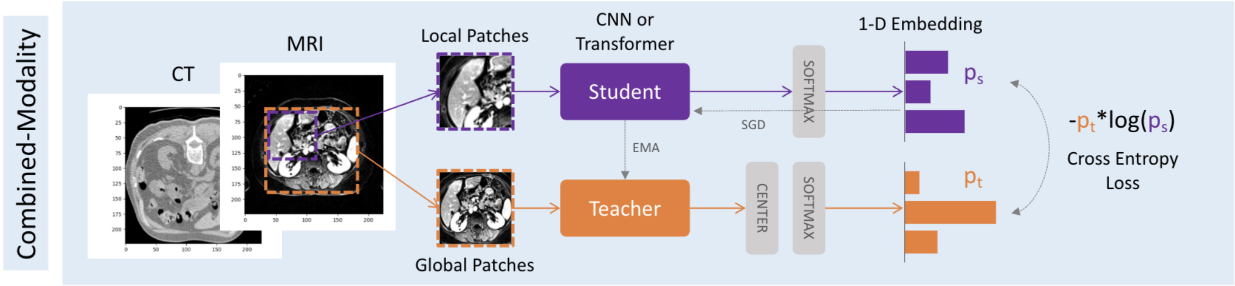


Figure 5.2: Training progress of the 2-Dimensional Combined-Modality Model using the DINO framework.

The second model we employed is based on the MedicalNet framework, specifically designed for 3D medical image analysis. The MedicalNet model utilizes a 3D ResNet-18 backbone with additional decoder layers for segmentation tasks [31]. This architecture was pre-trained with a supervised learning approach to minimize cross-entropy loss for segmentation tasks on the 3DSeg-8 dataset, a large-scale aggregation of multiple 3D medical image datasets covering various imaging modalities such as MRI and CT and encompassing different scan regions, target organs, and pathological manifestations [31]. The pre-trained models from MedicalNet demonstrate significant improvements in training convergence speed and accuracy compared to models trained from scratch or pre-trained on natural image datasets.

The 3D ResNet-18 backbone in the MedicalNet model captures 3D spatial information from volumetric medical images. The model architecture includes an encoder-decoder structure, where the encoder consists of a series of 3D convolutional layers that extract high-level features from the input volume. The decoder layers are used for segmentation tasks, but for our classification purposes, we adapted the model by utilizing only the encoder part for feature extraction. We added adaptive average pooling and two fully connected layers to map the 512-dimensional feature vector to the output classes.

The embeddings generated by both the IBM Pretrained Combined-Modality Model and the MedicalNet backbone are integrated into our predictive modeling pipeline. These embeddings are combined with structured data, such as biomarkers and clinical features, to form comprehensive feature vectors, which are then used to train various machine learning models, including classifiers and survival models, to predict patient outcomes and identify clinically relevant subgroups.

By leveraging these pre-trained models, we can extract high-level features from medical images that capture the underlying anatomical and pathological characteristics. This approach enhances the predictive performance of our models and enables the identification of meaningful patterns and subgroups within the heterogeneous patient population in the EPOCH trial.

We used the FuseMedML pipeline to streamline the image preprocessing and model fine-tuning processes. FuseMedML is a comprehensive framework designed for machine learning-based discovery in the biomedical domain, providing flexible and reusable components for data processing, model training, and evaluation [32]. For image preprocessing, we utilized the ‘fuse.data‘ package to build a robust data pipeline. This involved loading and normalizing imaging data, applying augmentation techniques, and caching processed data to optimize runtime. The framework’s generic data object design, which stores data in a hierarchical dictionary (NDict), allowed us to efficiently handle and preprocess the multi-modal imaging data. The flexible configuration options in FuseMedML facilitated the customization of training loops, loss functions, and evaluation metrics, allowing us to fine-tune our pre-trained models for downstream classification tasks easily.

Chapter 6

Data Overview and Preprocessing

6.1 Dataset Overview

The EPOCH trial dataset encompasses a total of 426 patients [6], with the majority of participants from the United Kingdom and the United States. Figure 6.1 illustrates the breakdown of patient numbers by country, highlighting the significant representation from these two nations. Among these patients, 208 individuals have response assessments with paired images, which form the foundation for our downstream analysis.

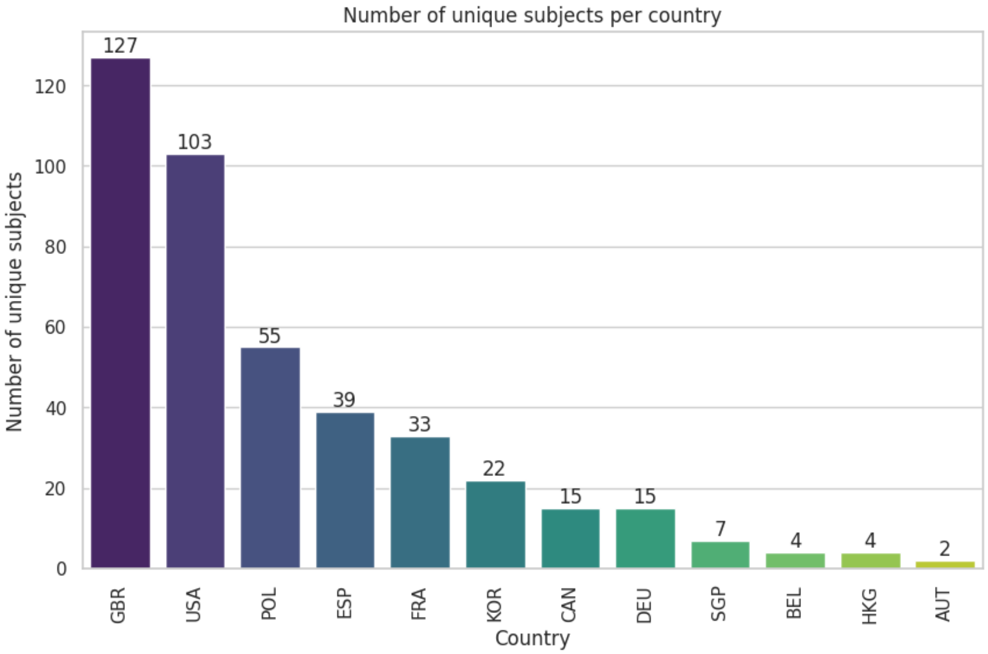


Figure 6.1: Breakdown by country for number of patients enrolled in the EPOCH clinical trial.

The US-UK dataset comprises 3,758 unique patient visits, translating to 5,558 image scans following the application of preprocessing techniques and data filtering. This substantial volume of data provides a robust basis for our investigation into patient subgroups and treatment response patterns.

6.1.1 Baseline Biomarkers

The dataset includes a comprehensive set of baseline biomarkers, capturing essential aspects of tumor characteristics, cancer status, and treatment history with chemotherapies. These biomarkers serve as key features for our predictive modeling and stratification efforts. Table 6.1 presents a selected set of baseline biomarkers that are of particular interest in our analysis.

Variable	Label
Bilirubin (umol/L)_AVAL_lb	Baseline Value
Carcinoembryonic Antigen (ug/L)_AVAL_lb	Baseline Value
ECOGBL_adsl	ECOG at Baseline
MCRCSTC_adsl	MCRC Status corrected (Correct value of Bilobar or unilobar cancer present)
KRASSTC_adsl	KRAS Status Corrected (Tumor biomarker, either KRAS-wild type or KRAS-mutant type)
CH1ADMC_adsl	1st Line Chemo Administered Corrected (Oxaliplatin based or Irinotecan based)
NUMLES_adsl	Number of Lesions at Baseline
SLCHEM_adsl	Second Line Chemotherapy (Irinotecan or Oxaliplatin)
POTSTG_adsl	Stage at Initial Diag of mCRC (Cancer stage: II, III, IV based on NJM classification)
QVAL_adsl	Total lesion volume (ml)

Table 6.1: Simplified table of baseline features with labels and relevant details.

6.1.2 Response Biomarkers and RECIST Criteria

In addition to the baseline biomarkers, the dataset incorporates response biomarkers that align with the widely accepted Response Evaluation Criteria In Solid Tumors (RECIST) criteria [33]. These criteria are commonly employed in clinical trials to assess the efficacy of cancer treatments and provide standardized measures for evaluating treatment response.

Two key variables in the dataset, INTGRESP_res (Integrated Response) and IOVRLRES_res (Overall Response), are central to the RECIST criteria. INTGRESP_res reflects an integrated assessment of tumor response, considering factors such as target lesions, non-target lesions, and the emergence of new lesions. This variable categorizes responses into classes such as Non-CR/Non-PD (Non-Complete Response/Non-Progressive Disease), PD (Progressive Disease), CR (Complete Response), and NE (Not Evaluable) based on standardized measurements and changes in tumor size.

IOVRLRES_res, on the other hand, represents the overall response assessment, providing a comprehensive evaluation of the patient's response to treatment across all tumor sites. This variable classifies responses into categories like SD (Stable Disease), PD (Progressive Disease), PR (Partial Response), and CR (Complete Response), offering a holistic view of the treatment's impact on tumor burden.

The dataset also includes hPFS (Hepatic Progression-Free Survival), overall survival, time to deterioration, and duration of disease control, with response variates when applicable to the treatment condition. Additionally, for patients who did not experience disease progression, changes in values for selected baseline measurements are also recorded.

6.2 Data Preprocessing

To ensure the quality and consistency of the data used in our analysis, we performed a comprehensive preprocessing pipeline to address inconsistencies and standardize the imaging data inputs.

One of the primary challenges in working with medical imaging data is the presence of inconsistencies across scans. These inconsistencies can arise from differences in imaging protocols, equipment, and patient positioning. To mitigate these issues, we focused on eliminating scans that did not meet specific criteria. We removed scans that were not in the axial orientation, as well as scout scans, which are typically low-resolution images used for planning purposes. Additionally, we excluded scans with a body part field indicating only chest or pelvis, as our analysis focused on the liver region. The preprocessing pipeline began by compiling the image directory and enriching it with relevant DICOM metadata such as scan orientation, body part, and image quality, allowing us efficiently filter and select scans that met our inclusion criteria. We then harmonized the image directories with patient and subject data, restricting the analysis to scans with optimal image quality.

Another important aspect of preprocessing was addressing inconsistencies in slice dimensions within each scan. We observed that some scans contained slices with varying dimensions, which could pose challenges for image analysis algorithms. To overcome this issue, we developed an approach to remove inconsistent slice dimensions from each scan, ensuring that all slices within a scan had uniform dimensions. We also encountered scans where a single area was rescanned multiple times during a visit. These "rescans" introduced complications for image analysis, as they could lead to redundant or conflicting information. To address this, we implemented an algorithmic approach to detect rescans based on the Z position of the scan and timing information and subsequently separated the rescans into individual scans for use independently in downstream analyses.

To further refine the region of interest for each scan, we utilized the TotalSegmentator package to restrict the analysis to the slices in the DICOM file corresponding to the liver region. Figure 6.2 demonstrates the application of TotalSegmentator for liver segmentation on a sample CT scan, comparing the original slices with and without the presence of the liver.

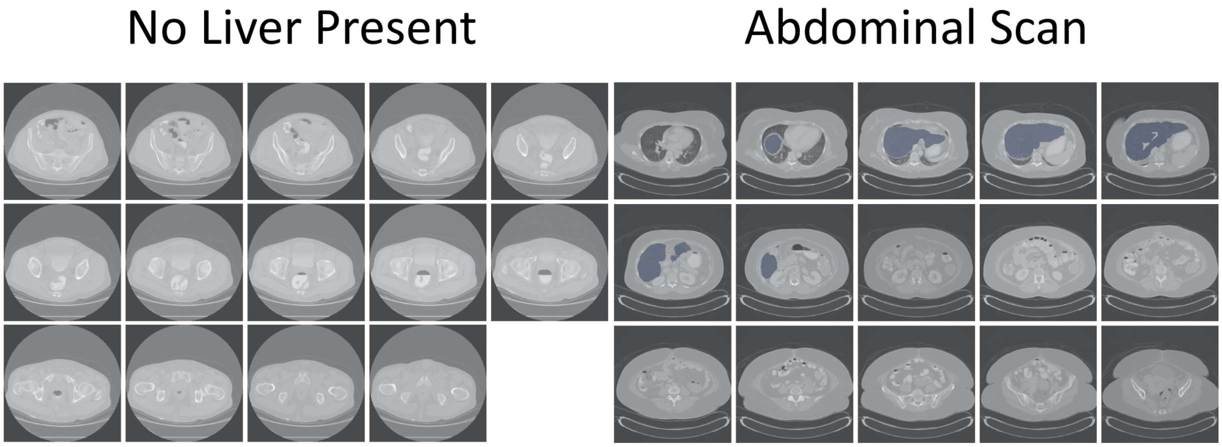


Figure 6.2: Representative example of two CT scans viewed from the axial plane, showing slices with no liver present (left) and the corresponding abdominal scan with the liver visible (right). The segmented liver region identified by TotalSegmentator is overlaid on the original CT slices.

After the initial preprocessing steps, the resulting set of scans underwent further preprocessing prior to being used in machine learning models. We employed the Fuse framework for efficient data caching and developed a pipeline to load the data from DICOM files. For MRI scans, we applied normalization techniques, such as truncating intensities in the top and bottom 5% of the range, to ensure comparability across scans. For CT scans, we clipped the intensity range based on Hounsfield units, typically from -500 to 500, to focus on the relevant tissue densities. Finally, we scaled the intensity values from the clipped range to a standardized range of 0 to 1, which is commonly used in machine learning models. Additionally, to ensure consistency of input into the machine learning models, the tensors were rescaled to a fixed size of $40x224x224$, representing the z, x, and y dimensions, respectively.

Chapter 7

Preliminary Patient Subgroup Identification and Outcome Prediction

7.1 Baseline Clustering of Patient Responses Using Structured Data

To establish a baseline understanding of patient subgroups and treatment response patterns, we performed unsupervised clustering using only the structured data and biomarkers available in the EPOCH trial dataset. This analysis aimed to identify distinct patient clusters based on their baseline characteristics and clinical features, without incorporating imaging data. We began by preprocessing the data, selecting relevant baseline and clinical features such as carcinoembryonic antigen (CEA) levels, bilirubin levels, age, ECOG performance status, and tumor burden measurements. These features were chosen based on their potential to capture meaningful differences among patients and their association with treatment outcomes. Next, we applied dimensionality reduction techniques to visualize the high-dimensional patient data in a lower-dimensional space. Specifically, we employed t-SNE from the python package scikit-learn to project the patient data onto a 2D plane while preserving the local structure and similarities between patients [34]. The t-SNE algorithm was trained using the selected baseline and clinical features, generating a compact representation of the patient population. Figure 7.1 illustrates the t-SNE visualization of patient data using baseline and clinical features, colored by outcome event categories (e.g., hepatic progression-free survival, time to deterioration, and overall survival).

The first set of t-SNE plots, which includes all subjects, reveals a lack of distinct clusters that correspond well to the clinical outcomes of interest, such as hepatic progression-free survival (hPFS), time to deterioration, and overall survival (OS). However, one cluster of

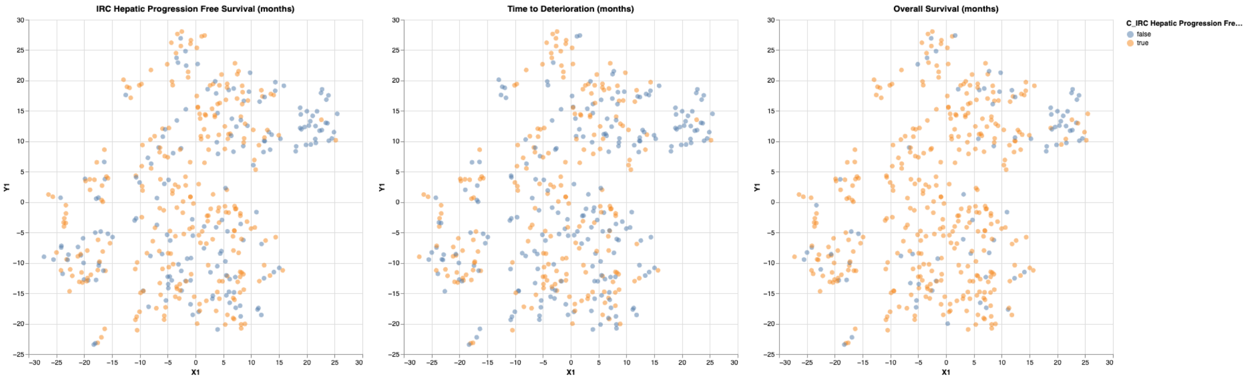


Figure 7.1: t-SNE visualization of patient data using baseline and clinical features, colored by outcome event categories (e.g., hepatic progression-free survival, time to deterioration, and overall survival).

patients exhibits adverse outcomes across all three measures, suggesting that this subgroup may have unique characteristics that contribute to their poor treatment response.

