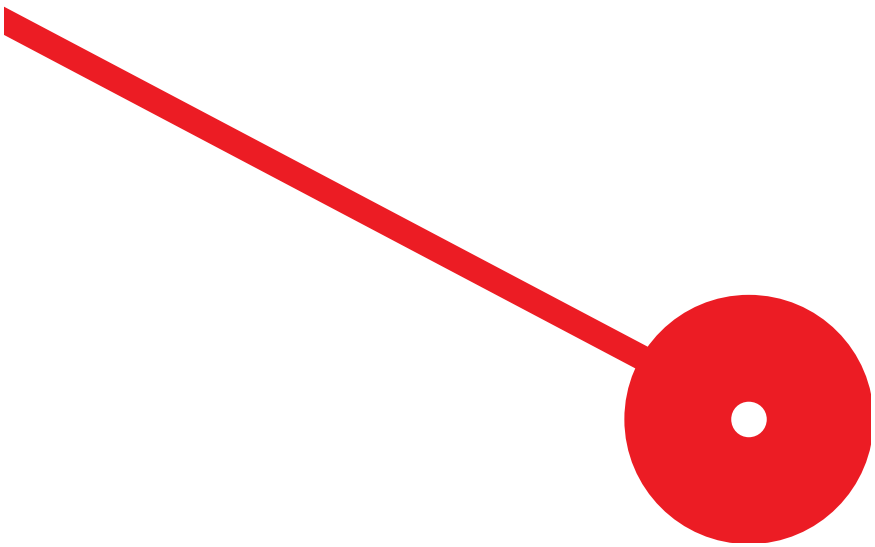
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Multimodal Data Integration in Oncology: A Case Study of LIHC

João Pedro Miranda Casais

10/2024



João Pedro Miranda Casais. Multimodal Data Integration in Oncology: A Case Study of LIHC 10/2024

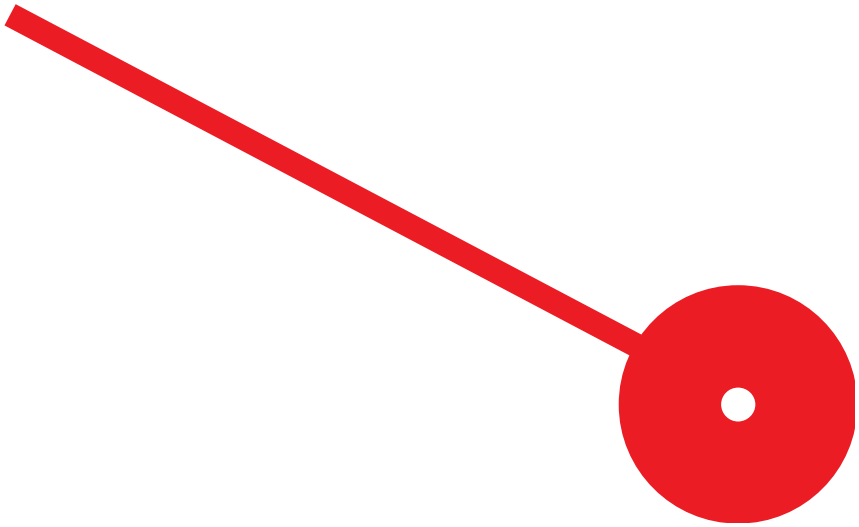
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Multimodal Data Integration in Oncology: A Case Study of LIHC

João Pedro Miranda Casais

Dissertação de Mestrado apresentada ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em Business Intelligence and Analytics, sob orientação da Doutora Patrícia Alexandra Gregório Ramos e do Doutor José Manuel Soares Oliveira.



João Pedro Miranda Casais. Multimodal Data Integration in Oncology: A Case Study of LIHC 10/2024

Dedicatória:

Gostaria de dedicar esta dissertação à minha sobrinha/afilhada que nasceu em 4 de outubro de 2024. Que a sua vida seja sempre iluminada e cheia de conquistas.

Agradecimentos:

Para além de tal, foi necessário e importante um empenhamento próprio, bem como várias pessoas que me acompanharam nesta etapa. Gostaria de agradecer a todos aqueles que me ajudaram neste percurso académico e que me sempre apoiaram e motivaram.

Quero agradecer a todos os professores da Pós-Graduação de Business Analytics e do Mestrado em Business Intelligence and Analytics por serem professores excelentes, por serem empenhados no âmbito académico e por demonstrarem sempre disponibilidade e interesse.

Agradeço a todos os meus colegas académicos por permitirem acompanhar e tornar a minha experiência académica incrível e única.

Gostaria de agradecer aos meus orientadores da dissertação, Doutora Patrícia Ramos, e ao co-orientador, Doutor José Oliveira, pela disponibilidade na transmissão de conhecimento, aconselhamento e sugestões sobre o trabalho decisivo para a motivação e à realização do presente trabalho;

Agradeço também à minha família e amigos pelo apoio incondicional prestado durante toda a minha vida.

Abstract:

This thesis focuses on the application of multimodal learning, specifically deep learning techniques, to enhance the diagnosis and treatment of Liver Hepatocellular Carcinoma. The increasing incidence and late detection of this aggressive form of cancer necessitate innovative approaches for early identification and personalized therapies. The core objective of this thesis was to evaluate the impact of integrating diverse data modalities, such as clinical, genetic, and imaging data, on improving diagnostic precision and therapeutic outcomes for Liver Hepatocellular Carcinoma patients. The study investigates the limitations of traditional unimodal data analysis in understanding the complex nature of Liver Hepatocellular Carcinoma. It posits that multimodal learning can leverage complementary strengths of different data types, leading to a more comprehensive understanding of tumor biology and individual patient profiles. The research explores the use of advanced deep learning architectures for processing and integrating these diverse data modalities.

A key component of the thesis involves a practical application using a dataset sourced from the American National Cancer Institute, focusing on predicting the vital status of Liver Hepatocellular Carcinoma patients. The dataset is preprocessed and analyzed using the AutoGluon-Multimodal (AutoMM) framework, an open-source automated machine learning library designed for multimodal learning tasks. Various model combinations incorporating image, text, and tabular data are evaluated based on classification performance metrics, including accuracy, Matthews Correlation Coefficient, precision, recall, F1 score, and ROC-AUC. The results demonstrate the superior performance of models combining tabular and text data in predicting the vital status of Liver Hepatocellular Carcinoma patients.

Keywords: Liver Hepatocellular Carcinoma | Precision Medicine | Machine Learning Models | Early Detection | Deep Learning

Resumo:

Esta tese centra-se na aplicação de aprendizagem multimodal, especificamente técnicas de aprendizagem profunda, para melhorar o diagnóstico e o tratamento do Carcinoma Hepatocelular do Fígado. A incidência crescente e a deteção tardia desta forma agressiva de cancro requerem abordagens inovadoras para a identificação precoce e terapias personalizadas. O principal objetivo desta tese foi avaliar o impacto da integração de diversas modalidades de dados, tais como dados clínicos, genéticos e imagiológicos, na melhoria da precisão do diagnóstico e dos resultados terapêuticos para os doentes com Carcinoma Hepatocelular do Fígado. O estudo investiga as limitações da análise tradicional de dados unimodais na compreensão da natureza complexa do Carcinoma Hepatocelular do Fígado. O estudo defende que a aprendizagem multimodal pode tirar partido dos pontos fortes complementares dos diferentes tipos de dados, conduzindo a uma compreensão mais abrangente da biologia do tumor e dos perfis individuais dos doentes. A investigação explora a utilização de arquitecturas avançadas de aprendizagem profunda para o processamento e a integração destas diversas modalidades de dados.

Uma componente fundamental da tese envolve uma aplicação prática utilizando um conjunto de dados provenientes do Instituto Nacional do Cancro americano, centrado na previsão do estado vital dos doentes com Carcinoma Hepatocelular do Fígado. O conjunto de dados é pré-processado e analisado utilizando a estrutura AutoGluon-Multimodal (AutoMM), uma biblioteca de aprendizagem automática de código aberto concebida para tarefas de aprendizagem multimodal. Várias combinações de modelos que incorporam dados de imagem, texto e tabulares são avaliadas com base em métricas de desempenho de classificação, incluindo exatidão, Coeficiente de Correlação de Matthews, precisão, recuperação, pontuação F1 e ROC-AUC. Os resultados demonstram o desempenho superior dos modelos que combinam dados tabulares e de texto na previsão do estado vital dos doentes com Carcinoma Hepatocelular do Fígado.

Palavras-chave: Carcinoma Hepatocelular do Fígado | Medicina de Precisão | Modelos de Aprendizagem Automática | Deteção Precoce | Deep Learning

Contents

List of Figures

List of Tables

List of Abbreviations

AI - Artificial Intelligence

AutoMM – AutoGluon – Multimodal

AutoML - Automated Machine Learning

CLIP - Contrastive Language-Image Pre-training

CNN - Convolutional Neural Networks

CT - Computed Tomography

DL - Deep Learning

DNA - Deoxyribonucleic Acid

EHR - Electronic Health Records

GNN - Graph Neural Networks

GPT-3 - Generative Pre-Training Transformer 3

LIHC - Liver Hepatocellular Carcinoma

MCC - Matthews Correlation Coefficient

MDL - Multimodal Deep Learning

ML - Machine Learning

MML - Multimodal Learning

MRI - Magnetic Resonance Imaging

NCI - National Cancer Institute

NLP - Natural Language Processing

PCA - Principal Component Analysis

PET - Positron Emission Tomography

RNN - Recurrent Neural Networks

ROI - Regions of Interest

TCGA - The Cancer Genome Atlas

WES - Whole-Exome Sequencing

WGS - Whole-Genome Sequencing

WSI - Whole Slide Imaging

1 Introduction

Liver Hepatocellular Carcinoma (LIHC), comprising a heterogeneous group of aggressive tumors, is rising in Europe, necessitating innovative diagnostic and therapeutic approaches. Late detection, attributed to limited knowledge of risk factors and a dearth of diagnostic biomarkers, underscores the urgency for advancements in early identification. These tumors exhibit high resistance to conventional chemotherapy, creating an imperative for novel therapeutic strategies. Targeted therapies, while promising for specific patient subgroups, face challenges of uneven distribution across Europe, demanding urgent implementation plans to ensure equitable access for patients.

In this context, the transformative potential of data-driven precision medicine, specifically leveraging Machine Learning (ML) models and Deep Learning (DL) approaches, is being explored to enhance diagnostic precision and prognostic insights in LIHC. The integration of Artificial Intelligence (AI) based strategies into medical imaging can contribute significantly to the development of personalized and effective care strategies, addressing the pressing need for early detection and innovative therapeutic interventions.

Similarly, in the field of Natural Language Processing (NLP), pre-training methods that learn directly from raw text have revolutionized the field. Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface has enabled task-agnostic architectures to zero-shot transfer to downstream datasets, removing the need for specialized output heads or dataset-specific customization. Flagship systems like Generative Pre-Training Transformer 3 (GPT-3) are now competitive across many tasks with bespoke models while requiring little to no dataset-specific training data.

This success in NLP suggests that scalable pre-training methods could result in similar breakthroughs in other fields, such as computer vision. Previous work has explored improving content-based image retrieval and learning more data-efficient image representations by predicting words in image captions. Recent advances have demonstrated the potential of transformer-based language modeling, masked language modeling, and contrastive objectives to learn image representations from text. For

instance, models like Contrastive Language-Image Pre-training (CLIP) have shown significant promise in learning from natural language supervision on a large scale. These models have been trained on vast amounts of publicly available data, demonstrating efficient learning and competitive performance on a wide set of tasks during pre-training.

The integration of these advanced AI techniques into medical imaging for LIHC could revolutionize the field, like the advancements seen in NLP and computer vision. By critically assessing and implementing AI-based strategies, we can enhance diagnostic precision and prognostic insights, ultimately contributing to the development of personalized and effective care strategies for patients with LIHC. This approach highlights the significant potential of leveraging data-driven precision medicine and AI to address complex challenges in oncology and beyond.

1.1 Motivation

LIHC, poses significant challenges due to its insidious nature, often leading to late-stage diagnosis and limited treatment options (Banales et al., 2016). Despite advancements in medical imaging, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), LIHCs remain elusive in accurate characterization and prognosis (Razumilava & Gores, 2014). The integration of AI into medical imaging has emerged as a promising avenue to address these challenges, but its application in LIHC remains underexplored, highlighting the critical need for comprehensive research and review.

The prognosis for LIHC patients is notably poor, with survival rates lagging behind those of many other cancers, underscoring the urgent need for improved diagnostic and therapeutic strategies (Bridgewater et al., 2014). Despite being considered rare, the global incidence of LIHC is on the rise, placing a significant burden on healthcare systems and societies worldwide (Blechacz & Gores, 2008). The etiology of LIHC is multifactorial, with risk factors such as chronic inflammation playing a central role in disease pathogenesis (Saha et al., 2014). Recognizing the complexity of LIHC, initiatives aim to advance precision medicine approaches tailored to individual patient profiles, including early detection, prevention, and personalized treatment plans.