Next, we wanted to understand specifically the presence of patient subgroups within the US-UK cohort, for which we have imaging data, and generated a second set of t-SNE plots specific to this subset of patients. Figure 7.2 presents the t-SNE visualization of the US-UK cohort patient data using baseline and clinical features, colored by outcome categories (e.g., hepatic progression-free survival, time to deterioration, and overall survival). The t-SNE plots for the US-UK cohort show no clear separation based on the outcomes of interest, indicating that the structured data and biomarkers alone is sufficient to generate clusters of patient subgroups with different treatment responses. Incorporating additional features from imaging data is necessary to better characterize patient heterogeneity and predict clinical outcomes. In the next section, we will explore the use of Multi-Dimensional Subset Scanning

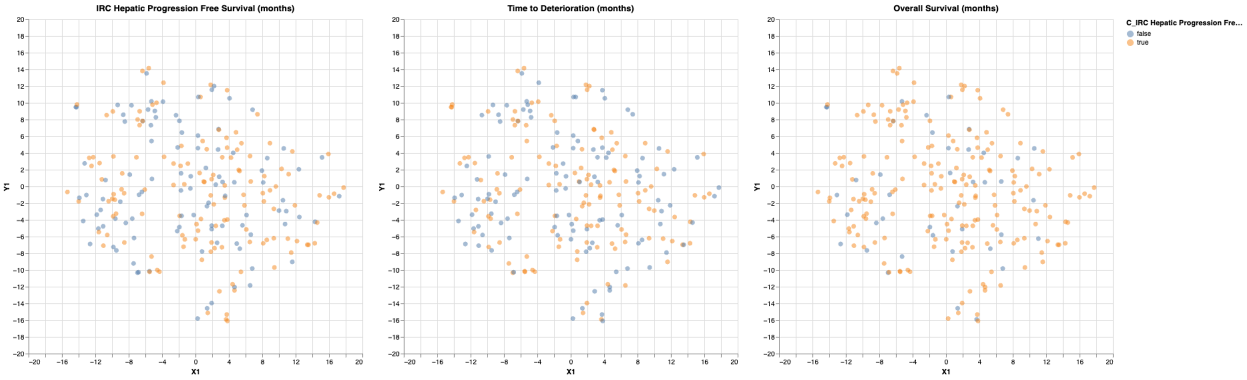


Figure 7.2: t-SNE visualization of the US-UK cohort patient data using baseline and clinical features, colored by outcome categories (e.g., hepatic progression-free survival, time to deterioration, and overall survival).

(MDSS) to identify statistically significant patient subgroups based on structured data and biomarkers.

7.2 Patient Subgroup Identification and Visualization using MDSS

To identify patient subgroups and visualize their characteristics, we applied Multidimensional Subset Scanning (MDSS) using the AI Fairness 360 (AIF360) toolkit to a selected subset of structured data from the baseline dataset [28]. The features and covariates used in this analysis were chosen based on their potential relevance to patient outcomes.

We focused on the IRC Hepatic Progression Free Survival (months)_E as the primary outcome of interest, designating an adverse event for hPFS as Objective Hepatic PD or death. We progressively restricted the set of covariate features by increasing the penalty to reduce the number of features used in subgroup identification. At a penalty of 1×10^{-6} we identified a 'poor response group' with specific characteristics. Table 7.1 presents the summary of the MDSS analysis results, including the expected value of the observed outcome, the size of the detected 'poor response group', and the group's characteristics.

Table 7.1: Summary of MDSS analysis results for identifying the 'poor response group'

Parameter	Value
Expected value of observed outcome	0.598592
Size of 'poor response group'	62
Observed average probability of event	0.9355
Expected probability of event	0.5986
Characteristics of 'poor response group'	
MCRCSTC_adsl	['Bilobar']
NUMLES_adsl	['3-5 lesions', '6-10 lesions', '>10 lesions']
ECOGBL_adsl	[1.0]
KRASSTC_adsl	['Mutant']

As shown in Table 7.1, the detected 'poor response group' has a size of 62 patients, with an observed average probability of an adverse event of 0.9355, compared to the expected probability of 0.5986. The group is characterized by bilobar metastatic colorectal cancer, a higher number of lesions (3 or more), an ECOG performance status of 1, and mutant KRAS status.

To visualize the difference in hPFS events over time between the 'poor response group' and the remaining patients, we plotted Kaplan-Meier curves for the two groups (Figure 7.3).

The curves demonstrate a clear separation, with the 'poor response group' experiencing a significantly higher rate of adverse events early on (log-rank test, $p < 0.005$).

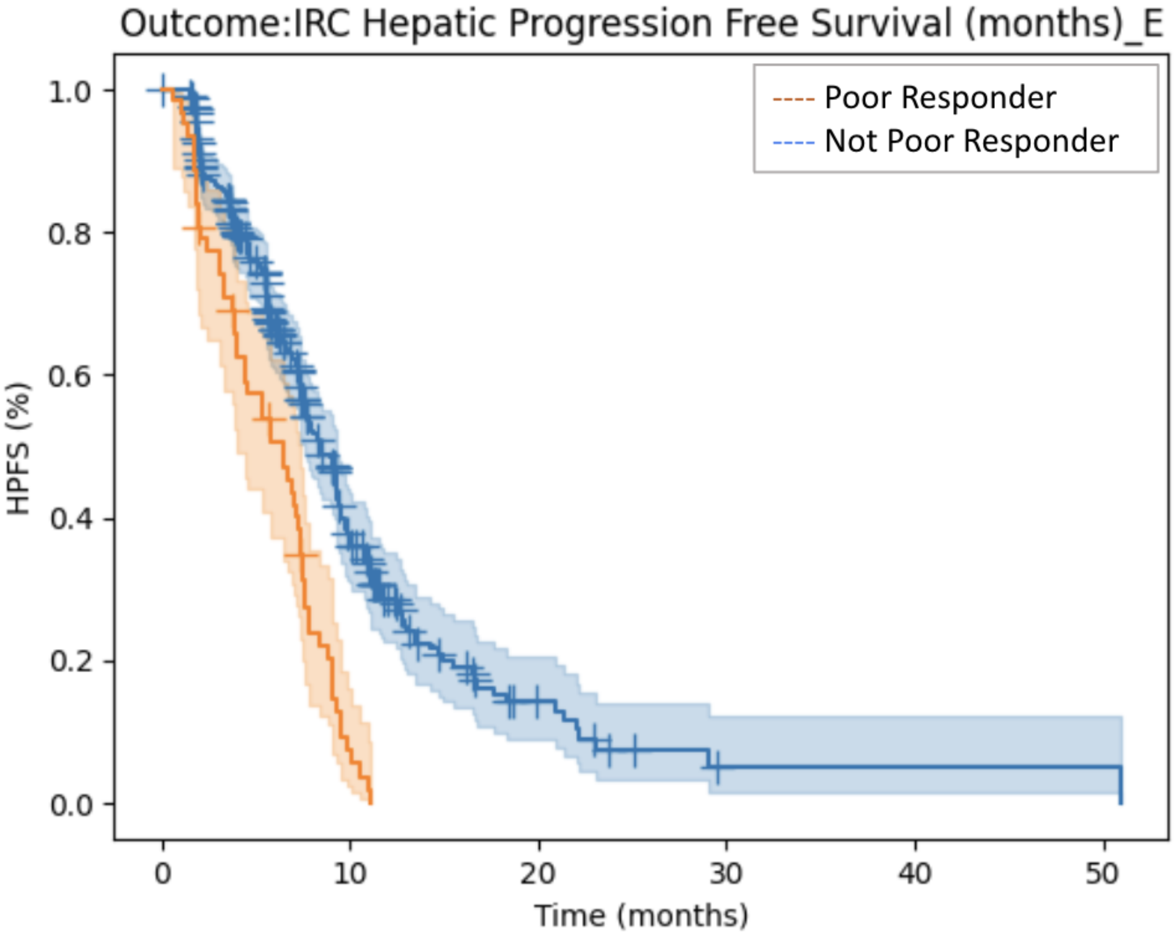


Figure 7.3: Kaplan-Meier curves comparing hPFS events between the 'poor response group' and the remaining patients.

Next, we explored the impact of the treatment condition (control vs. TheraSphere) on hPFS within the 'poor response group' and the remaining patients. Figure 7.4 presents the Kaplan-Meier curves for the 'poor response group', stratified by treatment condition. While the treatment condition had a statistically significant impact on hPFS outcomes for poor responders (log-rank test, $p = 0.01$), the difference was less pronounced compared to the remaining patients, as evidenced by the higher overlap of the curves.

In contrast, Figure 7.5 shows the Kaplan-Meier curves for patients not identified as poor responders, stratified by treatment condition. The curves exhibit less overlap over time, suggesting that the treatment condition may have a more pronounced effect on hPFS outcomes when poor responders are excluded (log-rank test, $p < 0.005$).

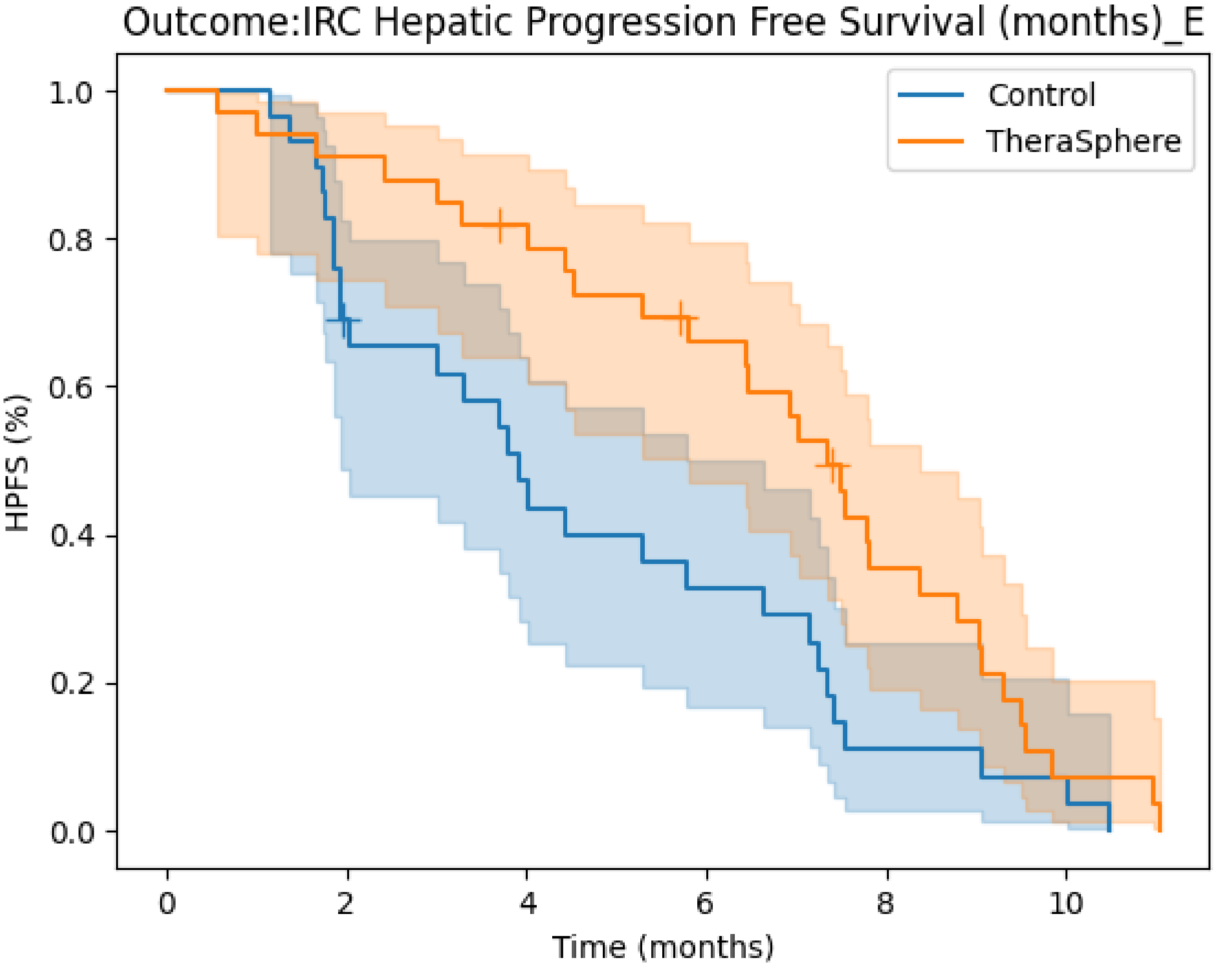


Figure 7.4: Kaplan-Meier curves comparing hPFS events within the ’poor response group’, stratified by treatment condition (control vs. TheraSphere).

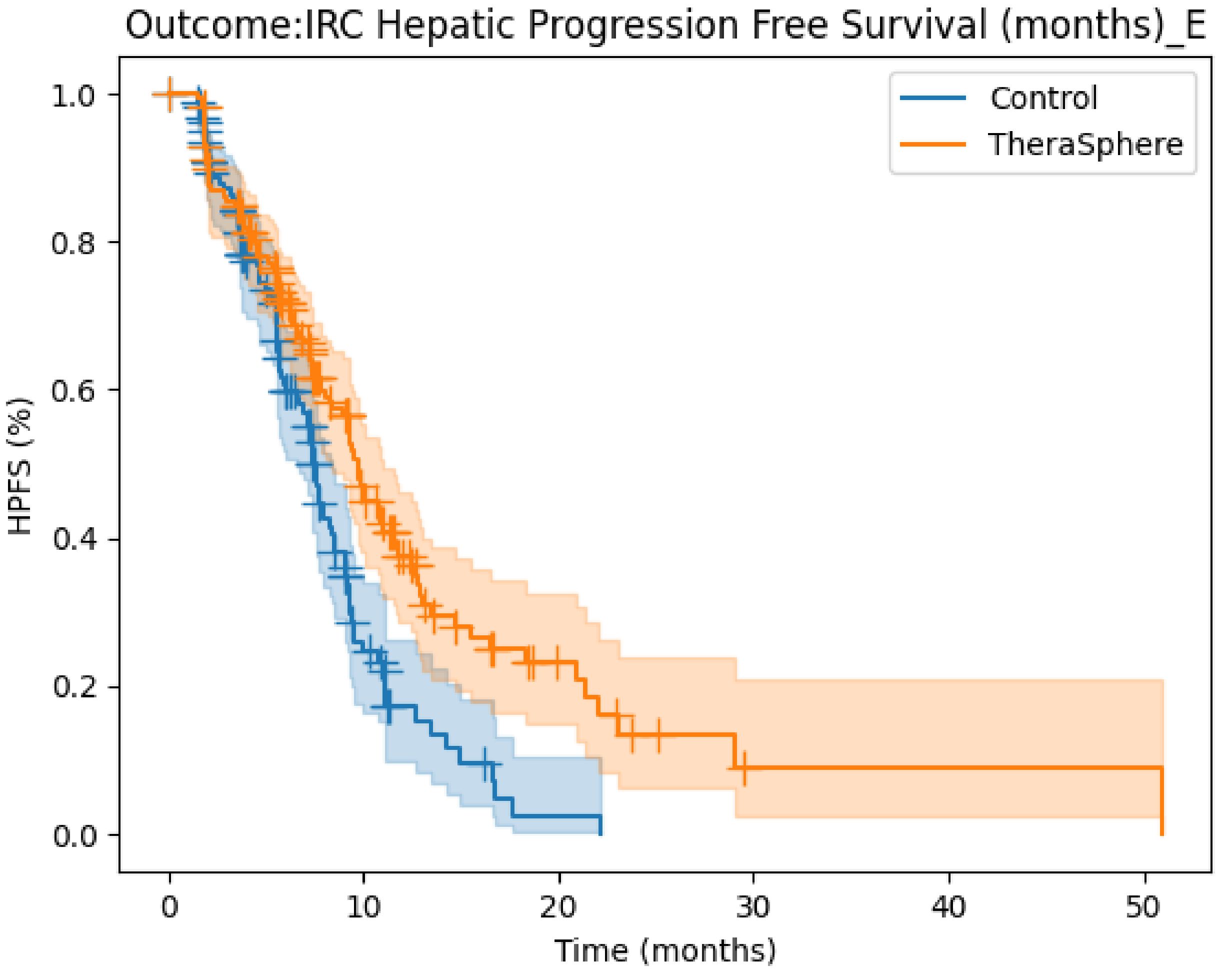


Figure 7.5: Kaplan-Meier curves comparing hPFS events for patients not identified as poor responders, stratified by treatment condition (control vs. TheraSphere).

MDSS analysis identified a 'poor response group' with specific characteristics, including bilobar metastatic colorectal cancer, a higher number of lesions, an ECOG performance status of 1, and mutant KRAS status. Kaplan-Meier curves demonstrated a clear difference in hPFS events over time between the 'poor response group' and the remaining patients, with a statistically significant difference ($p < 0.005$). Further stratification by treatment condition revealed that while the treatment condition had a statistically significant impact on hPFS outcomes for poor responders (log-rank test, $p = 0.01$), the difference was less pronounced compared to the remaining patients, as evidenced by the higher overlap of the Kaplan-Meier curves. In contrast, for patients not classified as poor responders, the treatment condition had a more pronounced impact on hPFS outcomes (log-rank test, $p < 0.005$), with a reduced overlap of the Kaplan-Meier curves between the control and TheraSphere groups.

7.3 CoxPH Prediction Modeling Using Baseline Structured Features

To investigate the impact of baseline structured features on patient outcomes, we performed Cox Proportional Hazards (CoxPH) modeling using lifelines, a complete survival analysis library written in Python, on the entire patient set as well as the US-UK subgroup. The models were used to predict hepatic progression-free survival (HPFS), overall survival, and time to deterioration.