Recent technological advancements have revolutionized the understanding and treatment of LIHC, particularly in the realm of molecular profiling and targeted therapies. Precision medicine holds immense promise, with a significant proportion of LIHC cases

harboring targetable molecular abnormalities that can be exploited for therapeutic benefit (Rizvi et al., 2018). Through the integration of AI into precision medicine frameworks researchers and clinicians can leverage AI-driven analytics to derive insights from diverse datasets encompassing clinical, genetic, and imaging data, thereby optimizing patient outcomes (Lowery et al., 2018).

Despite the potential of AI-driven imaging techniques in LIHC, several challenges persist on the path to clinical integration. Issues such as algorithmic bias, data privacy concerns, and the need for robust validation frameworks necessitate collaborative efforts from multidisciplinary stakeholders to ensure the ethical and effective implementation of AI technologies in clinical practice. Additionally, the development of AI algorithms tailored specifically for LIHC, coupled with rigorous clinical validation through randomized trials, is imperative to establish the reliability and efficacy of these tools in real-world healthcare settings (Saha et al., 2014).

1.2 Objectives

The potential of ML and DL to revolutionize diagnostics in healthcare is immense. In the context of LIHC, these technologies offer a promising avenue to enhance diagnostic precision and prognostic insights. ML models can analyze vast amounts of medical data to identify patterns and correlations that might be imperceptible to human clinicians. By leveraging DL techniques, particularly in imaging, it's possible to achieve more accurate early detection of LIHC, which is crucial given the disease's aggressive nature and resistance to conventional chemotherapy (Figure 1).

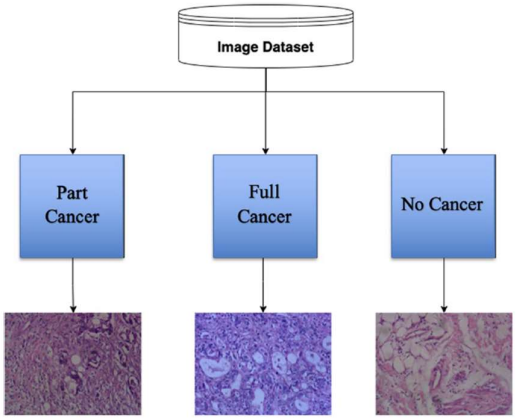


Figure 1 - Image dataset structure.

The integration of these advanced models into routine diagnostic workflows could significantly improve patient outcomes by enabling earlier and more accurate diagnoses. AI has the potential to revolutionize medical imaging, providing tools that can analyze images with greater accuracy and speed than traditional methods. In the context of LIHC, AI algorithms can be trained to recognize subtle signs of cancer in imaging studies such as CT scans and MRIs. These algorithms use DL techniques to identify patterns and anomalies that may indicate the presence of cancerous tissues. In Figure 2, the axial CT image shows a polypoidal gallbladder cancer arising from the fundus. The region of interest (red circle) is drawn over the lesion to extract the radiomic features shown in the table.

mean	sd	entropy	mpp	skewness	kurtosis
52.86000061	22.13999939	4.139999866	55.40000153	-1.309999943	2.710000038
28.54999924	71.79000092	4.980000019	69.06999969	-0.340000004	0.270000011
30.81999969	64.55999756	4.860000134	59.77999878	-0.860000014	1.870000005
33.27000046	66.90000153	4.909999847	62.47000122	-1.090000033	1.340000033
36.45999908	72.91000366	4.889999866	73.04000092	-0.960000019	0.310000002
38.04999924	76.94999695	5.019999981	78.56999969	-0.680000007	-0.360000014
95.63999939	31.37000084	4.150000095	95.63999939	-0.610000014	-0.639999986
25.85000038	67.5	4.690000057	64.27999878	-0.409999996	-0.479999989
40.56999969	87.68000031	4.730000019	94.04000092	-0.280000001	-0.939999998
50.15000153	102.0100021	4.71999979	107.5299988	0	-0.980000019
52.40000153	104.8199997	4.730000019	110.8799973	0.079999998	-0.920000017
46.83000183	98.01000214	4.699999809	103.6999969	-0.01	-0.790000021

Figure 2 - CT based radiomics in gallbladder cancer.

Additionally, AI can assist in the segmentation and quantification of tumors, helping to determine their size, shape, and progression (Helber et al., 2019). This detailed analysis supports more precise treatment planning and monitoring, ultimately improving patient outcomes. The integration of AI into medical imaging also facilitates remote diagnostics, enabling access to high-quality care in underserved regions.

The objective of this thesis is to conduct a comprehensive study on the impact of multimodal data integration in oncology, with a particular focus on Liver Hepatocellular Carcinoma. This research aims to explore how the integration of diverse data modalities—such as clinical, genetic, and imaging data—can enhance diagnostic precision and improve therapeutic outcomes for LIHC patients. By leveraging Deep Learning techniques across these modalities, the study seeks to assess the effectiveness

of data-driven approaches in advancing personalized medicine and developing more effective, targeted treatment strategies for LIHC.

1.3 Document Outline

The thesis begins with an introduction to the challenges posed by Liver Hepatocellular Carcinoma (LIHC), emphasizing its increasing incidence, late detection, and resistance to traditional therapies. The introduction highlights the need for innovative approaches, specifically data-driven precision medicine leveraging machine learning and deep learning techniques, to enhance diagnostic accuracy and improve patient outcomes.

The thesis then provides a foundational understanding of multimodal learning (MML) within the context of oncology. This section explores the different data modalities used in cancer research, including clinical data, radiological imaging, histopathological imaging, genomics, and proteomics. The chapter also explains the importance of data fusion strategies, such as early fusion, late fusion, and cross-attention for integrating diverse data modalities into a unified framework.

Next, the thesis outlines the methodology employed for the empirical study. It describes the use of AutoGluon-Multimodal (AutoMM), an open-source automated machine learning library designed specifically for multimodal learning tasks. The thesis details the key features of AutoMM, including data format and processing, intuitive APIs for training and evaluation, a late-fusion architecture for handling multiple data types, and support for advanced tasks beyond basic classification and regression.

The thesis then presents the empirical study, which focuses on predicting the vital status of LIHC patients using a dataset from the American National Cancer Institute (NCI). The study outlines the process of data extraction from the NCI database, data preparation steps such as converting image files, data identification and cleanup, data splitting for training and testing, and label encoding for the target variable. The study also explains the performance metrics used to evaluate the models, including accuracy, Matthews correlation coefficient (MCC), precision, recall, F1 score, and ROC-AUC.

The thesis concludes by presenting and analyzing the results obtained from the empirical study. It compares the performance of various model combinations using different data modalities, highlighting the superior performance of the model combining tabular and text data. The thesis discusses the limitations of the study, particularly the relatively small

dataset size, and suggests directions for future research, such as extending the analysis to other cancer types.

2 Fundamentals of Multimodal Learning

Within the field of oncology, multimodal learning (MML) represents a ground-breaking advance in technology/prediction using ML and AI. Due to the complexity and heterogeneity of cancer, integration of different types of data, modalities in particular, is essential for both treatment and diagnostic. For example, clinical records, diagnostic imaging, histopathological slides need additional consideration in conjunction with molecular data like genomics, proteomics and transcriptomics. Given the independence with which these data sources are traditionally analyzed, they fail to represent cancer in all its glory across different biological scales.

This transformative multilevel learning paradigm is exemplified by MML, where information from multiple distinct sources can be integrated to build more complete and effective models for cancer diagnosis, prognosis, and treatment prediction. The integration of multimodal data enables researchers to leverage the strengths of each modality together for a richer characterization of tumor biology and patient outcomes. Recent advances in ML, specifically with Graph Neural Networks (GNNs) and Transformers, have transformed multimodal data processing and fundamentally improved precision oncology.

The chapter situates MML in oncology within the context of basic science collaboration: a joint vision for the use of DL presents different types of data modalities and fusion strategies, as well as state-of-the-art algorithms driving discovery research in cancer. By appreciating these core principles, we may begin to understand the promise of MML for personalized oncology, and ultimately better patient outcomes.

2.1 Data Modalities in Oncology

In oncology, data modalities refer to the different types of data generated from various sources that capture unique and complementary aspects of cancer (Sleeman IV et al., 2021). These data modalities span multiple biological and clinical scales, ranging from patient medical records to advanced molecular profiling and imaging techniques. The integration of these data is critical for developing a comprehensive understanding of cancer biology, improving diagnosis, and personalizing treatment strategies. The main data modalities in oncology can be classified into three broad categories: clinical data, imaging data, and molecular data (Figure 3).

In the LIHC approach, leveraging multimodal data will enable more accurate responses by combining insights from MRI, CT scans, medical notes, patient diagnoses, and genetic information.

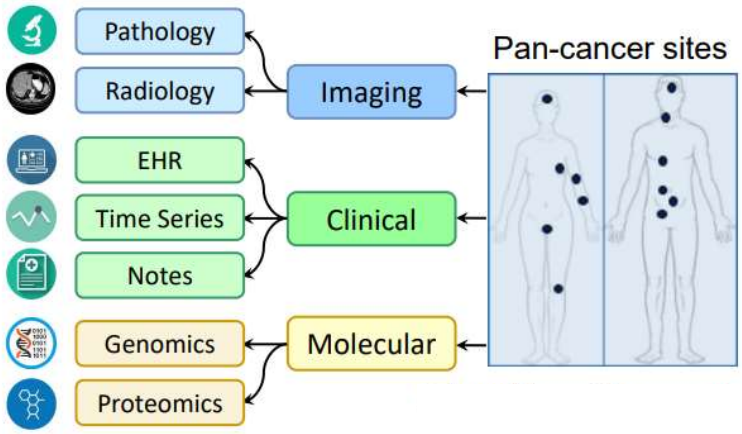


Figure 3 - Clinical modalities.

2.1.1 Clinical Data

Clinical data represents patient-specific information collected during routine medical care. This includes electronic health records (EHRs), which store a patient's medical history, including diagnoses, prescribed treatments, lab results, and demographic information. These records provide essential insights into a patient's overall health status, previous interventions, and disease progression over time. Clinical data may also include time-series information, such as repeated blood tests or vital sign measurements, which help monitor changes in the patient's condition (Asan et al., 2018). Time-series data is particularly valuable for tracking disease evolution and treatment response, offering a dynamic view of patient health (Quinn et al., 2019).

2.1.2 Imaging Data

Imaging plays a central role in oncology for diagnosis, staging, and treatment planning. Oncology imaging can be categorized into two main types: radiological imaging and histopathological imaging.

2.1.2.1 Radiological Imaging

Radiological imaging includes techniques like X-rays, CT, MRI, and positron emission tomography (PET). Radiological images provide detailed information about tumor size, shape, location, and metastasis in the body. They are vital for identifying the structural and functional abnormalities caused by cancer at the organ or tissue level. For example, PET scans can reveal metabolic activity in cancer cells, offering functional insights that complement the anatomical information provided by CT and MRI.

2.1.2.2 Histopathological Imaging

Histopathological imaging, also known as whole slide imaging (WSI), involves digitized images of tissue samples obtained from biopsies. Histopathology provides a microscopic view of the tissue architecture and cellular structures, allowing pathologists to assess the presence, type, and grade of cancer. Digitized pathology slides can be analyzed using advanced algorithms to detect patterns that are difficult for the human eye to discern, providing valuable information about tumor morphology and the tumor microenvironment (Rowe & Pomper, 2022; Waqas et al., 2023).

Each of these imaging techniques provides complementary information about the disease and combining them through multimodal analysis enhances the accuracy of cancer diagnosis and treatment planning.

2.1.3 Molecular Data

Molecular data modalities are perhaps the most granular and complex, capturing the biological and genetic underpinnings of cancer (J. Liu et al., 2021). These modalities include genomics, transcriptomics, proteomics, and epigenomics, collectively referred to as multi-omics data (Waqas et al., 2024).

2.1.3.1 Genomics

Genomics focuses on the study of the entire genome, identifying Deoxyribonucleic acid (DNA) mutations, copy number variations, and structural alterations that drive cancer progression (Boehm et al., 2022). Genomic sequencing, such as whole-genome sequencing (WGS) or whole-exome sequencing (WES), enables the detection of specific mutations associated with cancer.

2.1.3.2 Proteomics

Proteomics focuses on the protein expression profile within cancer cells. Since proteins are the functional molecules in biological systems, studying changes in protein abundance, modifications, and interactions can provide critical information about the biological processes altered in cancer. Advances in mass spectrometry and other proteomic techniques have made it possible to map protein networks that are disrupted in cancer cells (Boehm et al., 2022).

Each of these molecular data types provides a different layer of information about the biological state of cancer, and their integration through multimodal learning models enables a more comprehensive understanding of tumor behavior.

2.2 Data Fusion Strategies

In MML for oncology, data fusion is the process of integrating diverse data modalities such as clinical records, imaging, and molecular data into a unified framework that can provide a more comprehensive understanding of cancer. Effective data fusion strategies are crucial for leveraging the complementary strengths of different data sources and improving the accuracy of predictive models for diagnosis, prognosis, and treatment planning. Several data fusion strategies have been developed, each suited to different types of data and tasks. These strategies can be broadly classified into three categories: early fusion, intermediate fusion, and late fusion.

2.2.1 Early Fusion

Early fusion involves combining raw or extracted features from different modalities at the input level, before the learning process begins. In this approach, data from multiple modalities are concatenated or integrated into a single feature vector that is fed directly into the ML model (Sleeman IV et al., 2021). This strategy is particularly effective when the different modalities are closely related or when they provide complementary information that can be easily combined. (Figure 4).

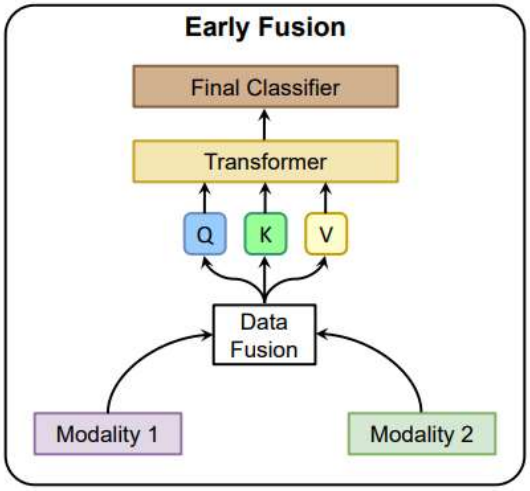


Figure 4 - Early fusion.

2.2.2 Late Fusion

Late fusion involves independent training models on each modality and combining their final predictions at the output stage. In this approach, each modality-specific model generates its own prediction, and these predictions are aggregated to make a final decision, often using techniques like majority voting, weighted averaging, or other decision-combination methods. (Figure 5).

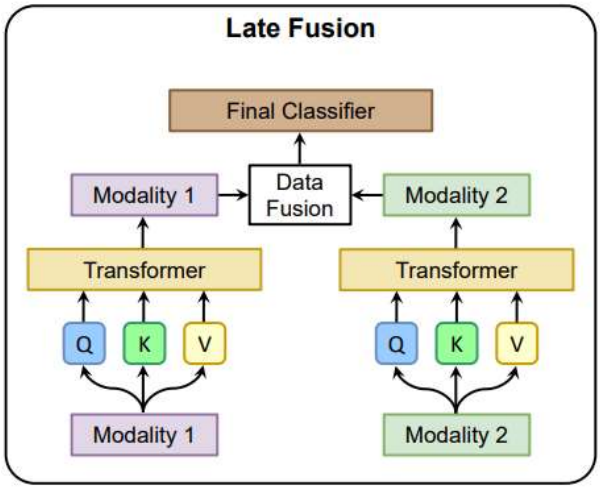


Figure 5 - Late fusion.

2.2.3 Cross-attention

Cross-attention involves attending to one sequence or set of data (the query) while using another sequence or set (the key and value) to provide contextual information (Figure 6). This mechanism enables the model to focus on relevant parts of one input based on the information provided by another input.

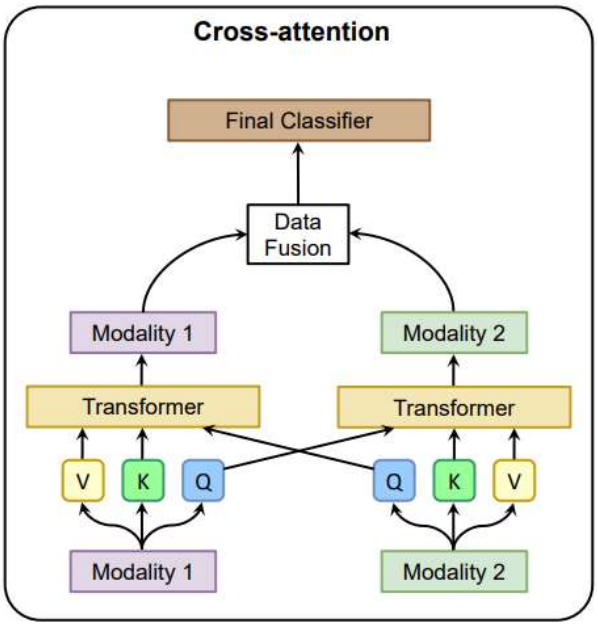


Figure 6 - Cross-attention.

2.3 MML for Oncology Datasets

2.3.1 Early Detection

The early detection of LIHC is crucial for improving patient outcomes, as treatments are more effective when the disease is identified at an initial stage. AI-powered diagnostic tools can significantly enhance early detection efforts. These tools employ ML algorithms to analyze medical images and other diagnostic data for early signs of LIHC. For instance, AI can detect minute changes in tissue structure and composition that may indicate the onset of cancer, even before symptoms appear. (Litjens et al., 2017) Additionally, AI can integrate data from multiple sources, such as imaging studies, blood tests, and genetic profiles, to provide a comprehensive assessment of cancer risk. This multi-modal approach increases the accuracy of early diagnosis, enabling timely interventions that can improve survival rates.

2.3.2 Personalized Therapy

Personalized therapy involves tailoring treatment strategies to the unique characteristics of each patient's cancer. This approach considers genetic, molecular, and clinical data to develop targeted therapies that are more effective and have fewer side effects than traditional treatments. AI and ML play a pivotal role in this process by analyzing large datasets to identify patterns and correlations that inform treatment decisions. For example, AI can help determine which patients are likely to respond to specific drugs based on their genetic profiles (X. Liu et al., 2021). This information can guide the selection of targeted therapies, immunotherapies, or combination treatments that are most likely to be effective for each patient. Personalized therapy not only improves treatment outcomes but also enhances the quality of life for patients by reducing the likelihood of adverse reactions and treatment failures.

2.3.3 Equitable Access

Ensuring equitable access to innovative therapies is essential for addressing disparities in healthcare. In Europe, the availability of advanced diagnostic and therapeutic tools can vary significantly between regions, leading to unequal health outcomes. AI-driven solutions can help bridge these gaps by standardizing care across different healthcare settings. For instance, AI-powered diagnostic tools can be deployed in remote or under-resourced areas to provide high-quality diagnostic services that might otherwise be unavailable. Telemedicine platforms enhanced with AI capabilities can facilitate consultations and treatment planning with specialists, regardless of geographic location. By democratizing access to cutting-edge healthcare technologies, AI can help ensure that all patients receive the best possible care, irrespective of where they live (Kalage et al., 2023).

2.3.4 Risk Factor Analysis

Understanding the risk factors associated with LIHC is key to developing preventive strategies and early intervention protocols. ML and NLP can be leveraged to analyze large datasets, including patient records, genetic data, lifestyle information, and environmental factors, to identify key risk factors for LIHC. These technologies can uncover complex relationships and interactions that contribute to cancer development, providing valuable insights for risk assessment. For example, AI can identify genetic mutations or biomarkers that are associated with an increased risk of LIHC, as well as

lifestyle factors such as diet, alcohol consumption, and exposure to certain chemicals. By identifying high-risk individuals, healthcare providers can implement targeted screening and preventive measures to reduce the incidence of LIHC (Radford et al., 2021).

Predicting how patients will respond to different therapeutic interventions is critical for optimizing treatment strategies. AI-based approaches can analyze vast amounts of clinical and molecular data to predict individual responses to various treatments. For instance, ML algorithms can identify biomarkers that indicate whether a patient is likely to respond to chemotherapy, targeted therapy, or immunotherapy. (Taori et al., 2020) This predictive capability allows clinicians to select the most appropriate treatment for each patient, reducing the trial-and-error approach often associated with cancer therapy (Choi et al., 2020). Furthermore, AI can continuously learn from new data, refining its predictions over time and improving the accuracy of treatment response assessments. This dynamic and adaptive approach ensures that patients receive the most effective therapies, minimizing the risk of adverse reactions and treatment failures.

2.3.5 Outcome Optimization

Optimizing patient outcomes involves continuously improving treatment strategies based on the latest data and insights. AI and ML play a crucial role in this process by providing real-time analysis and feedback on treatment efficacy (Radford et al., 2021). Predictive models can monitor patient progress and adjust treatment plans as needed, ensuring that interventions are responsive to the evolving nature of the disease. For example, ML algorithms can analyze longitudinal data to identify trends and patterns that indicate how a patient is responding to treatment. If a particular therapy does not produce the desired results, the model can suggest alternative approaches or adjustments to the current regimen. This iterative and data-driven approach enhances the precision and effectiveness of cancer treatment, ultimately leading to better patient outcomes. By leveraging NLP and ML, healthcare providers can ensure that treatment plans are continually optimized to meet the individual needs of each patient.

2.4 Graph Neural Networks in Multimodal Learning

GNNs have gained traction in multimodal learning, especially in fields like oncology, due to their ability to model complex relationships between diverse data types (Li et al., 2023). GNNs use graph structures to capture relationships between data points,

which can be incredibly beneficial when integrating multimodal data such as imaging, molecular, and clinical data in cancer research (Ektefaie et al., 2023).

2.4.1 The Graph Data

In GNNs, data is represented as a graph, $G = (V, E)$ where $V = \{v_1, v_2, \dots, v_n\}$ represents nodes, where node v has feature vector X_v, (data points or entities) and $E = \{(v_i, v_j) \mid v_i, v_j \in V\}$ represents edges (connections or relationships between data points). Each node is typically associated with a feature vector that represents specific data for that modality. For instance, in oncology, one node might represent a patient's genomic profile, while another node could represent an MRI image of a tumor, and the edges capture the interaction between these modalities.

2.4.1.1 Graph Types

Graphs in multimodal learning can take various forms: As illustrated in Figure 7, the common types of graphs include undirected, directed, homogeneous, heterogeneous, static, dynamic, unattributed, and attributed.

- **Undirected vs. Directed Graphs**: In undirected graphs (1), the relationships between nodes are symmetric (e.g., similarity between two images), while in directed graphs, relationships are asymmetric (e.g., the flow of information between medical records and image data).
- **Homogeneous vs. Heterogeneous Graphs**: Homogeneous graphs (2) have the same type of nodes, whereas heterogeneous graphs contain multiple node types (e.g., combining imaging data, clinical data, and molecular data in one graph) (Sun et al., 2022).
- **Static vs. Dynamic Graphs**: Static graphs (3) represent fixed relationships, while dynamic graphs evolve over time, capturing how relationships (such as tumor progression) change.
- **Attributed graphs**: Attributed graphs (4) play a crucial role by enhancing the expressiveness of the graph structure. Unlike basic graphs that only define relationships between nodes (via edges), attributed graphs go a step further by associating additional information (attributes) with both nodes and edges (Waikhom & Patgiri, 2022). This extra layer of information makes attributed

graphs highly relevant in complex tasks like multimodal learning, particularly in fields like oncology, where diverse data modalities (such as imaging, clinical records, and molecular data) need to be integrated.

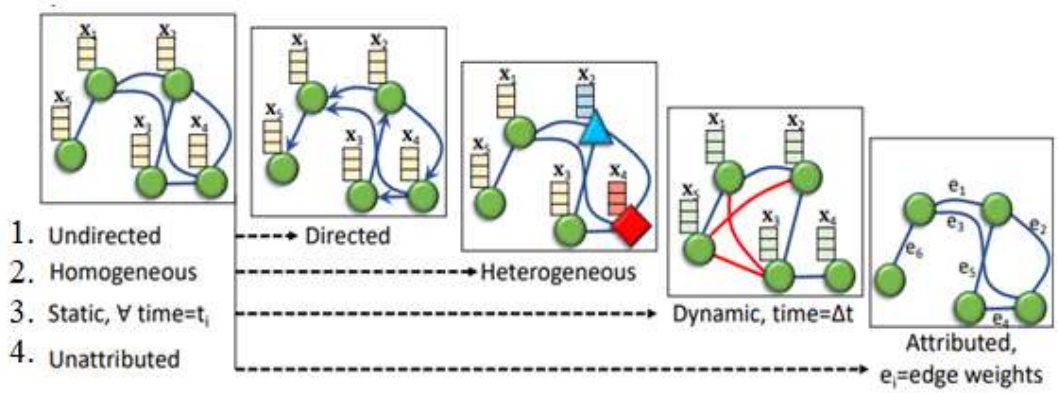


Figure 7 - Graph types.

2.4.1.2 Tasks for Graph Data

In Figure 8 we present three major types of tasks that can be defined on graph structures, each corresponding to a different level of analysis and prediction in graph neural networks. These tasks leverage the graph's inherent connectivity and structure to make more informed predictions, which are crucial for various domains, including healthcare and genomics.

1. **Node-level tasks**: These tasks focus on individual nodes within the graph. The goal is to classify or predict specific attributes of each node based on its features and the information available from its neighboring nodes. A typical application in healthcare might involve predicting the malignancy of a tumor by analyzing its image data. Here, the tumor (node) is connected to other relevant medical or biological entities (other nodes), and GNNs utilize this relational information to make predictions about the tumor's nature, such as whether it is cancerous or benign. Node-level tasks are essential for pinpointing specific details in a large network, such as identifying high-risk patients or abnormal cells.
2. **Edge-level tasks**: These tasks are concerned with the relationships or connections (edges) between pairs of nodes. The goal is to predict whether a certain connection

exists or to classify the type of relationship. For example, in the context of genomics, edge-level tasks could predict whether two genes interact based on their genomic features. Understanding these gene interactions is vital for identifying pathways involved in diseases or potential targets for drug development. Edge prediction tasks help unravel complex relationships within biological systems and networks, providing insights that are not always evident from analyzing individual nodes in isolation.