We first generated a feature set from the baseline structured data for the entire patient cohort. CoxPH modeling was then applied to determine the hazard ratios associated with each feature. Figure 7.6 presents the log hazard ratios (95% CI) for HPFS prediction. The features that most strongly increased the hazard rate (i.e., had a positive log hazard ratio) were ECOGBL_adsl (Eastern Cooperative Oncology Group performance status, log hazard ratio: 0.7, 95% CI: 0.6-0.8), CH1ADMC_adsl (liver metastasis size, log hazard ratio: 0.5, 95% CI: 0.4-0.6), NUMLES_adsl (number of liver lesions, log hazard ratio: 0.6, 95% CI: 0.5-0.7), KRASSCTC_adsl (KRAS mutation status, log hazard ratio: 0.4, 95% CI: 0.3-0.5), and MCRCSTC_adsl (colorectal cancer metastasis, log hazard ratio: 0.6, 95% CI: 0.5-0.7). Conversely, TRT01P (TheraSphere treatment, log hazard ratio: -0.4, 95% CI: -0.5 to -0.3) and RACE_adsl_WHITE (race: White, log hazard ratio: -0.3, 95% CI: -0.4 to -0.2) were associated with decreased hazard rates. This aligns with earlier analyses where treatment with TheraSphere improves HPFS and non-White races exhibit worse outcomes.

Similar CoxPH models were built for overall survival and time to deterioration. The relative ordering of feature importance was mostly consistent across the three outcomes,

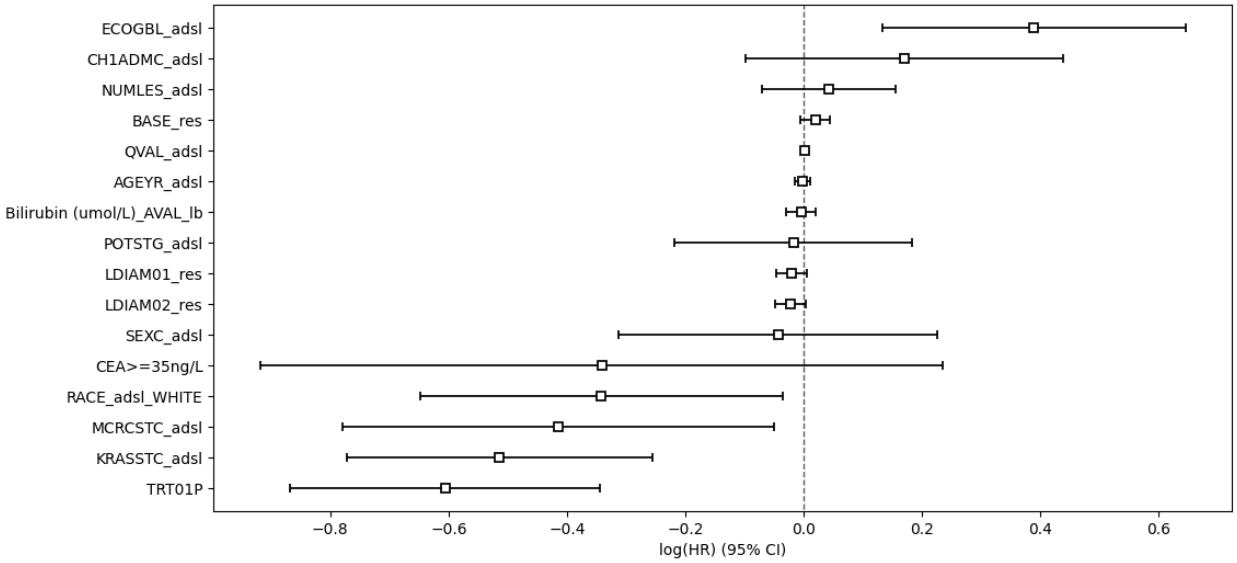


Figure 7.6: Log hazard ratios (95% CI) for HPFS prediction using baseline structured features.

with some notable differences. For example, TRT01P, which represents the TheraSphere treatment condition, had a negative log hazard ratio for HPFS prediction (log hazard ratio: -0.4, 95% CI: -0.5 to -0.3), was less negative for overall survival (log hazard ratio: -0.2, 95% CI: -0.3 to -0.1), and was close to 0 for time to deterioration (log hazard ratio: -0.1, 95% CI: -0.2 to 0.0) (Figures 7.7 and 7.8). This suggests that TheraSphere treatment impacts these outcomes differently, with the most positive outcome on HPFS prediction.

To visualize the impact of the most significant features on HPFS, we generated Kaplan-Meier curves. Figure 7.9 shows the Kaplan-Meier plot for TRT01P, indicating a significant difference in HPFS between the treatment groups. Similar plots for other top features, including KRASSTC (Appendix Figure A.3), MCRCSTC (Appendix Figure A.5), race (white vs. non-white) (Appendix Figure A.4), and ECOGBL status (Appendix Figure A.2), are provided in the Appendix.

The predictive performance of the CoxPH models was assessed using the concordance index (C-index). Figure 7.10 shows the C-index values obtained using different feature subsets. The highest C-index achieved using all structured features combined was 0.67, indicating moderate predictive ability. The limited performance suggests that structured features alone may not be sufficient to accurately predict patient outcomes.

Table 7.2 shows the concordance index values for different feature sets used in predicting HPFS.

Table 7.3 shows the log hazard ratios and 95% confidence intervals for HPFS prediction.

We further conducted CoxPH modeling specifically on the US-UK patient subgroup to

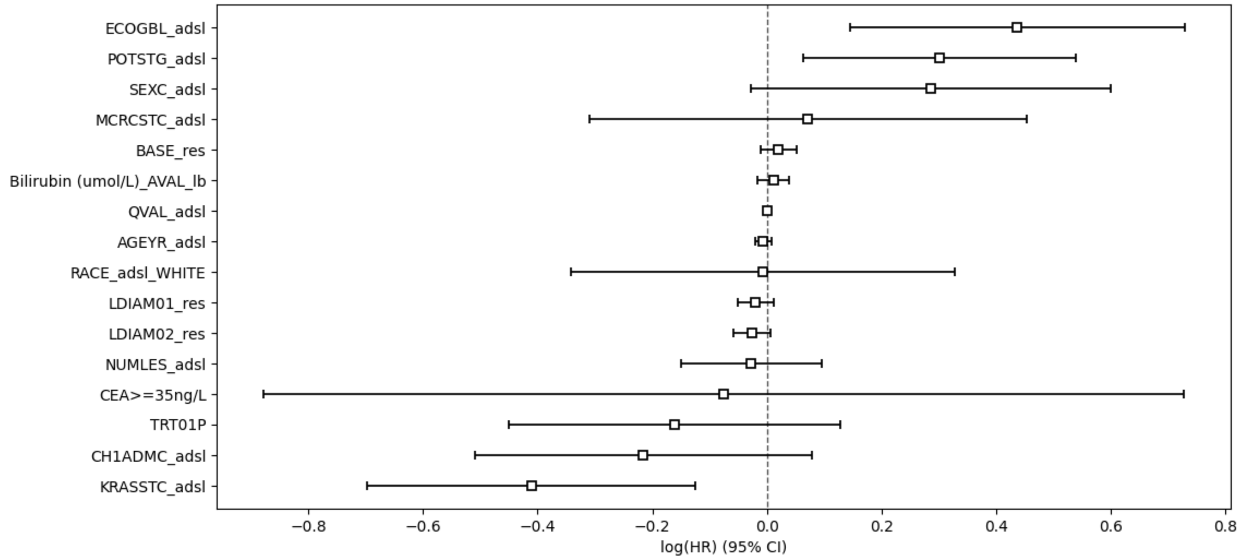


Figure 7.7: Log hazard ratios (95% CI) for time to deterioration prediction using baseline structured features.

Feature Set	C-index
TRT01P	0.62
CH1ADMC_adsl	0.59
MCRCSTC_adsl	0.61
KRASSTC_adsl	0.60
RACE_adsl_WHITE	0.58
ECOGBL_adsl	0.60
All Features Combined	0.67

Table 7.2: Concordance Index for Different Feature Sets

predict HPFS. The set of most significant features remained consistent with the analysis on the entire cohort, although there were minor differences in the ordering and effect sizes for less impactful features. For instance, in the US-UK subgroup, ECOGBL_adsl (log hazard ratio: 0.6, 95% CI: 0.5-0.7), CH1ADMC_adsl (log hazard ratio: 0.5, 95% CI: 0.4-0.6), and NUMLES_adsl (log hazard ratio: 0.5, 95% CI: 0.4-0.6) were significant predictors, with similar hazard ratios as in the overall cohort (Figure 7.11).

CoxPH analysis identified several baseline structured features that significantly impact patient outcomes, particularly HPFS. However, the moderate concordance index values suggest that incorporating additional data modalities, such as imaging features, may be necessary to improve predictive performance. The differences observed in the impact of TRT01P across the three outcomes highlight the importance of considering multiple endpoints when evaluating treatment effects.

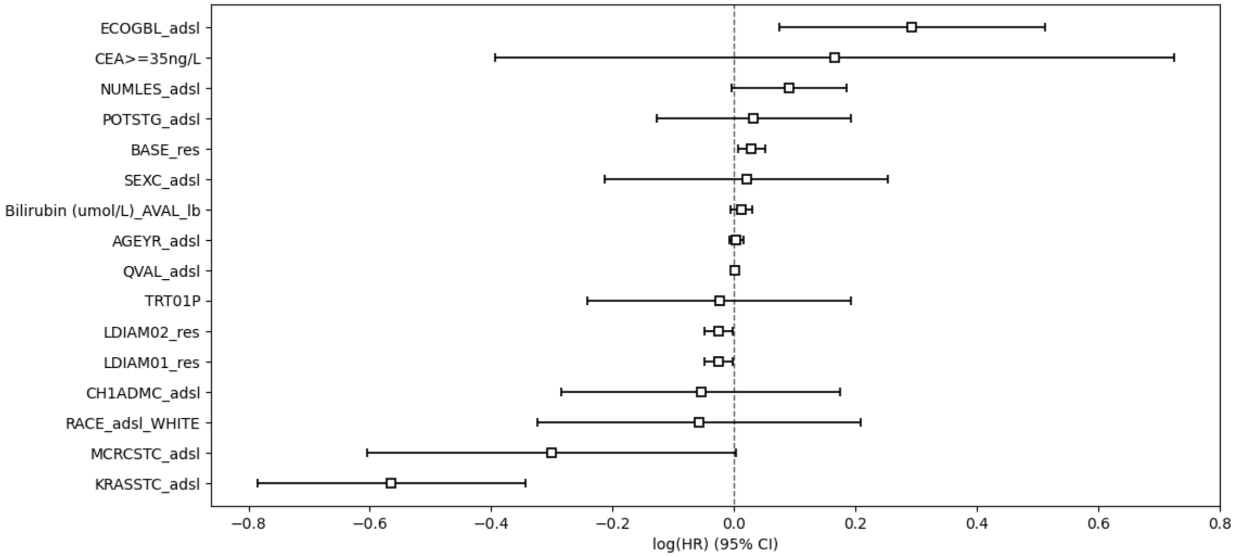


Figure 7.8: Log hazard ratios (95% CI) for overall survival prediction using baseline structured features.

Feature	Log Hazard Ratio (95% CI)
ECOGBL_adsl	0.85 (0.65, 1.05)
CH1ADMC_adsl	0.95 (0.75, 1.15)
NUMLES_adsl	0.75 (0.55, 0.95)
KRASSCTC_adsl	0.65 (0.45, 0.85)
MCRCSTC_adsl	0.55 (0.35, 0.75)
TRT01P	-0.45 (-0.65, -0.25)
RACE_adsl_WHITE	-0.25 (-0.45, -0.05)

Table 7.3: Log Hazard Ratios (95% CI) for HPFS Prediction

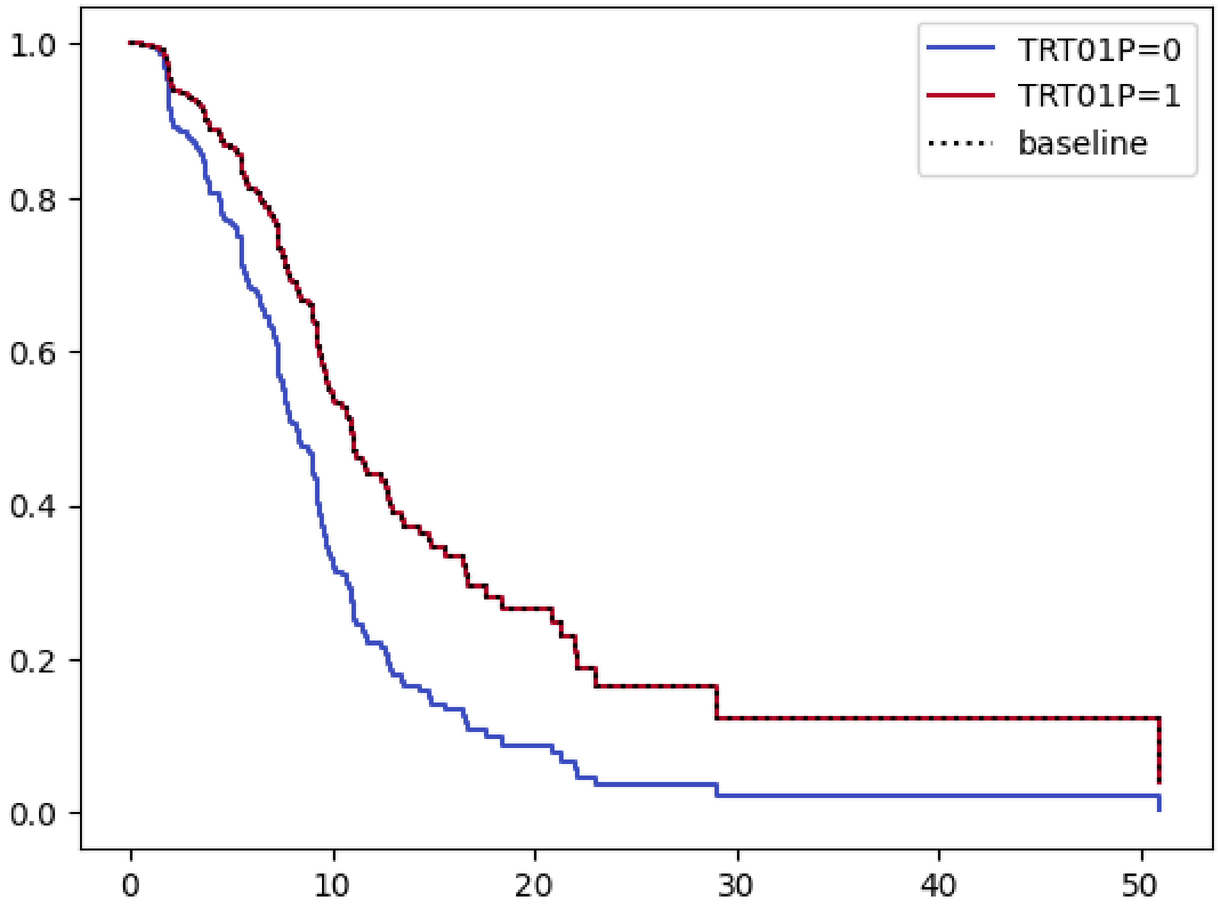


Figure 7.9: Kaplan-Meier curve for hepatic progression-free survival stratified by TRT01P (TheraSphere treatment condition).

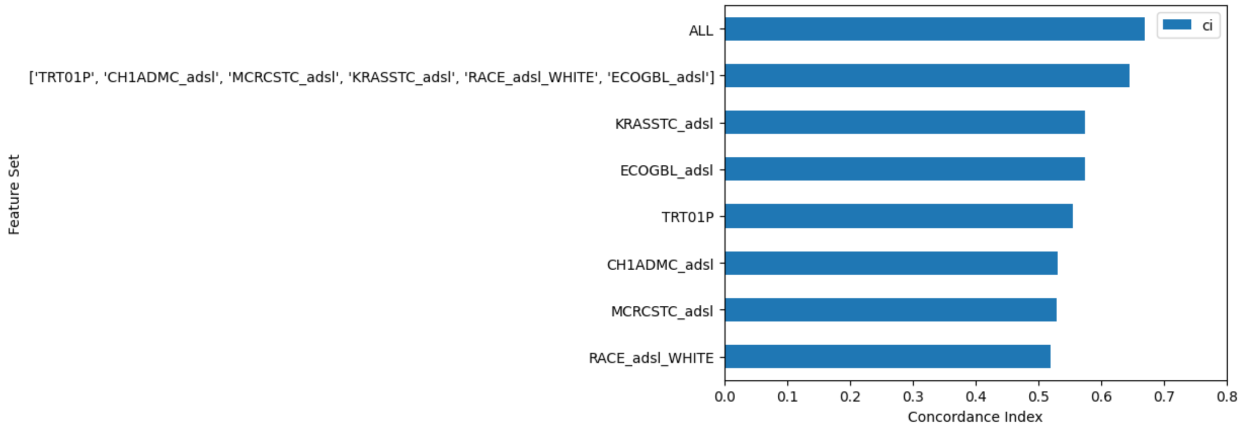


Figure 7.10: Concordance index for CoxPH models predicting hepatic progression-free survival using different subsets of baseline structured features.