3. **Graph-level tasks**: These tasks involve making predictions or classifications for the entire graph as a whole. Rather than focusing on individual nodes or edges, the goal is to predict attributes of the entire structure. In medical applications, this could be used for patient classification, such as predicting patient survival rates or responses to specific treatments based on their complete medical profile, which can be represented as a graph. Each patient would be modeled as a graph, incorporating various interconnected data points like clinical metrics, imaging data, and genomic information. GNNs can then analyze the graph holistically to predict outcomes such as survival probabilities or likely treatment success. Graph-level tasks are particularly useful for applications where understanding the broader, systemic picture is essential for decision-making, such as in personalized medicine.

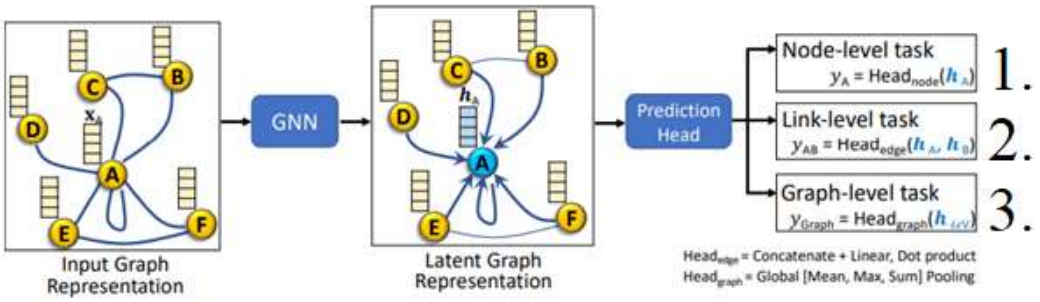


Figure 8 - Tasks for graph data.

By categorizing these tasks into node-level, edge-level, and graph-level, GNNs offer a versatile framework that can be applied to a wide range of problems in fields like oncology, genomics, and personalized healthcare, ultimately enabling more accurate predictions and better-informed clinical decisions.

2.4.2 ML for Graph Data

There are two main approaches to processing graph data in ML: traditional (shallow) methods and deep learning-based methods, as illustrated in Figure 9 (Wu et al., 2021).

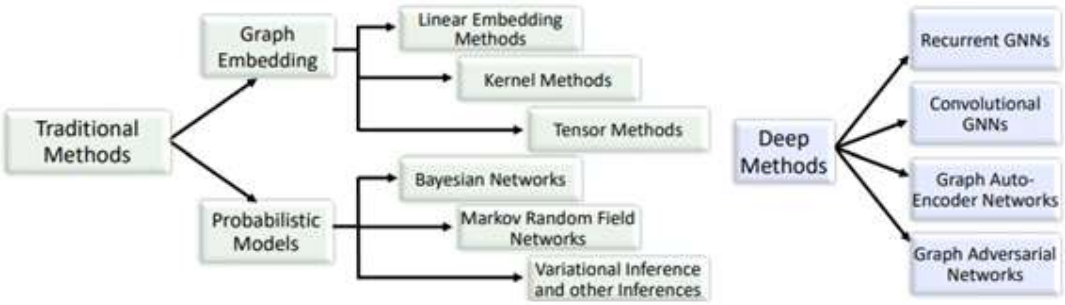


Figure 9 - Traditional methods and Deep Learning methods.

2.4.2.1 Traditional (Shallow) Methods

Shallow methods for learning graph representations involve techniques like graph embeddings, where nodes are represented in a low-dimensional space that captures their structural properties. Traditional ML techniques like Principal Component Analysis (PCA), linear models, or Bayesian networks can be used to analyze these embeddings. However, these methods have limitations when dealing with high-dimensional, complex, and heterogeneous multimodal data (Jiao et al., 2023).

2.4.2.2 DNN-based Methods - GNNs

GNNs are increasingly popular in the machine learning community due to their ability to handle non-Euclidean or relational datasets, which traditional deep learning methods like CNNs struggle with. GNNs aggregate information from neighboring nodes to update a node's representation, which enables them to encode higher-order structural information through message-passing. This process helps smooth features and provides a denoising mechanism.

GNNs can be divided into several categories based on their approach to information aggregation:

1. Recurrent GNNs (RecGNNs): These combine the structure of Recurrent Neural Networks (RNNs) with GNNs to capture both temporal dependencies

and local graph structures. RecGNNs repeat the information fusion process until equilibrium is reached.

2. Convolutional GNNs (ConvGNNs): These perform convolution operations on graphs, aggregating neighboring nodes' embeddings. ConvGNNs can be spatial, using weighted averages of neighboring vertices (e.g., GraphSAGE, GAT), or spectral, operating on the eigenvalues of the graph Laplacian matrix (e.g., ChebNet, GCN).
3. Graph Auto-Encoder Networks (GAEs): GAEs are unsupervised models used for tasks such as dimensionality reduction, anomaly detection, and graph generation. They map graphs into a lower-dimensional latent space through an encoder and reconstruct the graph through a decoder.
4. Graph Adversarial Networks (GraphANs): These networks, inspired by Generative Adversarial Networks (GANs), are designed to generate new graphs with similar properties to the input data. A generator creates new graphs, while a discriminator distinguishes between real and generated ones.

Other variations of GNNs include scalable, dynamic, hypergraph, and heterogeneous GNNs, each designed to address specific challenges in handling graph-structured data.

2.4.3 GNNs and ML using Unimodal Oncology Datasets

GNNs have been applied effectively in oncology to process unimodal datasets such as pathology, radiology, and molecular data. While traditional methods often fail to capture the complex interactions and dependencies between different types of cancer data, GNNs excel at modeling these relationships by representing data as graphs, making them a powerful tool in oncology applications (Ahmedt-Aristizabal et al., 2022).

2.4.3.1 Pathology Datasets

Pathology data, particularly histopathology images, provide detailed insights into the microscopic structure of tissues, revealing essential information about cancer. Traditionally, Convolutional Neural Networks (CNN) have been employed for pathology image analysis, but they are limited in capturing the structural relationships between cells or tissue regions (Rj et al., 2022). GNNs overcome these limitations by modeling histopathology data as graphs where:

- **Nodes**: Represent individual cells, tissue patches, or specific regions of interest (ROIs) within the histopathology slide.
- **Edges**: Represent spatial or functional relationships between the nodes, such as cell-to-cell interactions or shared tissue features.

In cancer diagnosis tasks, GNNs can represent tumor tissues as cell-graphs, where each node represents a cell, and edges represent their spatial proximity or interaction (Figure 10). The histograph model is an example that uses GNNs to classify cancerous versus non-cancerous regions in breast cancer histology data (Singh et al., 2021). By integrating both cellular features and spatial relationships, GNNs enable a richer analysis of tumor microenvironments and cellular heterogeneity, which is crucial for precise diagnosis and prognosis.

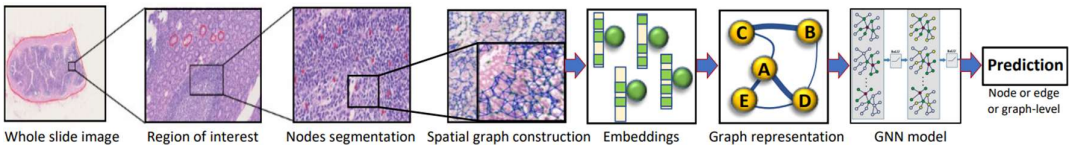


Figure 10 - Pathology datasets represent tumor tissues as cell-graphs.

2.4.3.2 Radiology Datasets

Radiology data, such as X-rays, MRI, CT, and PET scans, are critical in detecting and diagnosing various types of cancer (Zhu et al., 2024). In radiology datasets, GNNs can model interactions within or across different imaging modalities by treating individual image regions or anatomical structures as nodes (Figure 11):

- **Nodes**: Represent segmented image patches or ROIs from the radiological scan.
- **Edges**: Capture relationships such as proximity or anatomical connections (e.g., lymph nodes, blood vessels).

A common application of GNNs in radiology datasets is in tumor segmentation and classification. For instance, GNNs can be applied to MRI or CT scans to segment out tumors and then analyze their structure, connections to surrounding tissues, and overall progression (Z. Tian et al., 2020).

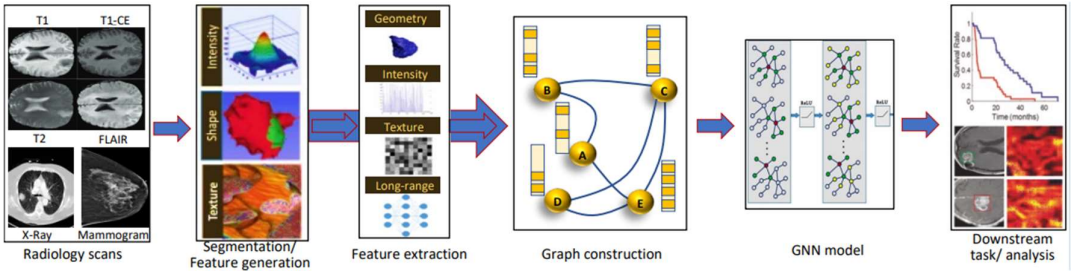


Figure 11 - Radiology datasets can model interactions within or across different imaging.

2.4.3.3 Molecular Datasets

Molecular data, such as genomics, transcriptomics, and proteomics, offer detailed information on the genetic and molecular underpinnings of cancer (Wang et al., 2021). GNNs can model molecular interactions by treating molecules or genes as nodes and their relationships as edges (Figure 12).

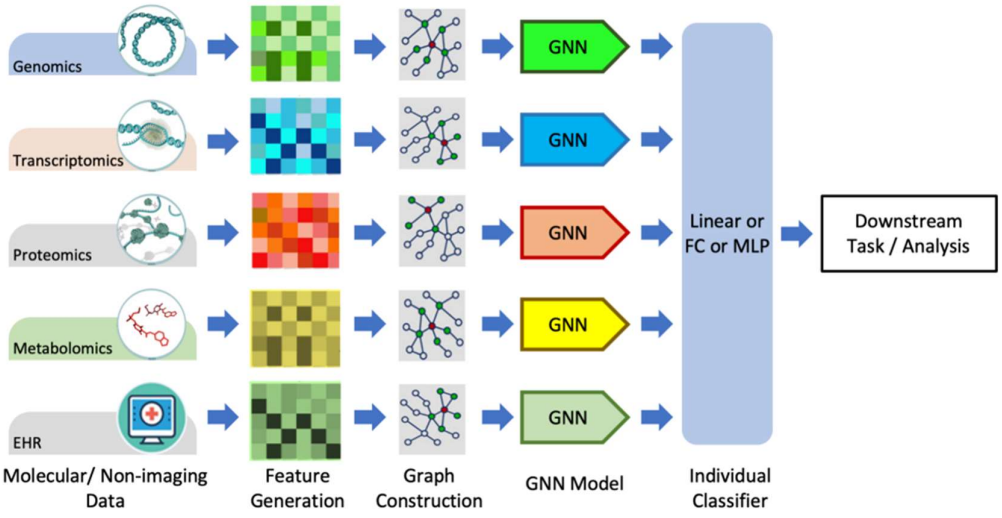


Figure 12 - Molecular datasets can model molecular interaction.

2.5 Transformers in MML

Transformers are a type of neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al. in 2017. They have revolutionized the field of NLP and have been extended to various other domains such as computer vision and audio processing. Transformers are known for their ability to handle sequential data

and capture long-range dependencies effectively, making them particularly powerful for tasks like machine translation, text generation, and more.

Transformers are attention-based deep neural networks initially designed for NLP (Vaswani et al., 2023), utilizing self-attention mechanisms to process data in parallel. They handle sequential data and capture long-range dependencies, making them suitable for tasks like language translation and modeling. Unlike recurrent neural networks (RNN) and CNNs, Transformers weigh the importance of input tokens, processing sequences of any length and capturing dependencies across various data types, including images, audio, and time-series. Pre-trained on large datasets and fine-tuned for specific tasks, they have led to models like BERT, GPT, and CLIP.

2.5.1 Model Architecture

A transformer is an encoder-decoder model that uses the attention mechanism. It can take advantage of parallelization and also process a large amount of data at the same time because of its model architecture (Figure 13) (Vaswani et al., 2023) .

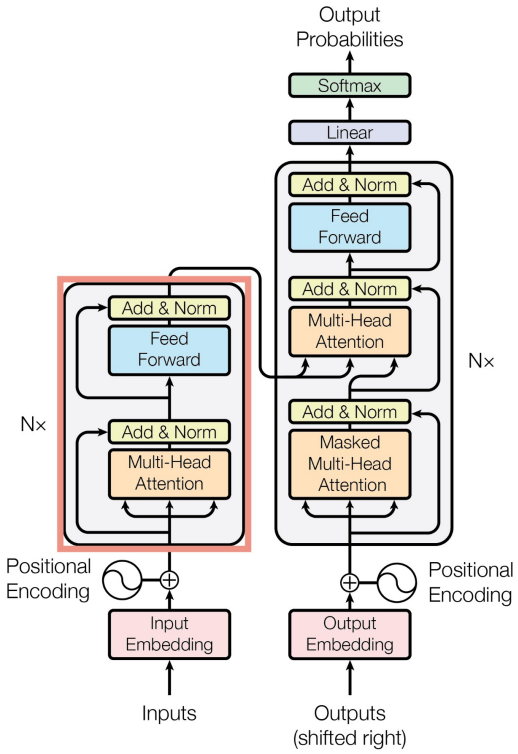


Figure 13 - The original Transformer architecture.

In the Figure 14, the transformer model consists of an encoder and a decoder. The encoder processes the input sequence and passes its representation to the decoder, which then decodes it for the relevant task (Vaswani et al., 2023).

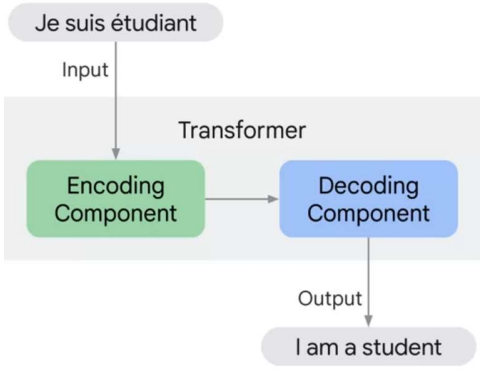


Figure 14 - Simple example of the encoder and decoder.

The encoders are all identical in structure but have different weights. Each encoder consists of two sub-layers (Figure 15).

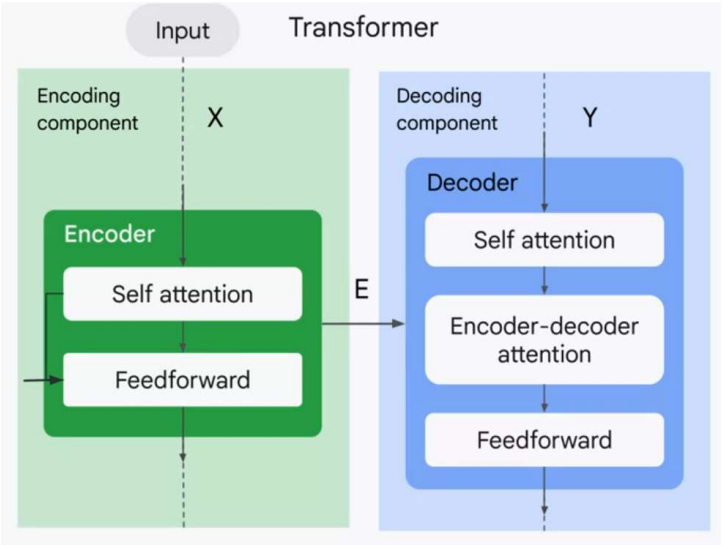


Figure 15 - Encode and decode process.

The first sub-layer is the self-attention mechanism, where the input first passes through the self-attention layer, enabling the encoder to focus on relevant parts of the surrounding words as it encodes the central word in the input sentence. The second sub-layer is a feedforward neural network, which processes the output from the self-attention

layer. The same feedforward network is applied independently to each position. The decoder also includes both the self-attention and feedforward layers, but an additional encoder-decoder attention layer is placed between them. This layer helps the decoder focus on relevant parts of the input sentence (Shi et al., 2021).

After embedding the words in the input sequence, each of the embedding vector flows through the two layers of the encoder. The word at each position passes through a self-attention process, then it passes through a feedforward neural network, the exact same network with each vector flowing through it separately. Dependencies exist between these paths in the self-attention layer. However, the feedforward layer does not have these dependencies. Therefore, various paths can be executed in parallel while they flow through the feedforward layer.

The self-attention mechanism in Transformers treats input data as a fully connected graph, making it ideal for processing various data modalities. This mechanism allows the model to learn and attend to global patterns in the data by focusing on different input tokens (or embeddings), which can represent different types of information such as images, text, or genomic data (Boehm et al., 2022). The input embedding is broken up into query, key, and value vectors. These vectors are computed using weights that the transformer learns during the training process (Xu et al., 2023). All of these computations happen in parallel in the model in the form of matrix computations (Figure 16).

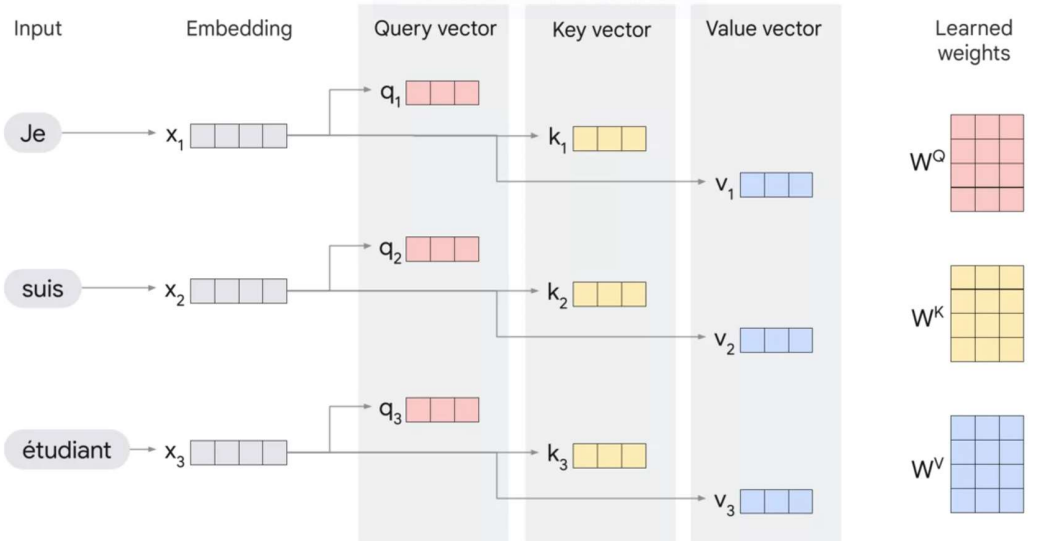


Figure 16 - The input embedding is broken up into query, key, and value vectors.

Once the query, key, and value vectors, the next step is to multiply each value vector by the softmax score in preparation to sum them up. The intuition here is to keep intact the values of the words you want to focus on and leave out irrelevant words by multiplying them by tiny numbers.

3 Methodology

AutoGluon-Multimodal (AutoMM) is an open-source Automated Machine Learning (AutoML) library made for learning from different types of data. AutoMM works with images, text, and tables, either on their own or together. It contains many features, such as classification, regression, object detection, semantic matching, and image segmentation.

3.1 Introduction

AutoML aims to make it easier to turn raw data into accurate predictions with less need for human expertise, though some gaps remain in certain areas (Zöller & Huber, 2021). AutoML frameworks bring together key steps in ML, like data preprocessing (Gada et al., 2021), model selection (Arango et al., 2024), training, and deployment (Paleyes et al., 2022) making it accessible to both technical and non-technical users. This helps users handle a wide variety of tasks without needing deep knowledge of ML.

The rapid growth of ML, especially with the introduction of foundation models trained on large datasets, has changed fields like computer vision and NLP (Bommasani et al., 2022). Fine-tuning these models for specific tasks is crucial, but there are few open-source AutoML tools to support this. Most existing AutoML tools focus on simple tasks like classification and regression using tabular data (Thornton et al., 2013), overlooking the more complex real-world challenges that often involve combining different types of data (e.g., images, text, and tables) (Baltrušaitis et al., 2019). While there are tools designed for specific tasks, the AutoML community still lacks a unified system that can handle diverse data types and tasks smoothly.

To address these challenges, AutoMM is a Python-based, open-source AutoML framework designed for machine learning with foundation models. As part of the AutoGluon ecosystem (Shchur et al., 2023), AutoMM simplifies fine-tuning foundation models on domain-specific data with just three lines of code. It leverages popular model repositories like Huggingface/transformers (Wolf et al., 2020), TIMM and MMDetection (Chen et al., 2019), supporting various data types such as text, images, and tables. This enables tasks like classification, regression, object detection (Zou et al., 2023), semantic matching (Reimers & Gurevych, 2019), and image segmentation (Minaee et al., 2020). Figure 17 illustrates the main features of the AutoMM framework.

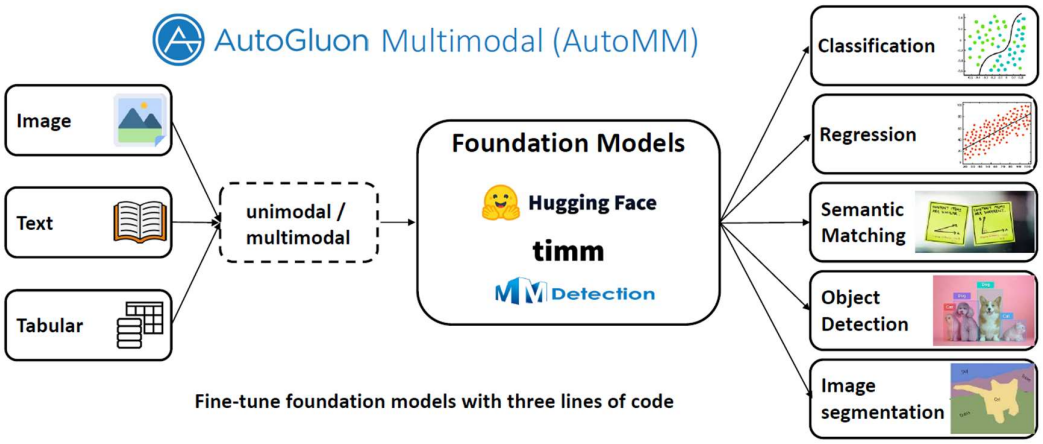


Figure 17 - AutoMM framework: supports both single and multiple types of data (left), allowing easy fine-tuning of foundation models (middle) for basic tasks like classification/regression and more advanced tasks (right).

Evaluating AutoMM presents two challenges: the absence of well-established benchmark datasets that span multiple data types and tasks, and the limited availability of AutoML libraries with comparable functionality.

3.2 AutoMM

Supporting diverse data types and tasks within an AutoML framework is challenging. The primary difficulty lies in automating these processes using a unified data format, APIs, model architecture, and training workflow. In this section, we outline the design and key features of AutoMM.

3.2.1 Data Format and Processing

AutoMM uses Pandas DataFrames to manage various data types, such as images, text, and tables (containing numeric or categorical data). DataFrames are widely used in data analysis due to their flexibility and ease of use, organizing data in a 2D table where rows can include text or image paths. This structure allows AutoMM to seamlessly handle mixed data types within a single DataFrame, even when multiple columns represent different aspects of the same dataset. Additionally, AutoMM can process raw or messy data, eliminating the need for users to perform extensive preprocessing.

Typically, handling different data types requires separate processing pipelines, resulting in redundant code and increased maintenance efforts. To address this, AutoMM uses a unified data pipeline that incorporates modules such as LightningDataModule,

torch DataLoader (Paszke et al., 2019), and a modality detector. The LightningDataModule generates data loaders for training, validation, and prediction, while the torch DataLoader processes the data using torch Dataset, based on the detected data types (Figure 18). This unified approach streamlines data handling across diverse data types.

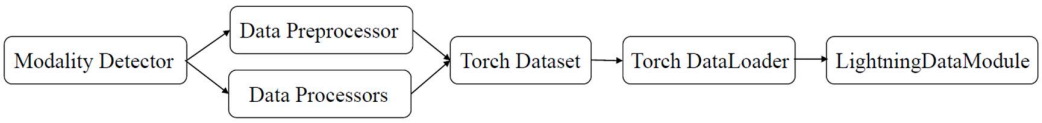


Figure 18 - The interdependence among data modules.

The data preprocessor manages general tasks such as filtering out irrelevant features, handling null values, and normalizing numerical data. Data processors are tailored to specific models and perform tasks like data augmentation and transforming individual samples during the loading process. Each model has its own set of individual samples and features, with each cell capable of storing different types of data, including numerical values and categorical variables. A data collator groups the processed samples into mini-batches for the model.

Data processors also handle specific tasks: numerical data is combined into a vector, categorical data is encoded, and images are augmented, converted into tensors, and normalized. If a sample contains multiple images, AutoMM processes each image separately before stacking them (Müller & Hutter, 2021). For text data, tokenization is performed, and the tokens from each field are combined. If the token sequence exceeds the model's limit, it is shortened by removing tokens from the longest text field.

3.2.2 APIs

AutoMM simplifies the fine-tuning of foundation models on single or multiple data types through an intuitive API. For instance, by using a DataFrame in "train.csv" with a label column named "label," users can train a model with just three lines of code:

```
1 from autogluon.multimodal import MultiModalPredictor
2 predictor = MultiModalPredictor(label="label")
3 predictor.fit("train.csv")
```

During training, AutoMM automatically detects the problem type (e.g., binary classification, multiclass classification, or regression), splits the data into training and validation sets, identifies data types, selects appropriate models, and fine-tunes them. The *fit()* method also offers customization options, allowing users to adjust training duration or hyperparameters, and supports continuous training by calling *fit()* again with new data.