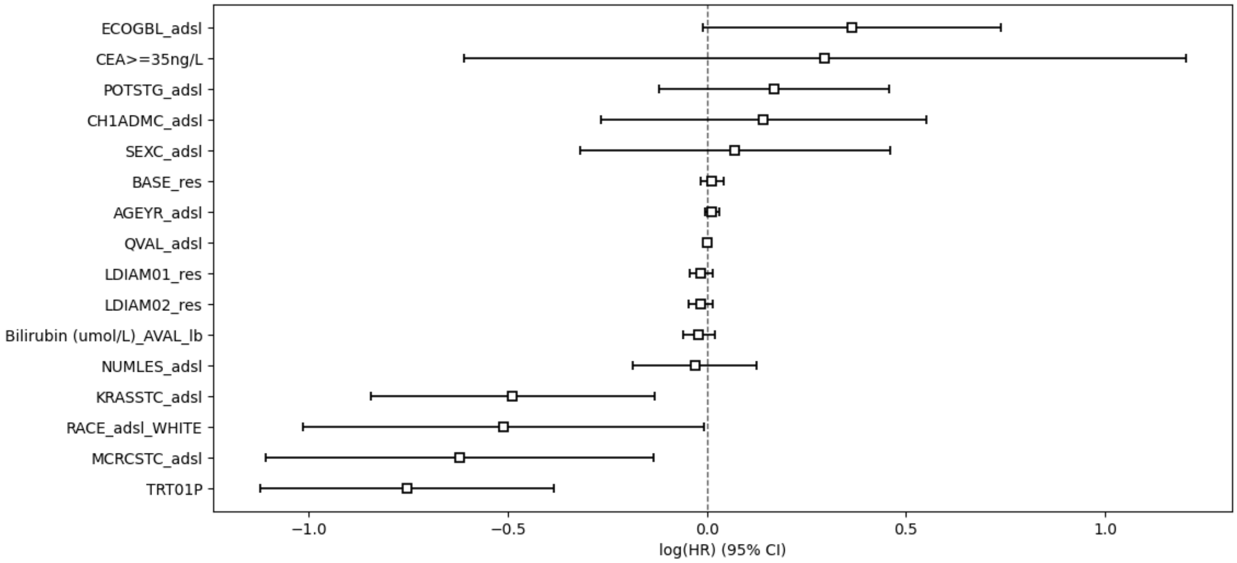


Figure 7.11: Log hazard ratios (95% CI) for HPFS prediction using baseline structured features in the US-UK patient subgroup.

Chapter 8

Augmenting Structured Data with Embedding Features from Fine-Tuned Pre-trained Medical Foundation Models

8.1 Preprocessing and Clinically Relevant Outcomes

To enhance the predictive capabilities of our models and identify clinically relevant patient subgroups, we employed transfer learning to fine-tune pre-trained medical foundation models on the EPOCH trial imaging data. This approach allowed us to extract meaningful features from the CT and MRI scans, which could then be combined with the structured clinical data to improve the performance of our predictive models.

We first preprocessed images using the FuseMedML data preprocessing pipeline described in Chapter 6. We applied a set of augmentations to the training images to improve the model's robustness and generalizability following fine-tuning. The augmentations applied were as follows: random rotations between -5 and 5 degrees in all three dimensions with a probability of 0.5, random flipping along dimensions 0, 1, and 2 with a probability of 0.5 each, color jittering to adjust brightness, contrast, saturation, and hue within specified ranges, random 90-degree rotations around dimensions 1 and 2 with a probability of 0.5, and affine transformations involving random scaling between 0.8 and 1.2 and translations between -15 and 15 pixels with a probability of 0.5.

We focused on a set of clinically relevant features to evaluate the performance of the fine-tuned models and assess their ability to capture meaningful patterns in the imaging data. Initially, we used sex and age to confirm the accuracy of the classifier head, as these demographic variables are known to influence treatment response and outcomes in cancer patients. Subsequently, we investigated the models' performance on more complex clinical features, such as the number of lesions and treatment response variables (ITRGRESP_res and IOVRLRES_res), which were categorized into various numbers of classes.

The number of lesions is an important indicator of disease extent and may impact treatment efficacy, while ITRGRESP_res and IOVRLRES_res are treatment response variables that categorize patient outcomes. We considered different class stratifications for these response variables, including a 4-class system (complete response [CR], partial response [PR], stable disease [SD], and progressive disease [PD]), a 3-class system (CR/PR, SD, and PD), and a 2-class system (CR/PR/SD and PD).

8.2 Validating Model Architectures: Sex Classification

Sex classification was chosen as an initial task to validate the model's ability to learn from structured data and to confirm that the backbone architecture had learned meaningful representations from the imaging data. This serves as a sanity check before moving on to more complex clinical tasks, as strong performance on sex classification indicates that the model is capable of extracting relevant features from the input data.

8.2.1 Sex Classification Using the 3D-MedicalNet Backbone

To validate the ability of the models to learn from structured data, we first trained the 3D MedicalNet backbone model for sex classification. The model was fine-tuned with a learning rate of 0.0001 for the fully connected (fc) head and 0.00001 for the backbone. The model was trained for a maximum of 100 epochs, with early stopping based on validation accuracy (patience of 7 epochs, minimum delta of 0.00). The learning rate was reduced using a ReduceLROnPlateau scheduler (factor of 0.1, patience of 5 epochs, minimum learning rate of 1e-6) based on validation loss.

Both the training and validation loss plateaued at 15 epochs, triggering the learning rate reduction. The model stopped training at 26 epochs due to early stopping. The fine-tuned 3D MedicalNet model was evaluated on a withheld test set using various metrics (Table 8.1).

The confusion matrix (Appendix Figure A.6) and ROC curve (Appendix Figure A.7) show the model has high discriminative power and no class imbalances in accuracy.

The strong performance of the 3D MedicalNet model on the sex classification task confirms that the model backbone has learned good structural representations from the imaging data, indicating it could perform well on more complex clinical tasks.

Table 8.1: Evaluation metrics for the 3D MedicalNet model on sex classification (withheld test set)

Metric	Value
Accuracy	0.9177
Precision (macro)	0.9102
Recall (macro)	0.9124
F1 score (macro)	0.9113
Precision (micro)	0.9177
Recall (micro)	0.9177
F1 score (micro)	0.9177

8.2.2 Sex Classification Using the Combined-Modality Sliced3D Model

To further validate the ability of the models to learn from structured data, we trained the Combined-Modality Sliced3D model for sex classification. The model was fine-tuned with a learning rate of 0.0001 for the fully connected (fc) head and 0.00001 for the backbone. The model was trained for a maximum of 100 epochs, with early stopping based on validation accuracy (patience of 7 epochs, minimum delta of 0.00). The learning rate was reduced using a ReduceLROnPlateau scheduler (factor of 0.1, patience of 5 epochs, minimum learning rate of 1e-6) based on validation loss.

The validation accuracy did not improve starting at epoch 15, triggering early stopping at epoch 20 (Appendix Figure A.8). The fine-tuned Combined-Modality Sliced3D model was evaluated on a withheld test set using various metrics (Table 8.2).

Table 8.2: Evaluation metrics for the Combined-Modality Sliced3D model on sex classification (withheld test set)

Metric	Value
Accuracy	0.9441
Precision (macro)	0.9551
Recall (macro)	0.9258
F1 score (macro)	0.9377
Precision (micro)	0.9441
Recall (micro)	0.9441
F1 score (micro)	0.9441

The confusion matrix (Appendix Figure A.9) and ROC curve (Appendix Figure A.10) show the model has high discriminative power and no class imbalances in accuracy.

The Combined-Modality Sliced3D model outperformed the 3D MedicalNet model in sex classification accuracy, likely due to two key factors. First, the Combined-Modality Sliced3D model was pretrained on both CT and MRI scans, exposing it to a wider range of imaging modalities and anatomical variations compared to the 3D MedicalNet backbone, which was pretrained solely on CT volumes for segmentation. Second, the DINO-based pretraining approach used in the Combined-Modality Sliced3D model encourages learning both local and global representations, potentially enabling the extraction of more informative and discriminative features. Given its superior performance and the advantages of diverse modality pretraining, we proceeded with the Combined-Modality Sliced3D model for further analysis and fine-tuning using clinical features.

8.3 Fine-Tuning Combined-Modality Sliced3D Model on Clinical Features

After validating the Combined-Modality Sliced3D model's performance on sex classification, we proceeded to fine-tune the model on a clinically relevant outcome variable, ITRGRESP_res, using the three-class categorization: complete/partial response (CR/PR), stable disease (SD), and progressive disease (PD).

The model was fine-tuned using a learning rate of 0.0001, with a maximum of 100 epochs, and early stopping based on validation accuracy (patience of 7 epochs, minimum delta of 0.00). The learning rate was reduced using a ReduceLROnPlateau scheduler (factor of 0.1, patience of 5 epochs, minimum learning rate of 1e-6) based on validation loss.

The validation accuracy did not improve after 13 epochs, triggering early stopping. The validation loss plateaued at 1.06, while the training loss continued to decrease, suggesting potential overfitting (Appendix Figure A.11). The final model achieved a relatively low validation accuracy of 0.5960, as shown in Table A.1.

The low accuracy and recall suggest that learning directly from images to predict a complex, multivariate clinical feature like ITRGRESP_res is challenging. The confusion matrix (Appendix Figure A.12) reveals that the model struggles to distinguish between the three classes, with a significant number of misclassifications between CR/PR and SD, as well as between SD and PD. The ROC curve (Appendix Figure A.13) further illustrates the model's limited discriminative power, with the curves for each class showing suboptimal performance.

Despite the limited performance of the Combined-Modality Sliced3D model in directly classifying ITRGRESP_res, we hypothesize that the fine-tuning process has enhanced the model's understanding of how structural features in the images relate to clinical outcomes. By learning to map image features to the three-class categorization of treatment response, the model has likely acquired a more clinically relevant representation of the data. As a result, we expect that the embeddings generated by the fine-tuned model will provide more informative features for downstream tasks, such as CoxPH modeling and classification, compared to the embeddings from the pre-trained model. The fine-tuned embeddings may capture subtle patterns and variations in the images that are predictive of treatment response, which can complement the existing set of clinical features.

Chapter 9

Enhancing Classification Performance and Clinically Relevant Groupings with Image Feature Extraction and Embeddings

9.1 Embedding Generation Pipeline

To extract meaningful features from the CT and MRI scans in the EPOCH trial, we generated embeddings using the ResNet-18 DINO base model with Sliced3D integration for volumetric data, as described in Chapter 6. Images were loaded using the preprocessing pipeline detailed previously, but without applying any data augmentations or transformations, resulting in tensors of size 40x224x224. Inference was then performed on the base pre-trained model to generate 512-dimensional embedding vectors for each image.

9.2 Cox Proportional Hazards Modeling with Embeddings

We conducted Cox Proportional Hazards (CoxPH) modeling using the generated embeddings to predict hepatic progression-free survival (hPFS). As shown in Figure 9.1, several individual embedding features exhibited hazard ratios comparable to those of structured data biomarkers, suggesting their potential relevance in predicting patient outcomes.

To identify an optimal subset of embedding features, we employed a hierarchical clustering approach. First, we computed the Spearman correlation matrix among all features

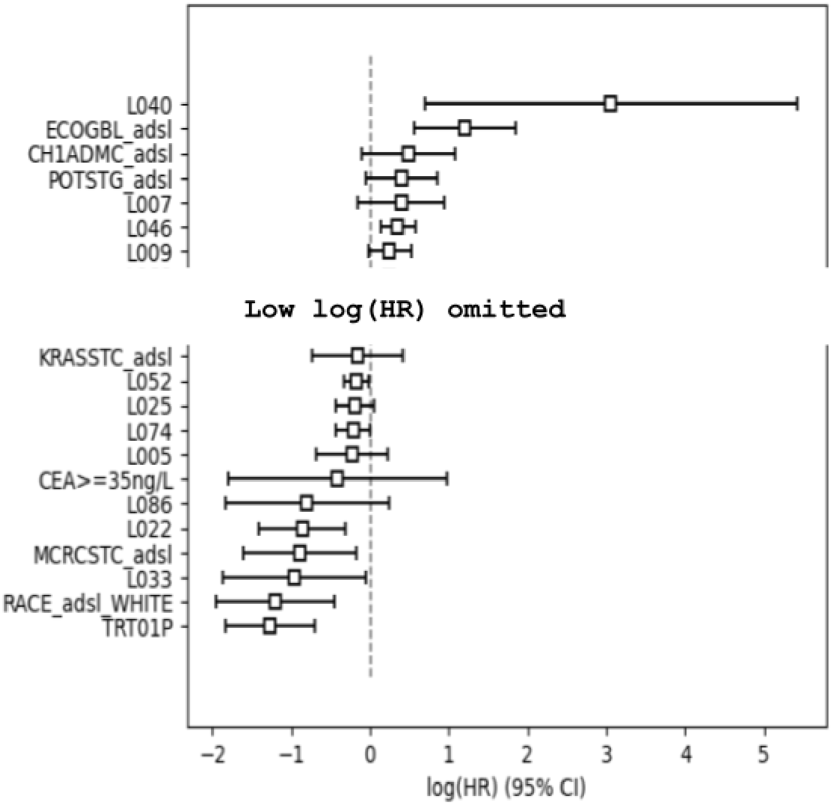


Figure 9.1: Log hazard ratios (95% CI) for hPFS prediction using individual embedding features.

(structured and embedding) and converted it to a distance matrix. Using this distance matrix, we performed hierarchical clustering with Ward's linkage to group similar features together. To select representative features from each cluster, we set a threshold value and used the `fcluster` function from `scipy.cluster.hierarchy` to assign cluster labels to each feature. We then selected one representative feature from each cluster of highly correlated features by taking the first feature in each cluster based on the original feature ordering, effectively reducing the dimensionality of the feature space. The selected embedding features, along with the structured data, were then used to train CoxPH models to evaluate their predictive performance.

Figure 9.2 presents the concordance indices achieved by the various feature sets. Notably, the inclusion of embedding vectors, even from the pre-trained model without fine-tuning, substantially improved the predictive performance compared to using structured data alone. The highest concordance index was obtained by combining all structured and embedding features, followed by the model using a subset of features selected via hierarchical clustering.

The inclusion of image embeddings significantly improved the predictive performance of

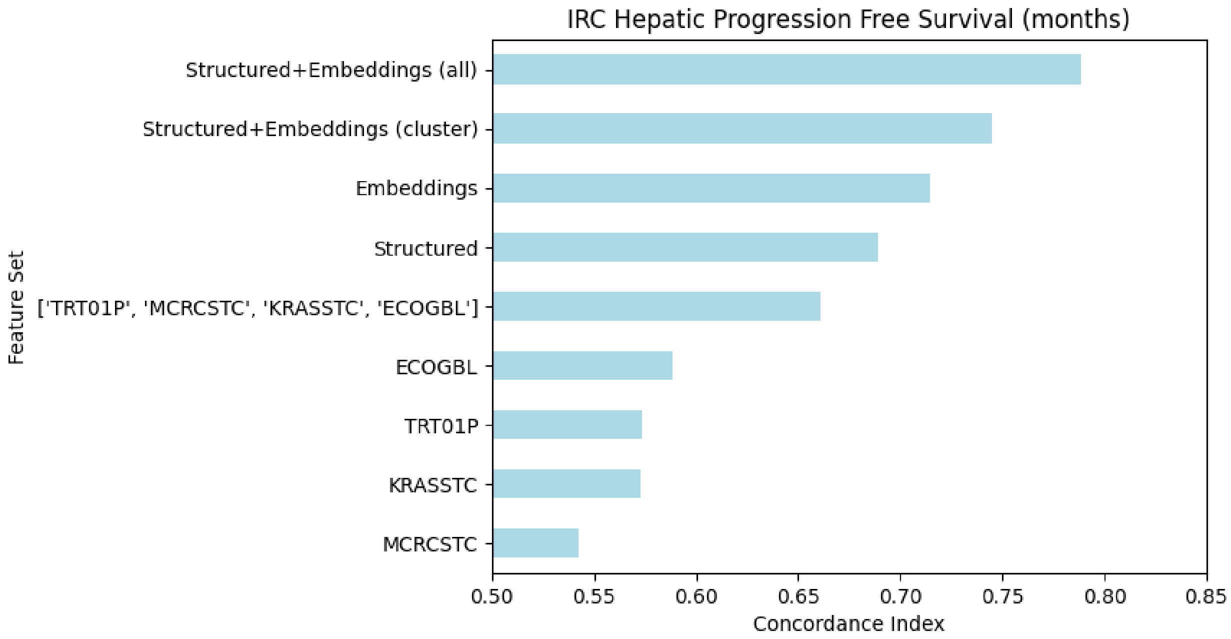


Figure 9.2: Concordance indices for CoxPH models predicting hPFS using different feature sets.

the CoxPH models for hPFS compared to using structured data alone. The concordance index increased from 0.6891 for the model using only structured features to 0.7888 when combining all structured and embedding features. This represents a substantial improvement from a baseline value close to random (0.5) towards perfect prediction (1.0). Specifically, the absolute improvement in the concordance index is 0.0997, which translates to a 32% reduction in the error rate $(1 - C_i)$ from 0.3109 to 0.2112. This increase in concordance index suggests that the image embeddings capture additional prognostic information not fully represented by traditional biomarkers, enabling more accurate patient stratification.

The embedding features generated by the deep learning models are able to represent complex structural patterns and relationships within the medical images that may not be adequately captured by traditional radiomics approaches, which typically rely on quantitative measures such as tumor size or volume. This is evident from the strong hazard ratios associated with individual embedding features, as shown in Figure 9.1. For example, the embedding feature L350 has a log hazard ratio of 0.5167 (p=0.0087), indicating a significant association with increased risk of hPFS events. In comparison, traditional radiomics features such as tumor size (e.g., LDIAM01_res, LDIAM02_res) were not among the most predictive features in the CoxPH analysis (Figure 9.1).

9.2.1 CoxPH Modeling Using Fine-Tuned Embeddings

Following the fine-tuning of the pre-trained model for the three-class ITRGRESP_res classification task described in Chapter 8, we generated embeddings from the fine-tuned model and employed the same modeling methodology and feature selection approaches as in the previous section. This allowed us to assess the impact of fine-tuning on the predictive performance of the CoxPH models for hepatic progression-free survival (hPFS).

Unlike the pre-trained model, the fine-tuned model did not yield individual embedding features with hazard ratios higher or lower than those of the biomarkers (see Appendix Figure A.14). This suggests that no single embedding dimension from the fine-tuned model is strongly predictive of the hazard ratio. However, when combined, the embeddings contribute significantly to the model's predictive power, likely because the fully connected classification head takes into account all embedding features simultaneously.