After training, various APIs are available for evaluation and making predictions:

```
score = predictor.evaluate("test.csv")
predictions = predictor.predict("new.csv")
probabilities = predictor.predict_proba("new.csv")
embeddings = predictor.extract_embedding("new.csv")
```

These APIs facilitate model evaluation, predictions, probability estimates (for classification), and feature embedding extraction. The prediction-related APIs do not require labels, making them ideal for deployment scenarios.

Additionally, one can save and reload models using:

```
predictor.save("save_path")
predictor = MultiModalPredictor.load("save_path")
```

The *load()* method allows one to resume interrupted training by specifying *resume=True* and calling *fit()* again.

3.2.3 Models

AutoMM is designed to apply foundation models across diverse real-world tasks. These models are initially pretrained on extensive datasets and then fine-tuned on smaller, task-specific datasets. Fine-tuning performs best when the data types used in pretraining align with those required for the task. For instance, BERT can be effectively fine-tuned for text-based tasks but cannot handle tasks that combine text with tables (Devlin et al., 2019). Most foundation models are specialized for a single data type—such as images (Dosovitskiy et al., 2021), text (He et al., 2023), or both (Radford et al., 2021)—whereas many real-world tasks involve a mix of data types, like images, text, and tables. To address this, AutoMM employs a late-fusion architecture (Figure 19) that processes each data type (image, text, or table) independently and then merges the results. If only one data type is present, the merging step is bypassed. This architecture also allows new data types to be added and integrated seamlessly with existing ones.

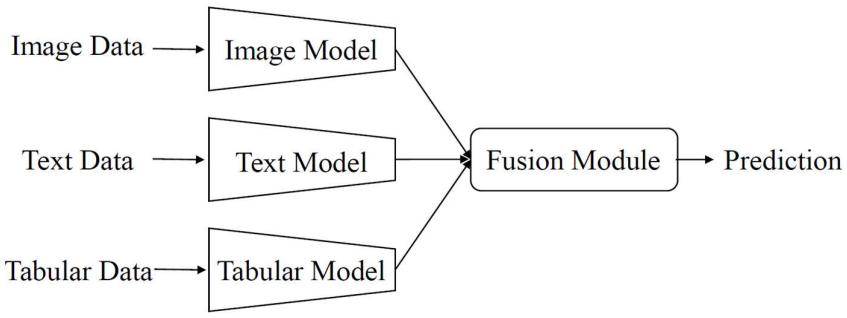


Figure 19 - Late-fusion model.

AutoMM supports several popular model libraries, including Huggingface, TIMM, and MMDetection, which offer a broad selection of pretrained models. These libraries provide models tailored for tasks involving text, images, and object detection—such as Electra and DeBERTa for text or Swin Transformer and ViT for images. Models are available in various sizes to help balance performance and speed. To simplify model selection, AutoMM offers three preset quality levels—best, high, and medium—determined through internal testing.

For fundamental tasks like classification or regression, AutoMM selects the top five models for each data type based on performance in single-modality evaluations. It then identifies the optimal combination of models and configurations for multimodal tasks, setting this combination as the default model.

3.2.4 Training

AutoMM trains a single late-fusion model with fixed, pre-determined hyperparameters (e.g., learning rate and weight decay) optimized offline across multiple benchmark datasets. Unlike some AutoML tools, such as Autosklearn 2.0 (Feurer et al., 2022), which emphasize continuous hyperparameter optimization or model ensembling (Erickson et al., 2020), AutoMM prioritizes offline tuning to avoid the high resource demands often associated with fine-tuning large multimodal models.

AutoMM ensures efficient training by leveraging the PyTorch-based Lightning framework, which streamlines the process with modular components like LightningDataModule (for data handling), LightningModule (for model definition), and Trainer (for process orchestration). Lightning's Callbacks further support logging, checkpointing, and early stopping, enhancing the training workflow.

To make fine-tuning large foundation models accessible, particularly for users with limited computational resources, AutoMM employs parameter-efficient fine-tuning techniques. By optimizing only a small subset of the model, these techniques significantly reduce memory and time requirements.

3.2.5 Deployment

Efficient model deployment is essential for real-world applications. To facilitate smooth offline predictor loading, AutoMM pre-saves required artifacts, such as Huggingface model configurations, to avoid errors when Internet access is unavailable. In production, low inference latency is crucial, particularly for online inference. While AutoMM leverages Lightning modules for training on large datasets, these can introduce latency with smaller batches. To improve inference speed, AutoMM provides a "realtime" option that switches to plain PyTorch models and processing.

AutoMM also integrates NVIDIA TensorRT, enabling optimized, low-latency, and high-throughput inference (R. Tian et al., 2022). Additionally, it manages inconsistencies in image formats between training and deployment by dynamically detecting image sub-types during inference, simplifying deployment across varied environments.

3.2.6 Advanced Tasks

AutoMM extends traditional AutoML beyond basic classification and regression to support advanced tasks like semantic matching, object detection, and semantic segmentation. Semantic matching evaluates the similarity between two items (e.g., images, texts, or an image-text pair) using a bi-encoder design to compare their embeddings (Figure 20). Object detection identifies and locates objects in images, while semantic segmentation assigns a category to each pixel in an image. Object detection outputs class labels and bounding boxes, while segmentation produces masks.

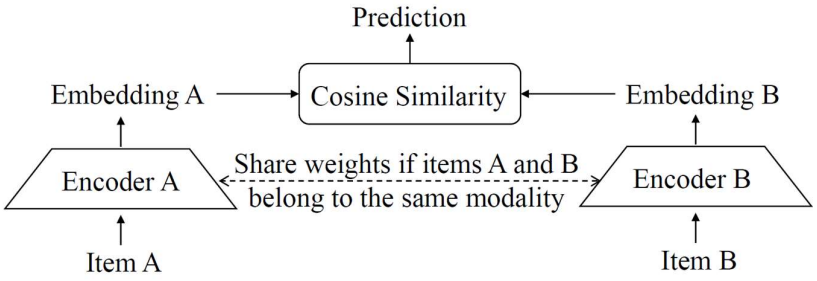


Figure 20 - Bi-encoder model.

Despite the complexity, users can execute these tasks with just three lines of code, specifying the problem type when initializing the predictor. AutoMM achieves this by fine-tuning foundation models from Huggingface and MMDetection, using models like CLIP (Radford et al., 2021) for image-text matching, DINO (Zhang et al., 2022) for object detection, and the Segment Anything Model (Kirillov et al., 2023) for segmentation.

3.3 Metrics

Metrics are selected based on the specific task. For regression tasks with continuous labels, the R-squared (R^2) metric is used to assess how well the model's predictions align with the observed data compared to the mean prediction:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

here y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

For binary classification tasks the $F1$ score is used, which represents the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

where

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

and

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

For multiclass classification tasks, the weighted $F1$ score ($F1_{weighted}$) is used, which calculates a weighted average of the F1 scores for each class to account for label imbalance:

$$F1_{weighted} = \sum_{i=1}^{C} w_i \, F1_i$$

Here, C represents the number of classes, $F1_i$ is the $F1$ score for class i, and w_i is the proportion of true instances for class i in the dataset.

All reported metrics are calculated on a holdout test set that is not used for training or hyperparameter tuning. Some datasets include a predefined test set, while others do not. For datasets without a test set, we randomly allocate 20% of the training data as the test set.

4 Empirical Study

In this chapter, an analysis of the retrieved information on Liver Hepatocellular Carcinoma (LIHC) is presented, along with a discussion of the study's results. The first step was to plan the type of analysis required to predict and develop insights into LIHC data, laying the groundwork for effective research. This process is fundamental to ensure accurate data processing, analysis, and meaningful results. The following sections detail each step taken in the research process.

The dataset used in this study was sourced from the American National Cancer Institute (NCI), ensuring both reliability and relevance to liver hepatocellular carcinoma research. The NCI was selected as the primary source due to its extensive and comprehensive data on various cancer types, including LIHC. Navigating the NCI website was essential for locating the specific data required for this analysis. This process involved identifying and accessing LIHC-related datasets specifically curated for research purposes, which facilitated a smooth data retrieval experience.

The extraction of image data from the NCI's database was a structured process, involving the downloading of raw LIHC image files. Data extraction protocols were carefully followed to maintain data integrity and ensure completeness, allowing for robust analysis in later stages. Once accessed, the LIHC image dataset underwent preliminary checks to ensure data quality. This data extraction process was vital for isolating relevant data attributes and filtering out extraneous information, setting the stage for the data preparation phase.

Data preparation involved transforming the raw LIHC data into a usable format. This step was essential to ensure compatibility with analytical tools and models used later in the study. To enable image-based analysis, the single nucleotide substitution (SNS) files were converted into PNG format. This transformation was necessary to integrate the data with machine learning models designed for visual data input, thereby supporting advanced predictive analysis.

Data identification entailed the categorization of various data types within the LIHC dataset. This classification helped streamline the analysis by isolating critical data subsets and identifying patterns that contribute to the development of the predictive model. Data cleanup was conducted to address any inconsistencies, missing values, and outliers within the dataset. This step improved data quality and enhanced the reliability of subsequent

analyses. Proper data cleanup ensured that the model's accuracy would not be compromised by noise or irrelevant data.

The proposed framework for the analysis was designed to facilitate an effective approach to predicting and understanding LIHC progression. This framework integrated machine learning algorithms and image-based data interpretation to produce a scalable solution that could be applied to LIHC data for enhanced insights and predictive modeling. Additionally, clinical data and notes, both textual and tabular, were gathered from Hugging Face Datasets[1], adding another layer of valuable information for the analysis.

The results analysis presents an in-depth evaluation of the outcomes derived from the proposed framework. Through this analysis, key insights into the nature of LIHC were uncovered, allowing for a better understanding of its progression and potential markers for early detection. The effectiveness of the framework was assessed based on prediction accuracy, relevance to clinical applications, and potential implications for future research on LIHC.

4.1 Dataset

As mentioned the dataset used in this study was sourced from a reliable provider, the American National Cancer Institute. This NCI website offers access to clinical data, medical notes, and imaging resources essential for LIHC research. The NCI provided the primary dataset used in this study. This source was chosen due to its extensive and comprehensive data on various cancer types, including liver hepatocellular carcinoma.

The Cancer Genome Atlas (TCGA) Multimodal Dataset also available in Hugging Face is a comprehensive collection of clinical data, pathology reports, and slide images for cancer patients. This dataset aims to facilitate research in multimodal machine learning for oncology by providing embeddings generated using state-of-the-art models such as GatorTron and UNI. This source was chosen to extract information about LIHC in the format of text and tabular.

[1] https://huggingface.co/datasets/Lab-Rasool/TCGA/tree/main

4.1.1 Extract the LIHC Images

The extraction of data from the NCI's database was a structured process, involving the downloading of raw LIHC imaging data files. Data extraction protocols were carefully followed to maintain data integrity and ensure completeness, allowing for robust analysis in later stages.

The NCI website initial page[2] contains all the information related to the multiple cancer types (Figure 21).

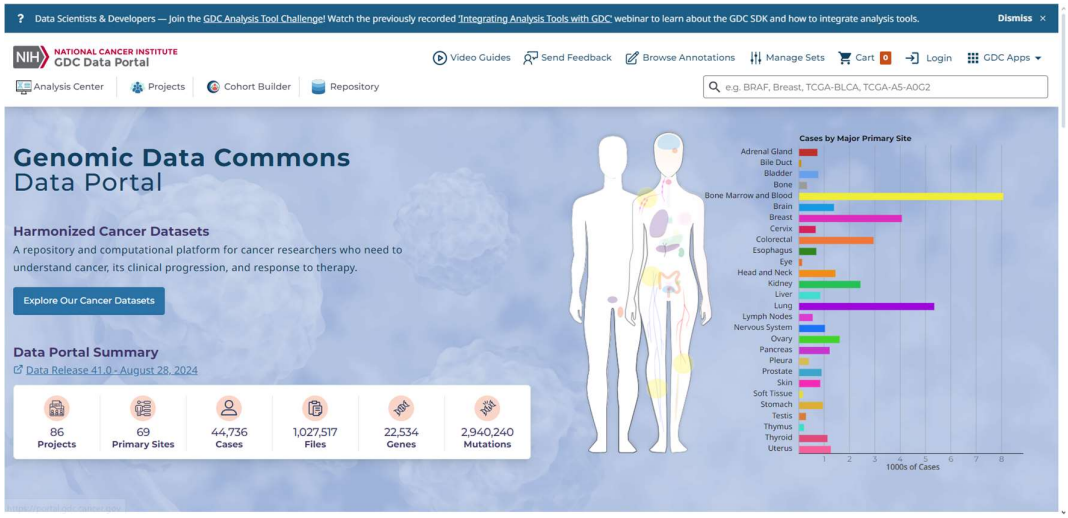


Figure 21 - National Cancer Institute GDC data portal main page.

Selecting the *Projects* option on the top left of the Figure 21 opens a new webpage with all projects available. For this study the "*TCGA-LIHC*" project was selected. This opens a new window with multiple data categories, including clinical data, medical notes, and imaging resources, essential for a multifaceted approach to LIHC research.

Once the relevant data was located, each file was reviewed for completeness and compatibility with the study's analytical framework. The availably data comprises 377 cases, but only 360 are public to use. The data files were downloaded, in high-resolution images. During this process, specific extraction protocols were followed to maintain data integrity and avoid any loss of critical information.

[2] https://portal.gdc.cancer.gov

4.1.2 Extract the LIHC Text and Tabular

The extraction of data from the TCGA database was a structured process that involved downloading transformed liver hepatocellular carcinoma data files. TCGA data is accessible through Hugging Face Datasets, which includes information on various cancer types, including LIHC. Specifically, the Clinical Data dataset (gatortron-base dataset) and the Pathology Report dataset (gatortron-base dataset) were retrieved for tabular and text data, respectively.

4.2 Proposed Framework

In this study, a robust framework was developed utilizing the AutoMM model to predict the vital status of patients with liver hepatocellular carcinoma. The vital status, categorized as either 'Death' or 'Alive', serves as our target variable. The dataset consists of 360 rows, with 80% allocated for training and 20% for testing the model's performance.

4.2.1 Data preparation

Data preparation involved transforming the raw LIHC data into a usable format. This step was essential to ensure compatibility with analytical tools and models used later in the study.

4.2.1.1 Convert the SNS files into PNG

To facilitate our analysis using AutoMM, it was necessary to convert the clinical images of SNS to JPEG format (Figure 22). Converting SNS data to JPEG format ensures that the image data is standardized and consistent. This consistency is important for training machine learning models, as it eliminates variability in image formats that could complicate the analysis process. JPEG format, while lossy, is sufficient for the majority of machine learning tasks and provides a good balance between file size and image quality.

JPEG files are optimized for both storage and processing efficiency. They offer a high level of compression, which reduces file sizes significantly compared to other formats like PNG. This efficient data processing is particularly beneficial when working with large datasets, as it reduces the computational load and speeds up the analysis.

Converting the clinical images of SNS to JPEG format was essential for ensuring compatibility with AutoMM, standardizing the data, and enhancing processing efficiency.

This conversion allowed us to fully leverage AutoMM's image-based machine learning capabilities, which was crucial for obtaining accurate and reliable results in our LIHC data analysis.

Figure 22 - Clinical image in PNG format.

4.2.1.2 Data Identification

Table 1 presents the set of tabular clinical data that was exported for the study. This dataset includes various patient and clinical characteristics essential for our analysis, providing valuable insights into the clinical and pathological aspects of the disease. The detailed fields captured in this dataset enable a thorough examination of factors influencing patient outcomes and treatment efficacy, laying the groundwork for our research.

Table 1 - Tabular clinical data.

Tabular Clinical Data
project_id
case_id
case_submitter_id
selected_normal_wgs
source_center
state
a260_a280_ratio
normal_tumor_genotype_snp_match
ribosomal_rna_28s_16s_ratio
rna_integrity_number
age_at_index
days_to_birth
days_to_death
ethnicity
gender

race
vital_status
year_of_birth
year_of_death
age_at_diagnosis
ajcc_clinical_m
ajcc_pathologic_n
ajcc_pathologic_stage
ajcc_pathologic_t
classification_of_tumor
days_to_diagnosis
icd_10_code
last_known_disease_status
morphology
primary_diagnosis
prior_malignancy
prior_treatment
progression_or_recurrence
site_of_resection_or_biopsy
synchronous_malignancy
tissue_or_organ_of_origin
tumor_grade
year_of_diagnosis
treatment_or_therapy
alcohol_history
is_ffpe
composition
days_to_collection
days_to_sample_procurement
initial_weight
pathology_report_uuid
portion_submitter_id
percent_lymphocyte_infiltration
percent_monocyte_infiltration
percent_necrosis
percent_neutrophil_infiltration
percent_normal_cells
percent_stromal_cells
percent_tumor_cells
percent_tumor_nuclei
analyte_type
days_to_last_follow_up
weight
sample_id
sample_submitter_id
portion_id

slide_id
slide_submitter_id
section_location
concentration
spectrophotometer_method
oct_embedded
preservation_method
no_matched_normal_low_pass_wgs
no_matched_normal_targeted_sequencing
no_matched_normal_wgs
no_matched_normal_wxs
selected_normal_low_pass_wgs
selected_normal_targeted_sequencing
selected_normal_wxs
creation_datetime
portion_number
specimen_type
ajcc_pathologic_m
ajcc_staging_system_edition
intermediate_dimension
longest_dimension
shortest_dimension
cigarettes_per_day
pack_years_smoked
years_smoked
analyte_type_id
figo_stage
analyte_id
analyte_submitter_id
aliquot_id
aliquot_submitter_id
sample_type
sample_type_id
tissue_type
tumor_descriptor
ajcc_clinical_t
ann_arbor_b_symptoms
ann_arbor_clinical_stage
ann_arbor_extranodal_involvement
aliquot_quantity
aliquot_volume
experimental_protocol_type
ajcc_clinical_n
ajcc_clinical_stage
primary_gleason_grade
secondary_gleason_grade

igcccg_stage
analyte_volume
masaoka_stage

Table 2 presents the exported text report data collected for this study, providing critical insights into the clinical documentation associated with LIHC patients. This data includes various attributes pertinent to the text reports, offering insights into the structure and content of the clinical documentation. Key fields in this dataset encompass identifiers, data formats, and details about the report text itself, enabling a comprehensive analysis of the clinical narratives associated with each case.

Table 2 - Text report data.

Text Report Data
id
data_format
access
file_name
md5sum
updated_datetime
data_type
data_category
state
type
file_size
created_datetime
PatientID
gdc_case_id
report_text

4.2.1.3 Data Cleanup

During the data preparation phase, we deleted specific columns in the dataset, such as IDs, to enhance the efficiency and effectiveness of our analysis. While these ID columns serve to uniquely identify records, they do not provide useful information for predicting outcomes. Including them could introduce noise and lead the model to learn from these identifiers instead of the actual data patterns. By removing unnecessary columns, we also reduce the dataset's size, making the training process more efficient and enabling the model to concentrate on the most relevant features, which improves its ability to generalize to new data.

Furthermore, eliminating irrelevant columns enhances the model's performance by ensuring that it trains on data that directly influences the target variable, resulting in more accurate and reliable predictions. A simpler dataset is easier to process and manage, as removing non-informative columns reduces the complexity of data transformations and preprocessing steps. By focusing on significant features that impact the target variable, we can develop a robust predictive model

4.2.1.4 Data Split

To ensure robust training and evaluation of our model, we carefully split our dataset of 360 entries into distinct training and testing sets. This splitting process is crucial for developing a model that generalizes well to unseen data. We allocated 80% of the dataset, which amounted to 288 entries, to the training set. This subset was used to train the AutoMM model, allowing it to learn patterns and relationships within the data.

The remaining 20% of the dataset, comprising 72 entries, was reserved for testing. This set was used to evaluate the model's performance, ensuring that the model's predictions could be validated against unseen data. By splitting the dataset in this manner, we were able to create a reliable framework for training and assessing the model, ultimately leading to more accurate and generalizable predictions of the vital status of LIHC patients.

4.2.1.5 Label Encoding

To prepare the target variable, vital status, for use in the AutoMM model, we implemented label encoding. Label encoding is a preprocessing step that converts categorical labels into a numerical format that machine learning models can interpret. In this study, the vital status of patients, which was originally categorized as 'Death' or 'Alive', was encoded into binary labels. Specifically, 'Alive' was encoded as 0, and 'Death' was encoded as 1. This conversion was essential for the model to process and predict the target variable effectively, ensuring that the vital status data could be seamlessly integrated into the predictive framework.

4.2.1.6 Distribution of Vital Status

The dataset consists of a total of 360 entries, with a distribution of 231 classified as 'Alive' and 129 as 'Dead', which translates to 64.2% and 35.8%, respectively. To ensure

that our model could learn effectively and generalize well, it was crucial to maintain this distribution in both the training and testing sets.

The dataset was normalized so that the training set, which contains 288 entries, mirrors the original distribution with 185 entries classified as 'Alive' and 103 as 'Dead', maintaining the same percentages of 64.2% and 35.8%. Similarly, the testing set, consisting of 72 entries, was also adjusted to reflect this distribution with 46 entries labeled as 'Alive' and 26 as 'Dead', resulting in percentages of 63.9% and 36.1%.

4.2.1.7 Performance Metrics for Classification Models

To evaluate the performance of classification models, several metrics can be employed, including precision, recall, F1 score, and Matthews correlation coefficient (MCC). A valuable tool for assessing these metrics is the confusion matrix, which is a square matrix that summarizes the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the classifier. This visualization provides insights into the model's performance and is illustrated in Figure 23

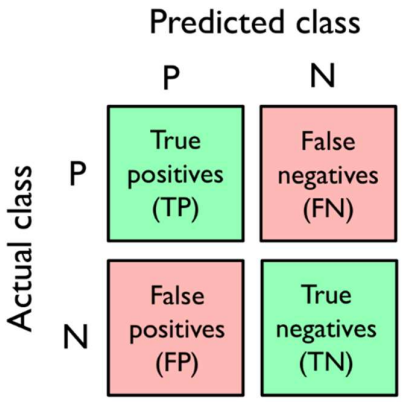


Figure 23 - The confusion matrix.

Both the prediction error and accuracy offer valuable insights into the number of misclassified examples. The prediction error is defined as the total number of false predictions divided by the overall number of predictions, while accuracy is calculated as the number of correct predictions divided by the total number of predictions (accuracy can also be calculated from the precision error):

$$\text{Prediction Error} = \frac{FP + FN}{FP + FN + TP + TN}$$

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} = 1 - \text{Prediction Error}$$

The true positive rate (TPR) and false positive rate (FPR) are performance metrics that are especially useful for imbalanced class problems:

$$True\ Positive\ Rate = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$False\ Positive\ Rate = \frac{FP}{N} = \frac{FP}{FP + TN}$$

The performance metrics precision and recall, mentioned in Chapter 3, are related to those TP and TN rates, and in fact, recall is synonymous with TPR:

$$Recall = True\ Positive\ Rate$$

In other words, recall quantifies how many of the relevant records (the positives) are captured as such (the true positives). Precision quantifies how many of the records predicted as relevant (the sum of true and false positives) are actually relevant (true positives):

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

To balance the up- and downsides of optimizing precision and recall, the harmonic mean of precision and recall is used, the so-called $F1$ score (as briefly mentioned in Chapter 3):

$$F1 = 2 \times \frac{\text{precision } \times \text{ recall}}{\text{precision} + \text{ recall}}$$

Lastly, the Matthews correlation coefficient (MCC) is a measure that summarizes the information in a confusion matrix and is particularly popular in biological research contexts. The MCC is calculated as follows:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

In contrast to precision, recall, and the F1 score, the MCC ranges between –1 and 1, and it takes all elements of a confusion matrix into account—for instance, the F1 score

does not involve the TN. While the MCC values are harder to interpret than the F1 score, it is regarded as a superior metric.

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are essential tools for evaluating the performance of classification models, particularly in scenarios involving imbalanced datasets. The ROC curve is a graphical representation that illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. As explained, the TPR indicates the proportion of actual positive cases correctly identified by the model, while the FPR reflects the proportion of negative cases incorrectly classified as positive. To generate the ROC curve, different thresholds are applied to the model's predicted probabilities, leading to a series of TPR and FPR values that are plotted on a two-dimensional graph. The resulting curve demonstrates how well the model distinguishes between the positive and negative classes. A model with perfect classification would achieve a TPR of 1 and an FPR of 0, resulting in a point in the top left corner of the ROC space. Conversely, a model that performs no better than random guessing would trace the diagonal line from (0,0) to (1,1). The ROC-AUC, which quantifies the overall ability of the model to discriminate between positive and negative classes, provides a single scalar value summarizing the ROC curve's performance. A ROC-AUC of 1 indicates perfect discrimination, while an ROC-AUC of 0.5 suggests no discrimination ability, equivalent to random chance. Generally, higher ROC-AUC values signify better model performance, making the ROC-AUC a valuable metric for comparing different classification models. In addition to providing insights into model performance, the ROC curve and ROC-AUC can help in selecting an optimal threshold for classification based on the specific requirements of the application. By adjusting the threshold, one can prioritize either maximizing the true positive rate (which may be critical in scenarios where failing to identify positive cases carries significant consequences) or minimizing the false positive rate (important in applications where false alarms must be reduced).

4.3 Results analysis

Analyzing multiple model combinations in this analysis is essential as it enables us to explore and understand how various data types—image, text, and tabular contribute to overall predictive performance. Each data type carries unique information and insights that enhance our understanding of the underlying patterns related to vital status in LIHC

patients. The diverse information sources provided by these models can highlight distinct aspects of patient conditions; for example, tabular data may present structured clinical features, text data can deliver nuanced insights from medical reports, and image data may capture visual characteristics pertinent to diagnosis or prognosis.