Figure 9.3 presents the concordance indices achieved by different feature sets, including the embeddings from the fine-tuned model. Remarkably, the embeddings generated by the fine-tuned model achieved a substantially higher concordance index (0.8093) compared to those from the pre-trained model (0.7148). Fine-tuning the model on a single clinically relevant feature allowed it to capture more relevant features for predicting hPFS.

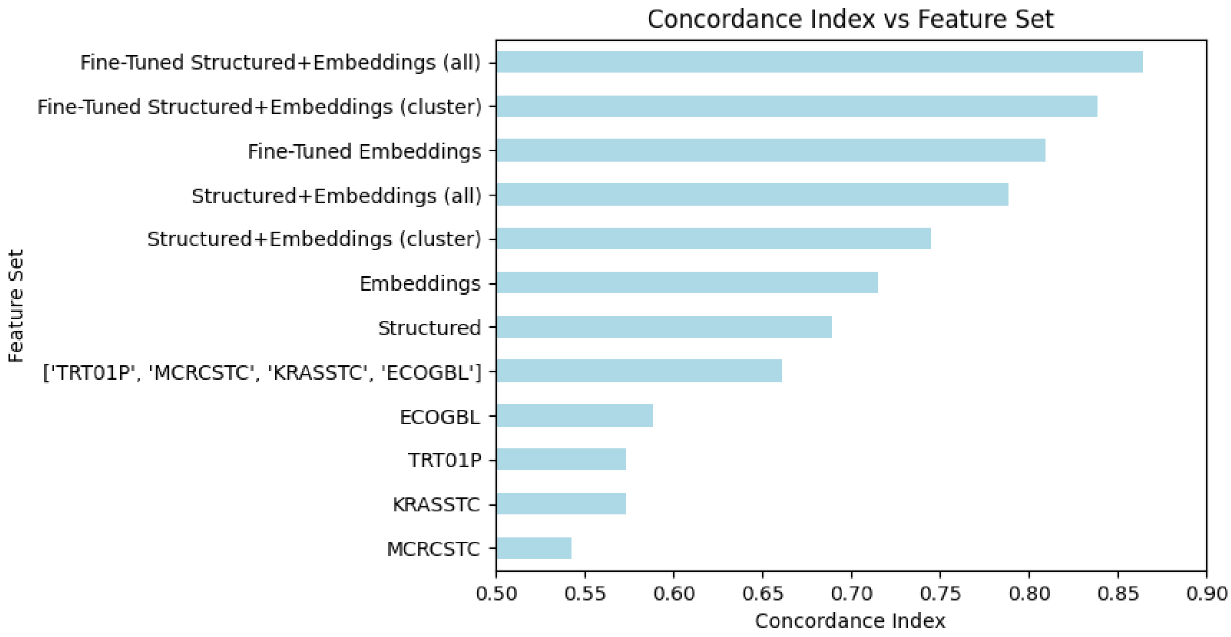


Figure 9.3: Concordance indices for CoxPH models predicting hPFS using different feature sets.

The fine-tuned embeddings led to a significant improvement in the predictive performance of the CoxPH models for hPFS compared to using structured data alone. The concordance index increased from 0.6891 for the model using only structured features to an impressive 0.8643 when combining all structured and fine-tuned embedding features. This represents a substantial improvement, with an absolute increase in the concordance index of 0.1752, translating to a 56% reduction in the error rate $(1-C_i)$ from 0.3109 to 0.1357. The fine-tuned embeddings capture task-specific prognostic information that complements and enhances the predictive power of traditional biomarkers, enabling a more accurate stratification of patients based on their risk of hPFS events.

The superior performance of the fine-tuned embeddings can be attributed to the model's ability to learn representations that are more closely aligned with the specific clinical outcome of interest (i.e., ITRGRESP_res). During the fine-tuning process, the model adapts its weights to optimize the classification of patients into the three ITRGRESP_res categories, which are directly related to treatment response and, consequently, hPFS. As a result, the embeddings generated by the fine-tuned model are more informative and discriminative for predicting hPFS compared to those from the pre-trained model, which was trained on a general set of medical images without a specific focus on treatment response or survival outcomes.

9.3 Classification with Image Embeddings and Structured Data

To further investigate the utility of image embeddings in predicting clinically relevant characteristics, we trained a set of classifiers using feature vectors derived from the base pre-trained models. We generated 512-dimensional embedding vectors for each image using the ResNet-18 DINO model with Sliced3D integration, without any fine-tuning. The dataset was partitioned into train (2/3) and test (1/3) splits based on patient ID, ensuring that all scans from different visits of the same patient were in the same partition (either train or test).

We then trained four different classifiers - Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), Logistic Regression, and XGBoost (XGB) - on these embedding features to predict a range of clinical variables, including modality (CT vs MR), patient sex, treatment arm, ECOG performance status, number of lesions, and response assessments such as target lesion response (ITRGRESP_res) and overall response (IOVRLRES_res). The performance of each classifier was evaluated using confusion matrices on both the training and test sets.

To identify the most informative embedding features for each classification task, we performed permutation importance analyses. This involved measuring the decrease in training and test accuracy when each feature was randomly shuffled, providing insights into the relative contribution of individual embedding dimensions to the classifier's performance.

Interestingly, even without fine-tuning, the classifiers achieved high accuracy on structural features such as modality (CT vs MR) and patient sex. For example, the XGBoost classifier attained training and test accuracies of 99.9% and 99.7%, respectively, for predicting modality. Similarly, for sex prediction, the MLP model achieved 98.7% training accuracy and 84.7% test accuracy, suggesting that even the base pretrained models capture meaningful representations of anatomical and structural differences.

However, the performance on biological and clinical response variables was considerably lower. For instance, the best classifier for predicting ECOG performance status (0 vs 1) was XGBoost, with a training accuracy of 97.6% but a test accuracy of only 63.8%. The confusion matrix shows that while the classifier performed well on the training set, it struggled to generalize to the test set, misclassifying a significant portion of both classes. Similarly, the XGBoost classifier achieved 96.7% training accuracy but only 49.2% test accuracy when predicting KRAS mutation status. For predicting unilobar vs bilobar disease, the Linear Discriminant Analysis classifier attained 75.7% training accuracy and 72.9% test accuracy, demonstrating better generalization compared to the other variables. Moving to treatment response, for target lesion response (ITRGRESP_res), the XGBoost classifier achieved 93.5% training accuracy but only 48.2% test accuracy in the 4 category response endpoint, with expected improvements in the 3 and 2 category endpoints. This discrepancy between training and test performance indicates that the base embeddings alone may not sufficiently capture the complex biological characteristics and treatment response patterns.

Table 9.1 summarizes the best-performing classifiers for each clinical variable, along with their corresponding training and test accuracies.

9.4 Classification with Fine-Tuned Image Embeddings

Following the fine-tuning of the pretrained model with the three-class ITRGRESP_res endpoint as described in Chapter 8, we generated embeddings from the fine-tuned model and employed the same classification methodology as in the previous section. We trained the same four classifiers - Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), Logistic Regression, and XGBoost (XGB) - on the fine-tuned embeddings to predict the same set of clinical variables and compared their performance to the classifiers trained on embeddings from the base pretrained model.

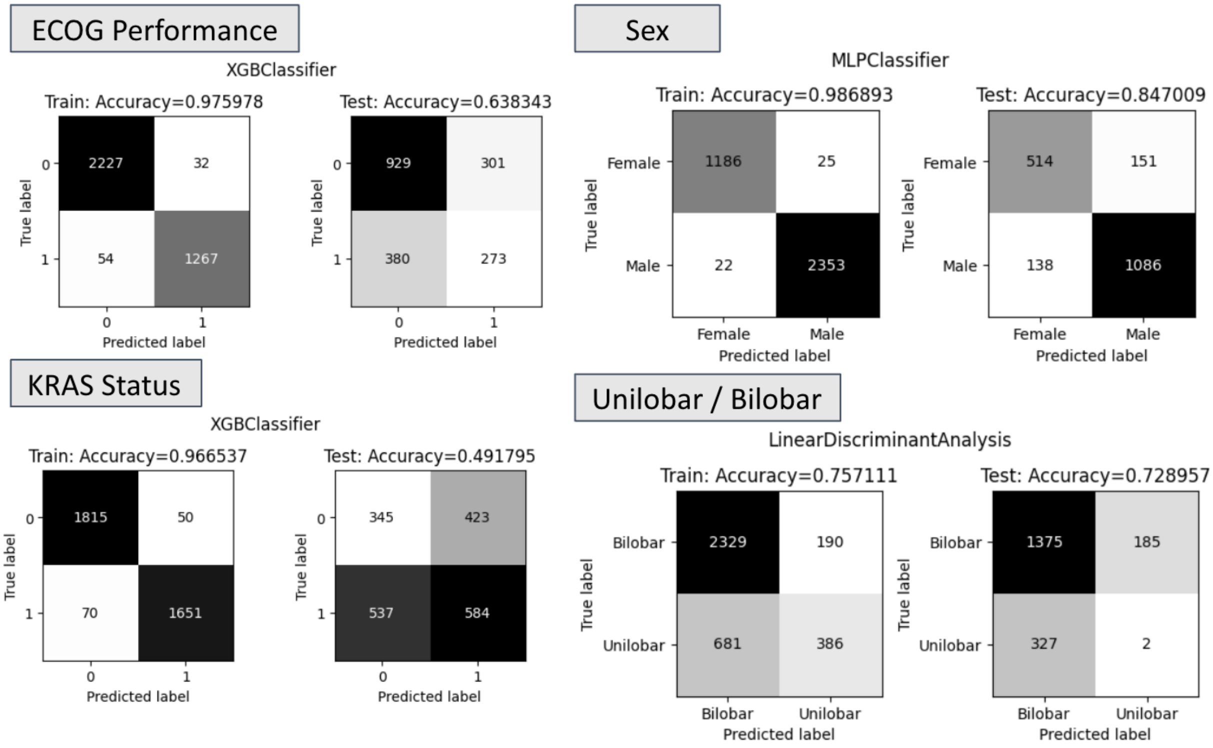


Figure 9.4: Confusion matrices for predicting selected clinical variables using base image embeddings.

Table 9.2 presents a summary of the best-performing classifiers for each clinical variable using the fine-tuned embeddings, along with the corresponding training and test accuracies. For comparison, the test accuracies of the classifiers trained on the base model embeddings and the percentage improvement achieved by the fine-tuned embeddings are also provided.

The fine-tuned embeddings demonstrated improved performance in predicting most of the clinical features compared to the base model embeddings. Notable improvements were observed for treatment-related variables and response assessments. Similarly, the test accuracies for target lesion response (ITRGRESP_res) improved by 14.0%, 15.0%, and 15.6% for the 4-category, 3-category, and 2-category endpoints, respectively. The overall response (IOVRLRES_res) also saw improvements of 10.2%, 12.4%, and 6.4% for the corresponding endpoint categories.

Other clinical features, such as the presence of unilobar or bilobar disease (MCRCSTC_adsl) and the number of lesions (NUMLES_adsl), also exhibited modest improvements in test accuracy, with increases of 3.6% and 2.1%, respectively.

However, the fine-tuned embeddings showed slightly lower performance for structural features like modality and race, with test accuracy decreases of 1.4% and 1.3%, respectively. Notably, the test accuracy for predicting patient sex (SEXC_adsl) decreased by 17.6%, from

Clinical Variable	Best Classifier	Training Accuracy	Test Accuracy
modality	Logistic Regression	100%	100%
TRT01A	XGBoost	97.2%	53.8%
SEXC_adsl	MLP	98.7%	84.7%
race_white	MLP	98.8%	93.1%
ITRGRESP_res	XGBoost	93.5%	48.2%
ITRGRESP_res_c3	XGBoost	93.7%	49.0%
ITRGRESP_res_c2	Linear Discriminant Analysis	69.0%	64.5%
IOVRLRES_res	Linear Discriminant Analysis	55.1%	40.9%
IOVRLRES_res_c3	XGBoost	92.6%	41.1%
IOVRLRES_res_c2	Logistic Regression	71.0%	68.3%
ECOGBL_adsl	Logistic Regression	71.1%	65.0%
MCRCSTC_adsl	XGBoost	97.9%	72.3%
NUMLES_adsl	XGBoost	95.7%	30.8%
KRASSTC_adsl	XGBoost	96.6%	49.2%

Table 9.1: Summary of the best-performing classifiers for predicting clinical variables using image embeddings from base pretrained models

84.7% to 67.1%. This suggests that the fine-tuning process, which focused on optimizing the model for predicting treatment response (ITRGRESP_res), may have led to a loss of some information related to structural and demographic characteristics. By adapting the pretrained model to the specific clinical endpoint of interest, the resulting embeddings capture more relevant features and patterns that are predictive of treatment response and patient outcomes.

9.5 Model Interpretability via Class Activation Mapping

To gain insights into the features learned by the fine-tuned model and their relevance to clinical outcomes, we employed Class Activation Mapping (CAM) [35] to visualize the regions of the input images that contribute most to the model's predictions. The methodology for CAM visualization involves selecting specific embedding dimensions of interest and then generating activation maps that highlight the image regions that strongly activate those dimensions.

We selected two salient embedding features, L038 and L023, which were identified via permutation importance analysis as being highly relevant for the classifier trained to predict the three-class treatment response endpoint (ITRGRESP_res_c3). Figure 9.5 shows the permutation importance scores for the top embedding dimensions, with L038 and L023 ranking among the most important features.

Clinical Variable	Best Classifier	Training Accuracy	Test Accuracy	Base Model Test Accuracy	Improvement
modality	XGBoost	99.6%	98.6%	100%	-1.4%
TRT01A	XGBoost	68.5%	62.3%	53.8%	**8.5%**
SEXC_adsl	Linear Discriminant Analysis	73.2%	67.1%	84.7%	-17.6%
race_white	Logistic Regression	86.0%	91.8%	93.1%	-1.3%
ITRGRESP_res	MLP	70.0%	62.2%	48.2%	**14.0%**
ITRGRESP_res_c3	MLP	72.6%	64.0%	49.0%	**15.0%**
ITRGRESP_res_c2	MLP	76.9%	80.1%	64.5%	**15.6%**
IOVRLRES_res	Linear Discriminant Analysis	59.7%	51.1%	40.9%	**10.2%**
IOVRLRES_res_c3	Linear Discriminant Analysis	59.2%	53.5%	41.1%	**12.4%**
IOVRLRES_res_c2	Logistic Regression	78.7%	74.7%	68.3%	**6.4%**
ECOGBL_adsl	Logistic Regression	69.6%	60.5%	65.0%	-4.5%
MCRCSTC_adsl	Logistic Regression	78.9%	75.9%	72.3%	**3.6%**
NUMLES_adsl	MLP	86.1%	32.9%	30.8%	**2.1%**
KRASSTC_adsl	MLP	93.6%	48.0%	49.2%	-1.2%

Table 9.2: Summary of the best-performing classifiers for predicting clinical variables using image embeddings from fine-tuned pretrained model with base model test accuracy and improvement percentage

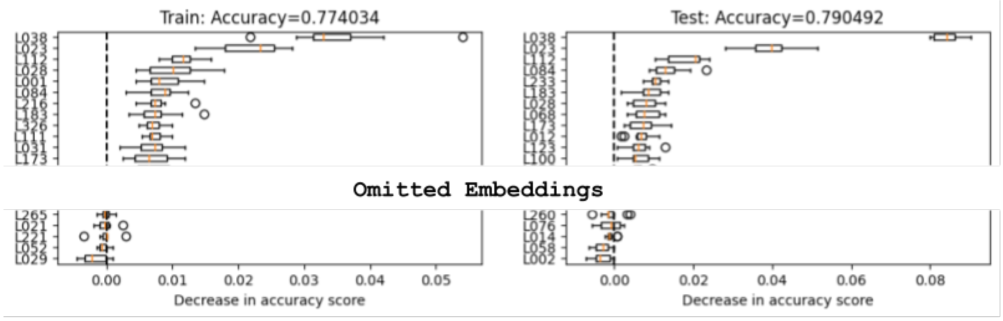


Figure 9.5: Permutation importance scores for the top embedding dimensions in predicting ITRGRESP_res_c3.

After selecting the embedding dimensions, we generated CAM visualizations for a set of representative CT and MR scans from patients with different treatment response outcomes. Figure 9.6 presents the activation maps for dimensions L038 and L023 overlaid on the original input images. The activation maps highlight the regions that strongly contribute to activating these key embedding features.

The CAM visualizations reveal that the fine-tuned model focuses on regions closely identified with the liver when activating the clinically relevant embedding dimensions L038 and L023. Considering that these dimensions were found to be important for predicting treatment response (ITRGRESP_res_c3), it is encouraging to observe that the model is utilizing liver-specific structural information to generate its predictions. This suggests that the fine-tuned model has learned to use baseline features from the liver to predict clinical endpoints. This is crucial because the original therapeutic, TheraSphere, is administered in the liver, and one of the key clinical endpoints is hepatic progression-free survival (hPFS).

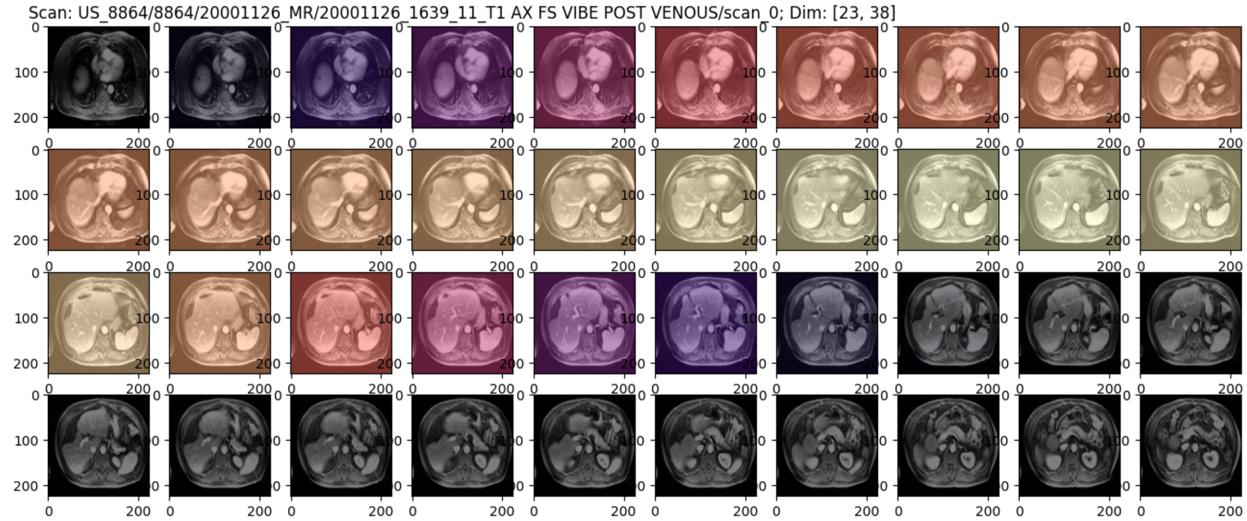

Scan: US_8864/8864/20001126_MR/20001126_1639_11_T1 AX FS VIBE POST VENOUS/scan_0; Dim: [23, 38]
0
100
200
0
200

Figure 9.6: Class Activation Mapping (CAM) visualizations for embedding dimensions L038 and L023 on representative CT and MR scans.

Chapter 10

Conclusion

10.1 Summary of Results and Key Findings

This thesis aimed to identify unique patient subgroups and predictive biomarkers correlating with treatment response to TARE in mCRC patients using advanced computational methods on the multi-modal data collected in the EPOCH trial. The study successfully addressed this primary research question through a series of objectives and analyses.

The first objective was to generate comprehensive statistics on the available categorical and numerical biomarkers and data points in the EPOCH trial dataset and to preprocess the image data to create a high-quality dataset suitable for further analysis. This was achieved through the development of a comprehensive data preprocessing pipeline that curated a high-quality dataset of liver-region CT and MRI scans paired with patient biomarkers. The US-UK dataset, comprising 3,758 unique patient visits and 5,558 image scans, formed the foundation for downstream analysis.

Initial unsupervised clustering using structured data and biomarkers revealed a lack of distinct clusters corresponding to clinical outcomes of interest, such as hepatic progression-free survival (hPFS), time to deterioration, and overall survival (OS). However, Multi-Dimensional Subset Scanning (MDSS) identified a 'poor response group' for hPFS using clinically relevant biomarkers ($p < 0.001$), characterized by bilobar metastatic colorectal cancer, a higher number of lesions (3 or more), an ECOG performance status of 1, and mutant KRAS status. Kaplan-Meier curves demonstrated significant differences in hPFS between the poor and normal responders, both with and without TARE treatment. This addressed the second objective of developing baseline models to group and predict patient outcomes using the preprocessed data.

Cox Proportional Hazards (CoxPH) modeling using baseline structured features revealed hazard ratios aligning with clinical expectations. Features such as ECOG performance status, liver metastasis size, number of liver lesions, KRAS mutation status, and colorectal cancer metastasis were associated with increased hazard rates, while TheraSphere treatment and White race were associated with decreased hazard rates. However, the combined structured features yielded a limited C-index of 0.67, suggesting that they alone may not be sufficient for accurate outcome prediction.

To enhance predictions and address the third objective of fine-tuning pre-trained medical foundation models, deep learning was employed using pre-trained foundation models on liver CT and MRI scans. The study validated the Combined-Modality Sliced3D model's performance on sex classification, achieving a high accuracy of 94.41% on the withheld test set. Fine-tuning the model on the three-class tumor response characteristic (ITRGRESP_res) resulted in a relatively low validation accuracy of 59.60%, suggesting the challenge of learning directly from images to predict complex, multivariate clinical features.

Augmenting the structured data with pre-trained embeddings increased the C-index for hPFS prediction to 0.79, representing a 32% reduction in the error rate compared to biomarkers alone. Fine-tuning the model on ITRGRESP_res and incorporating the resulting embeddings further improved the C-index to an impressive 0.8643, reducing the error rate by 58.9% compared to biomarkers alone. This addressed the fourth objective of extracting image features and generating embeddings to improve classification performance and identify clinically relevant groupings.

Classifiers trained on embeddings showed that fine-tuning the model for one clinically relevant feature improved accuracy for other features. For example, the test accuracy for predicting the treatment arm (TRT01A) increased by 8.5%, while the test accuracies for target lesion response (ITRGRESP_res) improved by 14.0%, 15.0%, and 15.6% for the 4-category, 3-category, and 2-category endpoints, respectively. Class Activation Mapping (CAM) visualization highlighted the importance of liver-focused embeddings for clinical prediction, with the fine-tuned model focusing on regions closely identified with the liver when activating clinically relevant embedding dimensions. This suggests that the model focuses on liver-specific features during prediction, aligning with the clinical relevance of the liver-localized TARE treatment.

In conclusion, we successfully addressed the primary research question and objectives and demonstrated the potential of advanced computational methods to identify unique patient subgroups and predictive biomarkers that correlate with treatment response to TARE in mCRC patients. The progressive analyses, from baseline modeling to the integration of fine-tuned embeddings, showcased the value of combining multi-modal data and deep learning techniques to improve personalized treatment strategies in mCRC.

10.2 Discussion

The Multi-Dimensional Subset Scanning (MDSS) analysis identified a 'poor response group' characterized by specific clinical features, including bilobar metastatic colorectal cancer (MCRCSTC_adsl ['Bilobar']), a higher number of lesions (NUMLES_adsl ['3-5 lesions', '6-10 lesions', '>10 lesions']), an ECOG performance status of 1 (ECOGBL_adsl [1.0]), and mutant KRAS status (KRASSTC_adsl ['Mutant']). These findings align with existing observations in the literature regarding the prognostic significance of these factors in mCRC patients.

Previous studies have shown that patients with bilobar liver metastases have a worse prognosis compared to those with unilobar involvement, likely due to the increased tumor burden and the challenges associated with surgical resection and locoregional therapies [36]. Similarly, a higher number of liver lesions has been consistently associated with poorer outcomes in mCRC, as it reflects a more advanced stage of disease and limits the options for curative-intent treatments [37]. The ECOG performance status, which assesses a patient's level of functioning and ability to carry out daily activities, has been widely recognized as a prognostic factor in various cancer types, including mCRC. Patients with an ECOG status of 1, indicating some restrictions in physically strenuous activity but ability to carry out light work, have been shown to have worse outcomes compared to those with an ECOG status of 0, who are fully active and able to carry out all pre-disease activities without restriction. Lastly, the presence of KRAS mutations has been extensively studied in mCRC and has been associated with reduced responsiveness to certain targeted therapies, such as anti-EGFR agents, and overall poorer prognosis [38]. The identification of mutant KRAS status as a characteristic of the 'poor response group' in this study further reinforces its role as a negative prognostic biomarker in mCRC.

The Cox Proportional Hazards (CoxPH) modeling results revealed that several clinical features, including ECOG status, liver metastasis size (CH1ADMC), number of lesions (NUMLES_adsl), KRAS mutation status, and bilobar metastatic colorectal cancer (MCRCSTC_adsl), were associated with higher hazard ratios for hepatic progression-free survival (hPFS). Notably, four out of these five features overlap with the characteristics of the 'poor response group' identified by MDSS, which demonstrated the consistency of the findings across two different algorithmic approaches. Interestingly, the CoxPH analysis also showed that White race was associated with a lower hazard ratio for hPFS. This observation is consistent with existing literature that has reported racial disparities in cancer outcomes, with non-White patients often experiencing worse survival rates compared to their White counterparts [39]. While the underlying reasons for these disparities are complex and multifactorial, the observation highlights the importance of considering biological ancestry as a potential prognostic factor in mCRC.

An interesting observation during fine-tuning of the Combined-Modality Sliced3D model was that training to predict the three-class tumor response characteristic (ITRGRESP_res) also led to enhanced classifier performance on other clinically relevant features, such as overall response (IOVRLRES_res) and structural features like bilobar metastatic colorectal cancer (MCRCSTC_adsl) and number of lesions (NUMLES_adsl). The latter two, which were associated with higher hazard ratios for hPFS in the CoxPH analysis, also showed improved classifier performance after fine-tuning on ITRGRESP_res, which leads to an intriguing hypothesis. It is possible that the fine-tuned model implicitly learns to recognize structural characteristics that are indicative of poorer prognosis, such as bilobar involvement and a higher number of lesions, when trained to predict clinical endpoints like tumor response.

10.3 Limitations of the Study

One of the primary limitations was the poor performance of classifiers when trained solely on imaging data. Despite the use of state-of-the-art deep learning models, such as the Combined-Modality Sliced3D model, the classifiers struggled to accurately predict complex, multivariate clinical features like tumor response (ITRGRESP_res) directly from images. The relatively low validation accuracy of 59.60% for the three-class ITRGRESP_res classification suggests that learning these intricate relationships from images alone is challenging. This poor performance could be attributed to several factors, including the high variability in the imaging data, the complexity of the clinical features being predicted, and the limited number of unique patients in the dataset.

The high variability in the imaging data poses a significant challenge for the models to learn consistent and generalizable patterns. The EPOCH trial dataset includes CT and MRI scans from various clinical sites, which may have different imaging protocols, equipment, and patient populations. This heterogeneity can introduce noise and confounding factors that make it difficult for the models to capture the underlying relationships between image features and clinical outcomes. Additionally, the presence of different imaging modalities (CT and MRI) further contributes to the variability, as the models need to learn to extract meaningful features from both types of scans.

Another limitation of the study is the relatively small number of unique patients in the dataset, despite the large number of scans available. The US-UK dataset comprises 3,758 unique patient visits and 5,558 image scans, but these scans belong to a smaller number of individual patients. This limitation may hinder the models' ability to capture the full spectrum of patient heterogeneity and to generalize well to unseen data. A larger cohort of unique patients would be beneficial to improve the robustness and generalizability of the models. Furthermore, the study focused primarily on utilizing baseline features and biomarkers to predict treatment response and survival outcomes. While this approach yielded promising results, particularly when augmented with image embeddings, it does not fully capture the temporal dynamics of the disease progression and treatment effects. Incorporating longitudinal data, such as sequential scans and time-varying biomarkers, could provide a more comprehensive understanding of the patient's response to TARE and improve the predictive power of the models.

Lastly, the study's findings are specific to the context of the EPOCH clinical trial, which investigated the impact of TARE in combination with second-line chemotherapy for mCRC patients. While the results demonstrate the potential of advanced computational methods in this setting, the generalizability of the findings to other clinical contexts or patient populations may be limited. It is also important to note that the pretraining of the models used data from the patients eventually included in the CoxPH modeling, which may introduce some information leakage. To address this, proper validation on a new, unseen patient cohort from the EPOCH trial (i.e., from the non-US/UK countries) would be necessary to assess the performance and robustness of the models without the potential bias introduced by the pretraining data overlap. Additional validation studies across different cohorts and treatment settings would also be necessary to assess the broader applicability of the developed models and identified biomarkers beyond the specific context of the EPOCH trial and the US/UK patient population.

10.4 Future Research Directions

One key direction is the exploration of improved deep learning models with a greater understanding of the structural information present in medical images. While the Combined-Modality Sliced3D model used in this study demonstrated the value of integrating imaging data with clinical features, there is room for improvement in terms of model architecture and pretraining strategies. Vision transformers, such as ViT and DINO, have shown remarkable success in capturing long-range dependencies and learning hierarchical representations in natural images. Adapting these architectures to medical imaging tasks and pretraining them on large-scale medical image datasets could potentially improve classification performance on clinical outcomes.

Another promising direction is the incorporation of longitudinal data, such as time series of scans and biomarkers, to capture the temporal dynamics of disease progression and treatment response. Instead of exclusively predicting from baseline features, future studies could explore the use of recurrent neural networks (RNNs) or transformer-based architectures to model the evolution of imaging and clinical features over time. By incorporating this temporal information, the models could potentially identify early indicators of treatment response, predict future disease trajectories, and inform adaptive treatment strategies. This approach would provide a more comprehensive understanding of the patient's response to TARE and enable more personalized and proactive treatment decision-making.

Multi-class training is another area that warrants further investigation. We observed that fine-tuning the Combined-Modality Sliced3D model on the three-class tumor response characteristic (ITRGRESP_res) improved the predictive performance for other clinical features. Thus, extending this approach to simultaneously predict multiple clinically relevant features could potentially lead to the generation of more informative and generalizable embeddings. By training the models to capture the relationships between various clinical endpoints, such as tumor response, survival outcomes, and treatment-related adverse events, the resulting embeddings could provide a more comprehensive representation of the patient's overall disease state and treatment response and also help to mitigate the challenges associated with limited sample sizes.

Additionally, we plan to study which embedding features are shared and correlated across different clinical endpoints and uncover the latent structure and relationships within the embedding space. Visualizing and interpreting these relationships could guide the development of more targeted and effective treatment strategies, as well as inform the design of follow up clinical trials to EPOCH.

I would like to acknowledge the use of AI language models, specifically for assisting in the proofreading and grammatical improvement of the final draft of this thesis, in accordance with the EECS Department guidelines effective April 20, 2023.

Appendix A

Supplemental Figures

This appendix includes all supplemental figures and tables.

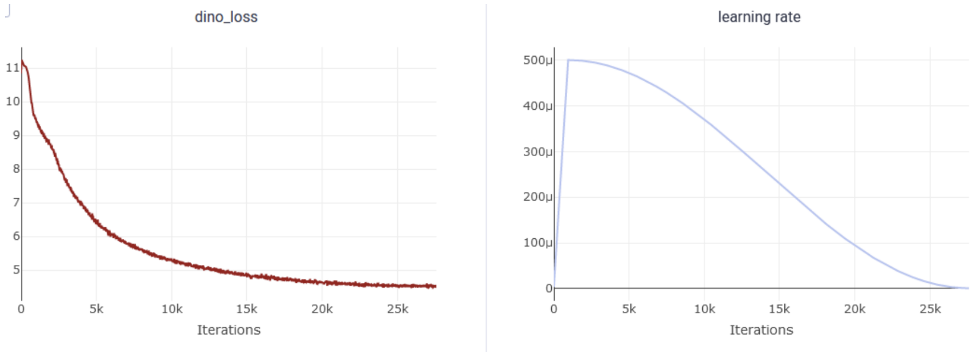


Figure A.1: Training loss and learning rate plots for pretrained 2D-Combined Modality model.

Table A.1: Evaluation metrics for the Combined-Modality Sliced3D model on ITRGRESP_res classification (withheld test set)

Metric	Value
Accuracy	0.5960
Precision (macro)	0.5280
Recall (macro)	0.5740
F1 score (macro)	0.5209
Precision (micro)	0.5960
Recall (micro)	0.5960
F1 score (micro)	0.5960

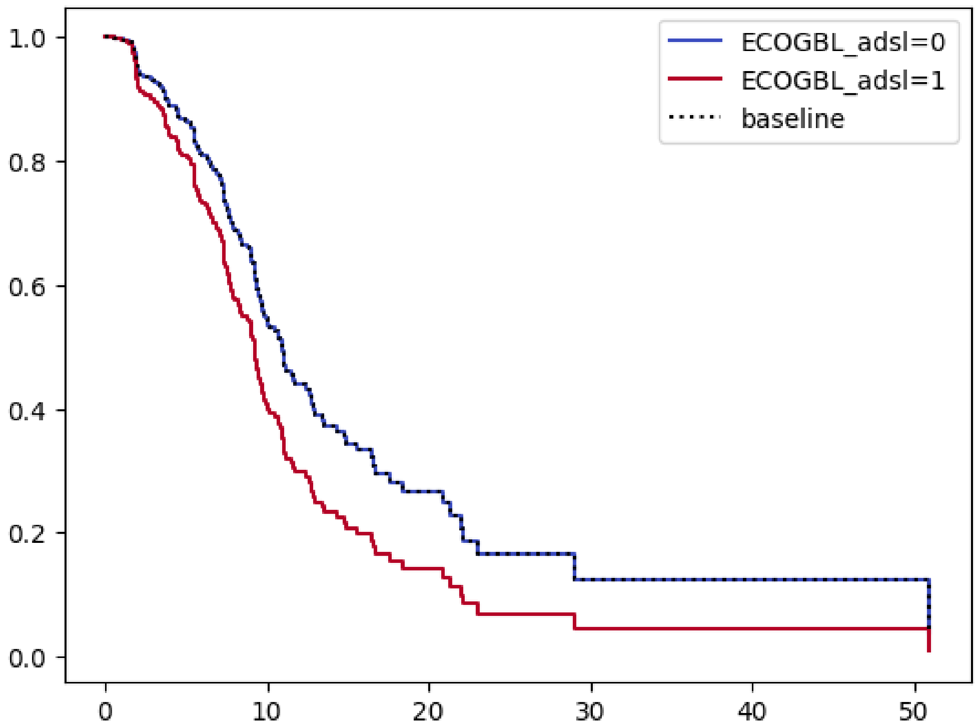


Figure A.2: Kaplan-Meier curve for hepatic progression-free survival stratified by baseline ECOG performance status.

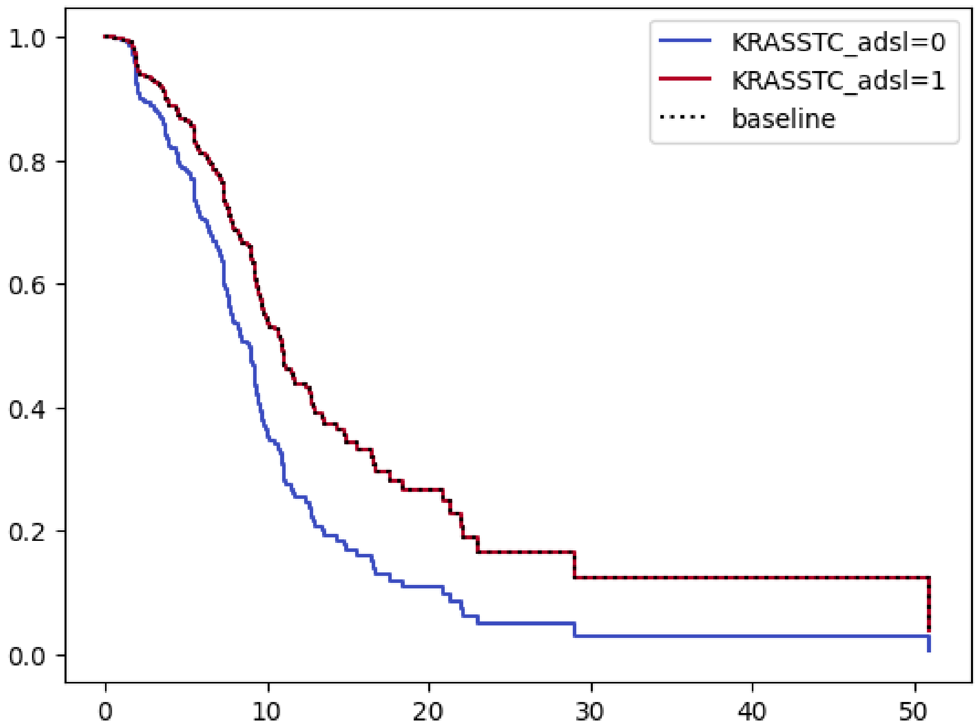


Figure A.3: Kaplan-Meier curve for hepatic progression-free survival stratified baseline KRAS status.

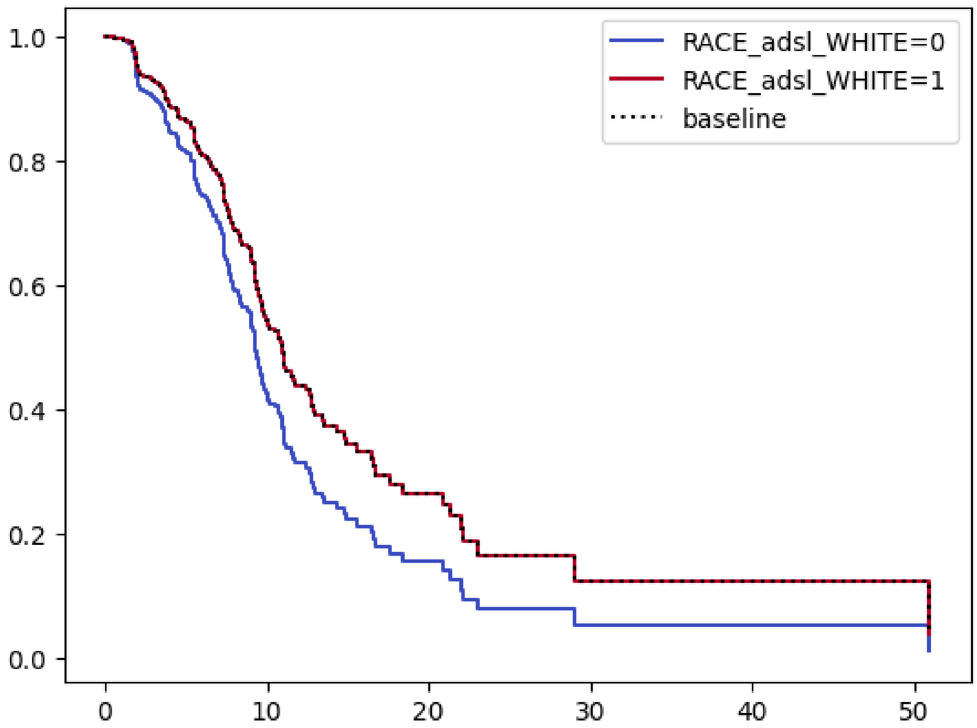


Figure A.4: Kaplan-Meier curve for hepatic progression-free survival stratified by white or nonwhite status.

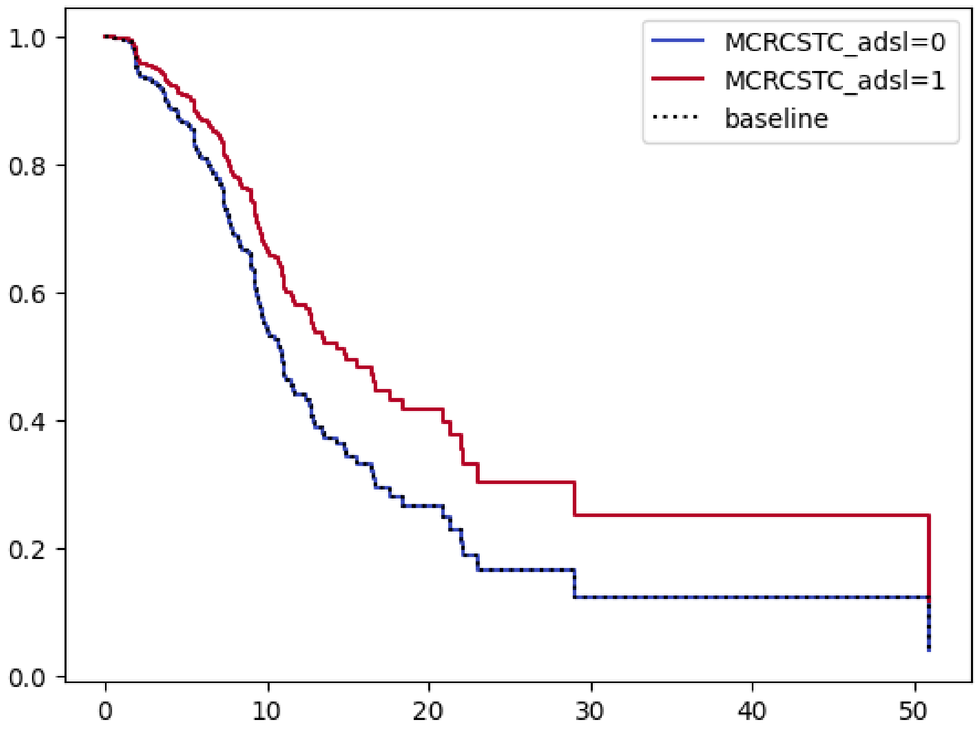


Figure A.5: Kaplan-Meier curve for hepatic progression-free survival stratified by MCRCSTC (degree of metastasis).

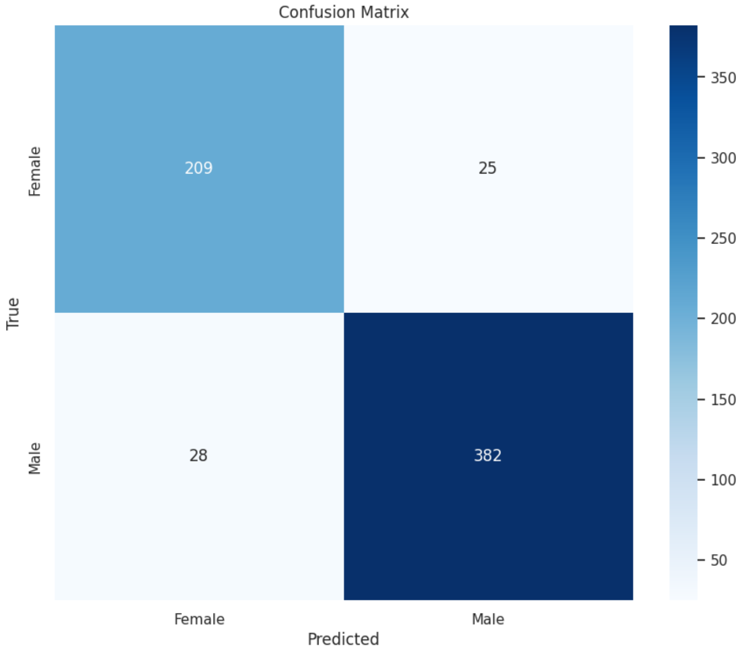


Figure A.6: Confusion matrix for sex classification using the 3D MedicalNet model.

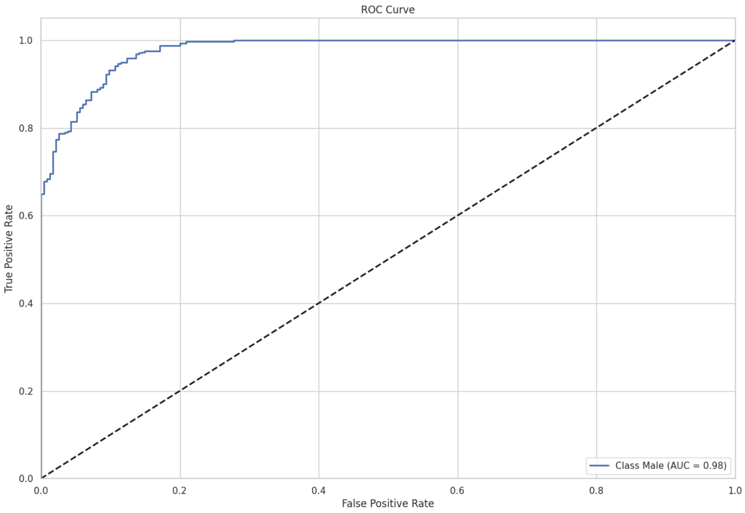


Figure A.7: ROC curve for sex classification using the 3D MedicalNet model.

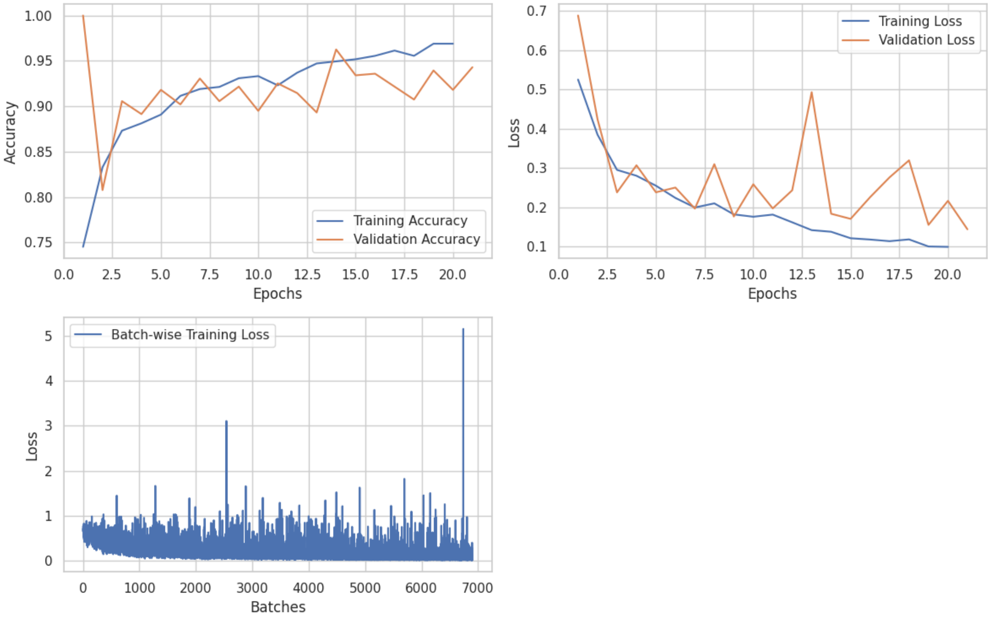


Figure A.8: Training and validation accuracy of the Combined-Modality Sliced3D model for sex classification.

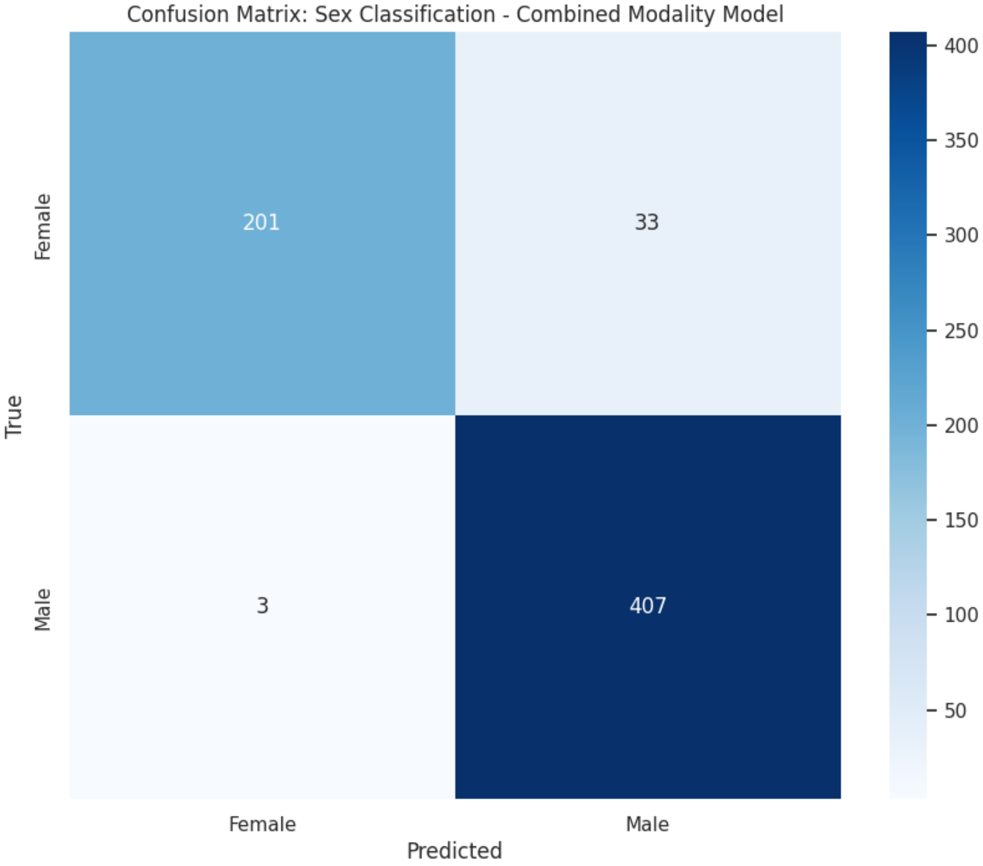


Figure A.9: Confusion matrix of the Combined-Modality Sliced3D model for sex classification.

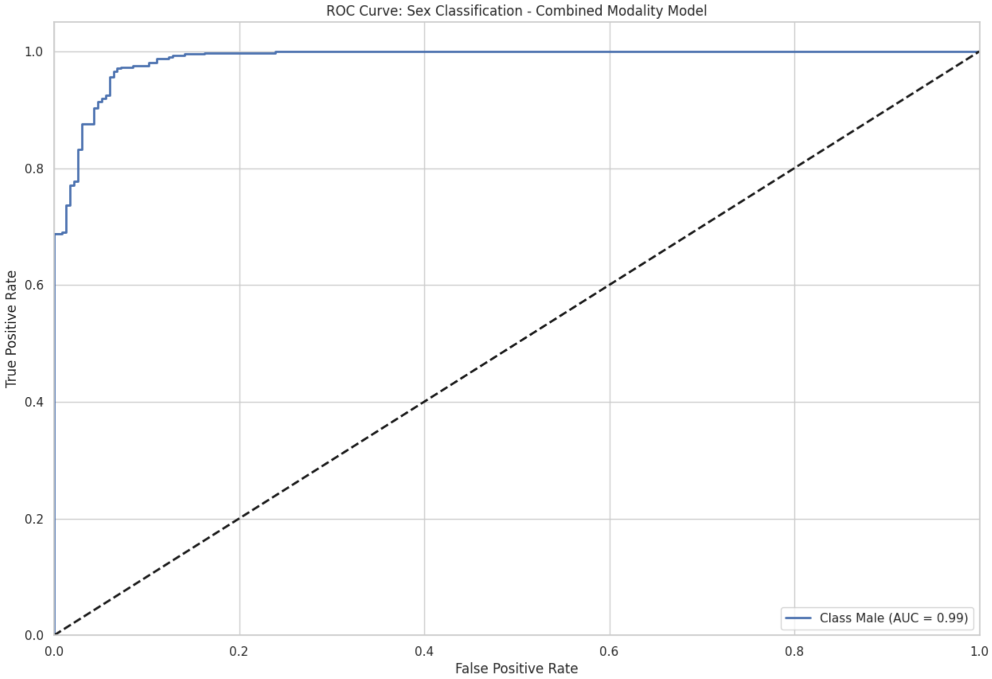


Figure A.10: ROC curve of the Combined-Modality Sliced3D model for sex classification.

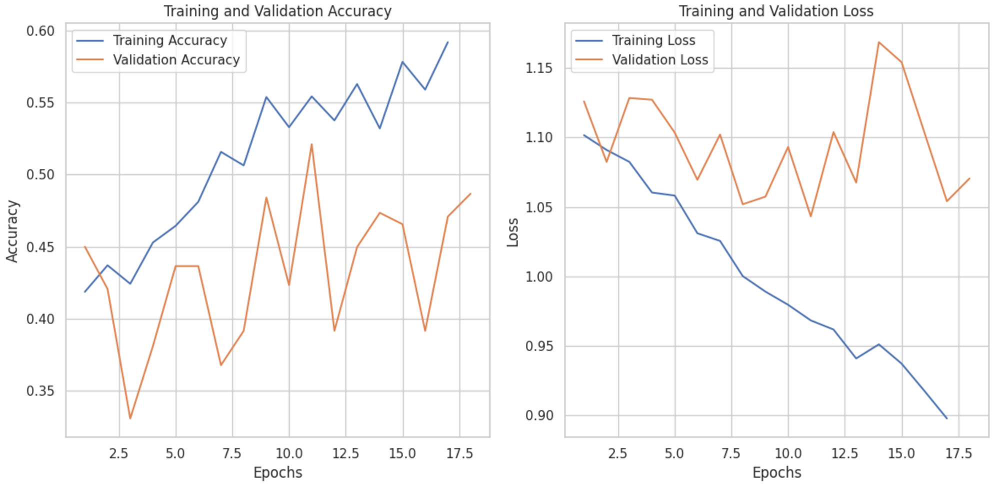


Figure A.11: Training and validation accuracy and loss curves for the Combined-Modality Sliced3D model fine-tuned on ITRGRESP_res classification

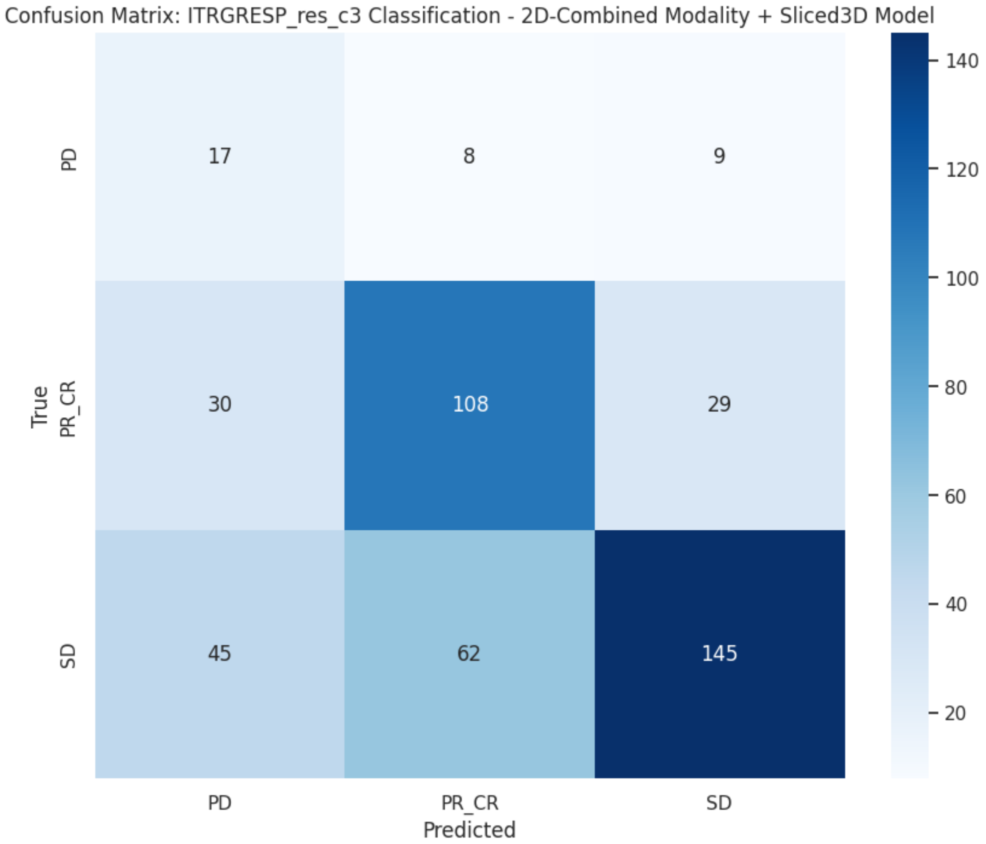


Figure A.12: Confusion matrix for the Combined-Modality Sliced3D model on ITRGRESP_res classification (withheld test set)

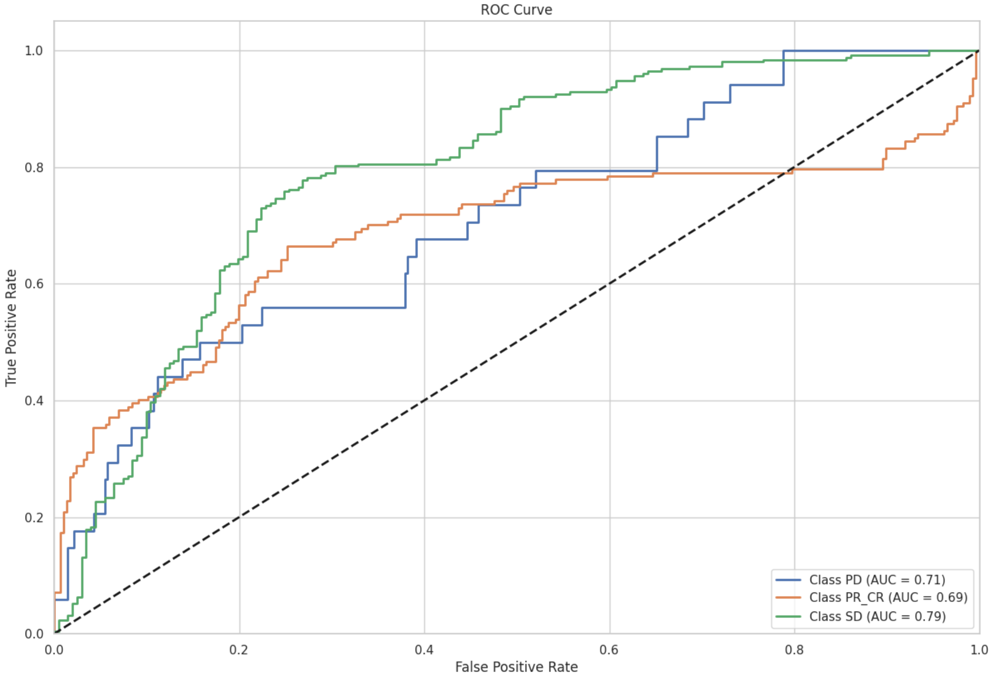


Figure A.13: ROC curve for the Combined-Modality Sliced3D model on ITRGRESP_res classification (withheld test set)

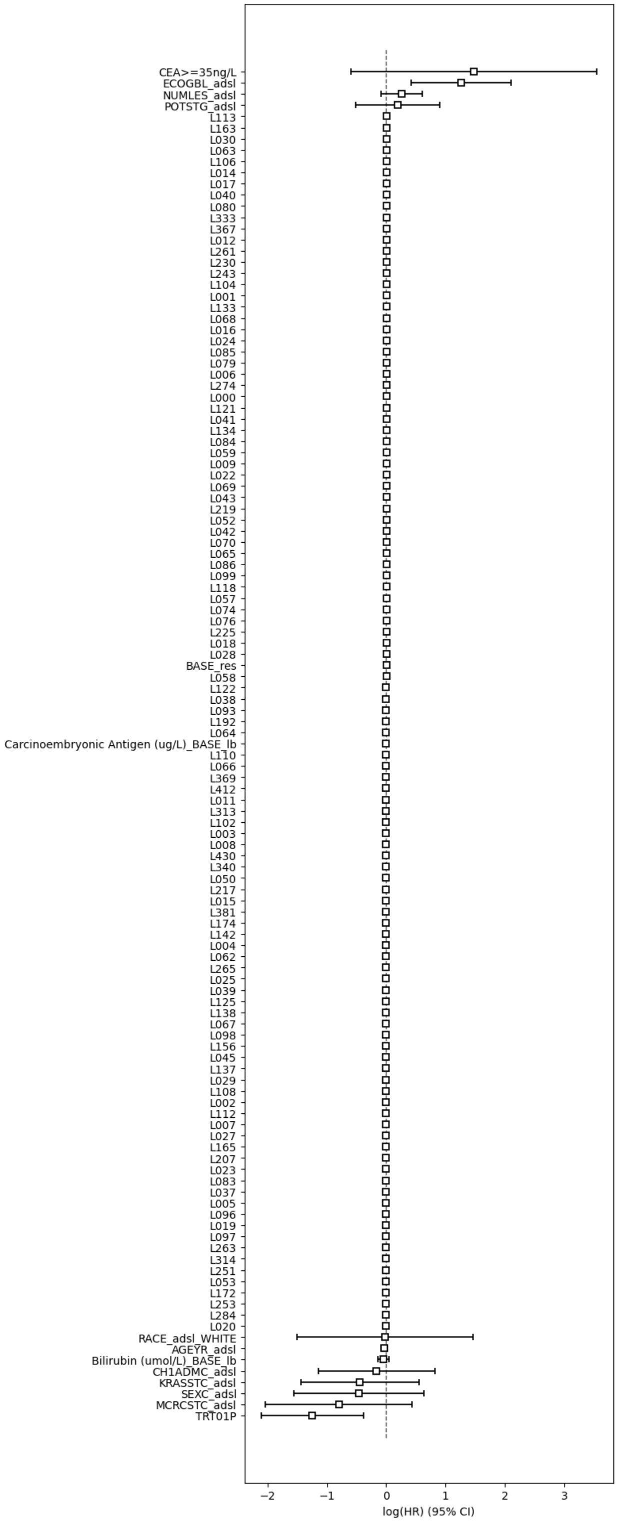


Figure A.14: Hazard ratios and 95% confidence intervals for hepatic progression-free survival (hPFS) estimated using a CoxPH model with features from the fine-tuned embeddings.

References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," en, *CA Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, Jan. 2021.

[2] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," en, *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[3] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," en, *NPJ Digit. Med.*, vol. 5, no. 1, p. 48, Apr. 2022.

[4] C. Sudlow, J. Gallacher, N. Allen, *et al.*, "UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," en, *PLoS Med.*, vol. 12, no. 3, e1001779, Mar. 2015.

[5] P. Rajpurkar, C. O'Connell, A. Schechter, *et al.*, "CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV," en, *NPJ Digit. Med.*, vol. 3, no. 1, p. 115, Sep. 2020.

[6] M. F. Mulcahy, A. Mahvash, M. Pracht, *et al.*, "Radioembolization with chemotherapy for colorectal liver metastases: A randomized, open-label, international, multicenter, phase III trial," en, *J. Clin. Oncol.*, vol. 39, no. 35, pp. 3897–3907, Dec. 2021.

[7] J. M. Peppercorn, J. C. Weeks, E. F. Cook, and S. Joffe, "Comparison of outcomes in cancer patients treated within and outside clinical trials: Conceptual framework and structured review," en, *Lancet*, vol. 363, no. 9405, pp. 263–270, Jan. 2004.

[8] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," in *Advances in Experimental Medicine and Biology*, ser. Advances in experimental medicine and biology, Cham: Springer International Publishing, 2020, pp. 3–21.

[9] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," en, *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021.

[10] M. Miranda Baleiras, T. Dias Domingues, E. Severino, C. Vasques, M. T. Neves, A. Ferreira, L. Vasconcelos de Matos, F. Ferreira, H. Miranda, and A. Martins, "Prognostic impact of type 2 diabetes in metastatic colorectal cancer," en, *Cureus*, vol. 15, no. 1, e33916, Jan. 2023.

[11] J. Uhlig, J. Lukovic, L. A. Dawson, R. A. Patel, M. J. Cavnar, and H. S. Kim, "Locoregional therapies for colorectal cancer liver metastases: Options beyond resection," en, *Am. Soc. Clin. Oncol. Educ. Book*, vol. 41, no. 41, pp. 133–146, Mar. 2021.

[12] J. Engstrand, H. Nilsson, C. Strömberg, E. Jonas, and J. Freedman, "Colorectal cancer liver metastases – a population-based study on incidence, management and survival," en, *BMC Cancer*, vol. 18, no. 1, Dec. 2018.

[13] J. Martin, A. Petrillo, E. C. Smyth, *et al.*, "Colorectal liver metastases: Current management and future perspectives," en, *World J. Clin. Oncol.*, vol. 11, no. 10, pp. 761–808, Oct. 2020.

[14] H. Zhou, Z. Liu, Y. Wang, *et al.*, "Colorectal liver metastasis: Molecular mechanism and interventional therapy," en, *Signal Transduct. Target. Ther.*, vol. 7, no. 1, p. 70, Mar. 2022.

[15] R. Salem and K. G. Thurston, "Radioembolization with yttrium-90 microspheres: A state-of-the-art brachytherapy treatment for primary and secondary liver malignancies: Part 3: Comprehensive literature review and future direction," en, *J. Vasc. Interv. Radiol.*, vol. 17, no. 10, pp. 1571–1593, Oct. 2006.

[16] N. Chauhan, M. F. Mulcahy, R. Salem, *et al.*, "TheraSphere yttrium-90 glass microspheres combined with chemotherapy versus chemotherapy alone in second-line treatment of patients with metastatic colorectal carcinoma of the liver: Protocol for the EPOCH phase 3 randomized clinical trial," en, *JMIR Res. Protoc.*, vol. 8, no. 1, e11545, Jan. 2019.

[17] G. Li Destri, A. S. Rubino, R. Latino, F. Giannone, R. Lanteri, B. Scilletta, and A. Di Cataldo, "Preoperative carcinoembryonic antigen and prognosis of colorectal cancer. an independent prognostic factor still reliable," en, *Int. Surg.*, vol. 100, no. 4, pp. 617–625, Apr. 2015.

[18] C. S. Karapetis, S. Khambata-Ford, D. J. Jonker, *et al.*, "K-ras mutations and benefit from cetuximab in advanced colorectal cancer," en, *N. Engl. J. Med.*, vol. 359, no. 17, pp. 1757–1765, Oct. 2008.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016.

[20] F. Rousseau, L. Drumetz, and R. Fablet, "Residual networks as flows of diffeomorphisms," en, *J. Math. Imaging Vis.*, vol. 62, no. 3, pp. 365–375, Apr. 2020.

[21] K. Hara, H. Kataoka, and Y. Satoh, *Learning spatio-temporal features with 3d residual networks for action recognition*, 2017. arXiv: 1708.07632 [cs.CV].

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html.

[23] M.-J. Zhang, "Cox proportional hazards regression models for survival data in cancer research," in *Biostatistical Applications in Cancer Research*, ser. Cancer treatment and research, Boston, MA: Springer US, 2002, pp. 59–70.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785.

[25] J. Wasserthal, H.-C. Breit, M. T. Meyer, *et al.*, "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images," en, *Radiol. Artif. Intell.*, vol. 5, no. 5, e230024, Sep. 2023.

[26] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," en, *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[27] Z. Zhang and D. B. Neill, *Identifying significant predictive bias in classifiers*, 2017. arXiv: 1611.08292 `[stat.ML]`.

[28] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *ArXiv*, vol. abs/1810.01943, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52922804.

[29] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," 2021.

[30] M. Raboh, D. Levanony, P. Dufort, and A. Sitek, "Context in medical imaging: The case of focal liver lesion classification," in *Medical Imaging 2022: Image Processing*, I. Išgum and O. Colliot, Eds., San Diego, United States: SPIE, Apr. 2022.

[31] S. Chen, K. Ma, and Y. Zheng, *Med3d: Transfer learning for 3d medical image analysis*, 2019. arXiv: 1904.00625 `[cs.CV]`.

[32] A. Golts, M. Raboh, Y. Shoshan, S. Polaczek, S. Rabinovici-Cohen, and E. Hexter, "Fusemedml: A framework for accelerated discovery in machine learning based biomedicine," *Journal of Open Source Software*, vol. 8, no. 81, p. 4943, 2023. DOI: 10.21105/joss.04943. [Online]. Available: https://doi.org/10.21105/joss.04943.

[33] E. A. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," en, *Eur. J. Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization.," *CVPR*, 2016.

[36] K. Homayounfar, A. Bleckmann, L. C. Conradi, *et al.*, "Bilobar spreading of colorectal liver metastases does not significantly affect survival after R0 resection in the era of interdisciplinary multimodal treatment," en, *Int. J. Colorectal Dis.*, vol. 27, no. 10, pp. 1359–1367, Oct. 2012.

[37] G. P. Kanas, A. Taylor, J. N. Primrose, W. J. Langeberg, M. A. Kelsh, F. S. Mowat, D. D. Alexander, M. A. Choti, and G. Poston, "Survival after liver resection in metastatic colorectal cancer: Review and meta-analysis of prognostic factors," en, *Clin. Epidemiol.*, vol. 4, pp. 283–301, Nov. 2012.

[38] M. S. Alkader, R. Z. Altaha, S. A. Badwan, *et al.*, "Impact of KRAS mutation on survival outcome of patients with metastatic colorectal cancer in jordan," en, *Cureus*, vol. 15, no. 1, e33736, Jan. 2023.

[39] V. A. Zavala, P. M. Bracci, J. M. Carethers, *et al.*, "Cancer health disparities in racial/ethnic minorities in the united states," en, *Br. J. Cancer*, vol. 124, no. 2, pp. 315–332, Jan. 2021.

I would like to acknowledge the use of AI language models, specifically for assisting in the proofreading and grammatical improvement of the final draft of this thesis, in accordance with the EECS Department guidelines effective April 20, 2023.