By comparing the performance of these different models, we can identify the strengths and weaknesses inherent to each data type. For example, while image data may effectively capture visual elements, it may not yield robust predictions on its own compared to a combination of text and tabular data. This comparative approach helps us pinpoint the most informative sources for our specific predictive goals. Furthermore, using a combination of data types can lead to improved predictive accuracy. Some models may perform better in identifying specific classes, such as 'Alive' or 'Dead', while others may excel at overall accuracy. Understanding these dynamics enables us to construct a more comprehensive model that maximizes predictive capabilities by integrating the strengths of each data type.

Ultimately, employing multiple models allows for robust decision-making in clinical settings, leading to better patient outcomes. By discerning which combination of data yields the best results, we can make informed clinical decisions and guide further research in predicting vital status and related health outcomes. In summary, the exploration of various models is crucial for determining the best approach to utilize the different data types effectively and achieve the most reliable and accurate results in LIHC research.

Table 3 presents a comparative analysis of classification performance metrics, including accuracy, Matthews correlation coefficient (MCC), precision, recall, F1 score, and ROC-AUC, for all model combinations across various data types.

Table 3 - Comparative analysis of classification performance metrics for all model combinations across various data types.

Model Type	Accuracy	MCC	Precision	Recall	F1 Score	ROC-AUC
Tabular Data Only	0.500	0.193	0.412	0.885	0.561	0.783
Text Data Only	0.514	0.164	0.412	0.808	0.546	0.642
Image Data Only	0.611	0.052	0.417	0.192	0.263	0.625
Tabular + Text Data	0.750	0.479	0.633	0.731	0.679	0.840
Tabular + Image Data	0.667	0.255	0.546	0.462	0.500	0.631
Text + Image Data	0.500	0.122	0.400	0.770	0.526	0.662
Tabular + Text + Image	0.639	0.274	0.500	0.654	0.567	0.634

Starting with the individual data types, the model utilizing only image data achieved the highest accuracy of 61.1%, despite having a notably low MCC of 0.052. This suggests that while the model may correctly classify some instances, it struggles to provide a balanced assessment of true and false positives and negatives. The precision and recall for the image-only model are also concerning, particularly the low recall of 19.2%, indicating that it misses a significant number of relevant instances.

In contrast, the tabular data only model had a similar accuracy of 50% but demonstrated a higher recall of 88.5%, indicating that it is effective at identifying relevant cases, though its overall precision is lower at 41.2%. This disparity suggests that while the model is catching most positive instances, it may also be incorrectly labeling a large number of negative instances as positive.

The model that combined tabular and text data significantly outperformed the others, with an accuracy of 75.0% and an MCC of 0.479. This combination achieved the highest precision (63.3%) and a strong recall (73.1%), resulting in a solid F1 score of

0.679. The ROC-AUC of 0.840 further emphasizes the model's robustness in distinguishing between classes, indicating a well-balanced performance across metrics.

On the other hand, combining tabular and image data yielded an accuracy of 66.7%, with a lower precision of 54.6% and a recall of 46.2%. This suggests that while the model benefits from integrating multiple data types, the specific contributions of each type may not be synergizing effectively.

The performance of models incorporating text and image data remained low, with an accuracy of 50.0% and an MCC of 0.122. This combination shows poor performance overall, particularly in precision and recall metrics.

Finally, the model combining tabular, text, and image data achieved an accuracy of 63.9%, which, while lower than the tabular and text combination, reflects the complexities introduced by integrating three data types. The MCC of 0.274 indicates some level of correlation between the observed and predicted classifications, although further improvement is necessary.

In conclusion, the results suggest combining tabular and text data yields the most effective predictive capability for determining the vital status of patients with LIHC, maximizing performance across the evaluated metrics. In contrast, the inclusion of image data appears to complicate matters without yielding significant benefits, underscoring the need for careful consideration of data types in model design and evaluation.

5 Conclusion

In this study, we explored the predictive capabilities of various data types—tabular, text, and image concerning the vital status of patients with LIHC. The analysis involved multiple models, each leveraging different forms of data to uncover insights about patient outcomes. Our findings underscore the importance of utilizing a diverse array of data sources, as each type contributes unique information that enhances our understanding of the factors influencing vital status.

The dataset comprised a total of 360 entries, with a distribution of 231 classified as 'Alive' and 129 as 'Dead'. To ensure the integrity of our analysis, we normalized the data to maintain consistent proportions of alive and deceased patients across both the training and testing sets. This careful attention to distribution was critical for preventing biases that could distort the model's learning process and evaluation.

Throughout the analysis, we utilized AutoMM, an automated machine learning framework designed to streamline the process of model selection and optimization. AutoMM's ability to handle multiple data types and facilitate efficient training processes made it a valuable tool in our efforts to predict vital status. However, one of the challenges we encountered was the relatively low number of data entries. This limitation posed difficulties in achieving robust model training and evaluation, potentially impacting the accuracy and reliability of our predictions.

Through our examination of individual and combined models, we determined that the Tabular + Text Data model yielded the highest accuracy at 75%. This finding highlights the efficacy of integrating both structured clinical features and nuanced textual insights to improve predictive performance. In contrast, models relying solely on image data performed less favorably, indicating that visual information alone may not suffice for accurate predictions in this context.

The study's results also revealed variations in performance among different combinations of data. For instance, the model utilizing tabular and text data together outperformed other combinations, while the integration of text and image data led to diminished predictive capabilities. These observations reaffirmed the value of conducting a comprehensive analysis that assesses how various data types interact and contribute to model performance.

Ultimately, the exploration of multiple models in this research serves as a foundation for understanding the best approaches to harness diverse data sources in predicting vital status for LIHC patients. By identifying the strengths and limitations of each data type, we can refine our methodologies to enhance predictive accuracy and reliability, thus informing clinical decision-making and guiding future research endeavors.

However, one of the challenges we encountered was the relatively low number of data entries. This limitation posed difficulties in achieving robust model training and evaluation, potentially impacting the accuracy and reliability of our predictions. Looking ahead, a future improvement of this project is to extend this analysis to encompass all cancer types, thereby broadening the applicability of our findings and contributing to a more comprehensive understanding of cancer prognostication. This work not only contributes to the ongoing efforts in cancer research but also highlights the importance of multi-modal data integration in achieving better patient outcomes. As we continue to advance our understanding of LIHC and its complexities, the insights gained from this analysis will be pivotal in driving improvements in patient care and outcomes in the field of oncology.

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., & Petersson, L. (2022). A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics*, *95*, 102027. https://doi.org/10.1016/j.compmedimag.2021.102027

Arango, S. P., Ferreira, F., Kadra, A., Hutter, F., & Grabocka, J. (2024). *Quick-Tune: Quickly Learning Which Pretrained Model to Finetune and How* (arXiv:2306.03828). arXiv. https://doi.org/10.48550/arXiv.2306.03828

Asan, O., Nattinger, A. B., Gurses, A. P., Tyszka, J. T., & Yen, T. W. F. (2018). Oncologists' Views Regarding the Role of Electronic Health Records in Care Coordination. *JCO Clinical Cancer Informatics*, *2*, 1–12. https://doi.org/10.1200/CCI.17.00118

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2018.2798607

Banales, J. M., Cardinale, V., Carpino, G., Marzioni, M., Andersen, J. B., Invernizzi, P., Lind, G. E., Folseraas, T., Forbes, S. J., Fouassier, L., Geier, A., Calvisi, D. F., Mertens, J. C., Trauner, M., Benedetti, A., Maroni, L., Vaquero, J., Macias, R. I. R., Raggi, C., Alvaro, D. (2016). Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the Study of Cholangiocarcinoma (ENS-CCA). *Nature Reviews. Gastroenterology & Hepatology*, *13*(5), 261–280. https://doi.org/10.1038/nrgastro.2016.51

Blechacz, B., & Gores, G. J. (2008). Cholangiocarcinoma: Advances in pathogenesis, diagnosis, and treatment. *Hepatology (Baltimore, Md.)*, *48*(1), 308–321. https://doi.org/10.1002/hep.22310

Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J., & Shah, S. P. (2022). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews. Cancer*, *22*(2), 114–126. https://doi.org/10.1038/s41568-021-00408-3

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bridgewater, J., Galle, P. R., Khan, S. A., Llovet, J. M., Park, J.-W., Patel, T., Pawlik, T. M., & Gores, G. J. (2014). Guidelines for the diagnosis and management of intrahepatic cholangiocarcinoma. *Journal of Hepatology*, *60*(6), 1268–1289. https://doi.org/10.1016/j.jhep.2014.01.021

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., … Lin, D. (2019). *MMDetection: Open MMLab Detection Toolbox and Benchmark* (arXiv:1906.07155). arXiv. https://doi.org/10.48550/arXiv.1906.07155

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2020). *On Empirical Comparisons of Optimizers for Deep Learning* (arXiv:1910.05446). arXiv. https://doi.org/10.48550/arXiv.1910.05446

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). *Multimodal learning with graphs* (arXiv:2209.03299). arXiv. https://doi.org/10.48550/arXiv.2209.03299

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data* (arXiv:2003.06505). arXiv. https://doi.org/10.48550/arXiv.2003.06505

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2022). *Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning* (arXiv:2007.04074). arXiv. https://doi.org/10.48550/arXiv.2007.04074

Gada, M., Haria, Z., Mankad, A., Damania, K., & Sankhe, S. (2021). *Automated Feature Engineering and Hyperparameter optimization for Machine Learning*. 981–986. https://doi.org/10.1109/ICACCS51430.2021.9441668

He, P., Gao, J., & Chen, W. (2023). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing* (arXiv:2111.09543). arXiv. https://doi.org/10.48550/arXiv.2111.09543

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). *EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification* (arXiv:1709.00029). arXiv. https://doi.org/10.48550/arXiv.1709.00029

Jiao, L., Chen, J., Liu, F., Yang, S., You, C., Liu, X., Li, L., & Hou, B. (2023). Graph Representation Learning Meets Computer Vision: A Survey. *IEEE Transactions on Artificial Intelligence*, *4*(01), 2–22. https://doi.org/10.1109/TAI.2022.3194869

Kalage, D., Gupta, P., Gulati, A., Yadav, T. D., Gupta, V., Kaman, L., Nada, R., Singh, H., Irrinki, S., Gupta, P., Das, C., Dutta, U., & Sandhu, M. (2023). Multiparametric MR imaging with diffusion-weighted, intravoxel incoherent motion, diffusion tensor, and dynamic contrast-enhanced perfusion sequences to assess gallbladder wall thickening: A prospective study based on surgical histopathology. *European Radiology*, *33*(7), 4981–4993. https://doi.org/10.1007/s00330-023-09455-w

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything* (arXiv:2304.02643). arXiv. https://doi.org/10.48550/arXiv.2304.02643

Li, F., Nian, Y., Sun, Z., & Tao, C. (2023). *Advancing Biomedicine with Graph Representation Learning: Recent Progress, Challenges, and Future Directions*. https://doi.org/10.48550/arXiv.2306.10456

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, J., Pandya, P., & Afshar, S. (2021). Therapeutic Advances in Oncology. *International Journal of Molecular Sciences*, *22*(4), Article 4. https://doi.org/10.3390/ijms22042008

Liu, X., Khalvati, F., Namdar, K., Fischer, S., Lewis, S., Taouli, B., Haider, M. A., & Jhaveri, K. S. (2021). Can machine learning radiomics provide pre-operative

differentiation of combined hepatocellular cholangiocarcinoma from hepatocellular carcinoma and cholangiocarcinoma to inform optimal treatment planning? *European Radiology*, *31*(1), 244–255. https://doi.org/10.1007/s00330-020-07119-7

Lowery, M. A., Ptashkin, R., Jordan, E., Berger, M. F., Zehir, A., Capanu, M., Kemeny, N. E., O'Reilly, E. M., El-Dika, I., Jarnagin, W. R., Harding, J. J., D'Angelica, M. I., Cercek, A., Hechtman, J. F., Solit, D. B., Schultz, N., Hyman, D. M., Klimstra, D. S., Saltz, L. B., & Abou-Alfa, G. K. (2018). Comprehensive Molecular Profiling of Intrahepatic and Extrahepatic Cholangiocarcinomas: Potential Targets for Intervention. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *24*(17), 4154–4161. https://doi.org/10.1158/1078-0432.CCR-18-0078

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). *Image Segmentation Using Deep Learning: A Survey* (arXiv:2001.05566). arXiv. https://doi.org/10.48550/arXiv.2001.05566

Müller, S. G., & Hutter, F. (2021). *TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation* (arXiv:2103.10158). arXiv. https://doi.org/10.48550/arXiv.2103.10158

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Comput. Surv.*, *55*(6), 114:1-114:29. https://doi.org/10.1145/3533378

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S.

(2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (arXiv:1912.01703). arXiv. https://doi.org/10.48550/arXiv.1912.01703

Quinn, M., Forman, J., Harrod, M., Winter, S., Fowler, K. E., Krein, S. L., Gupta, A., Saint, S., Singh, H., & Chopra, V. (2019). Electronic health records, communication, and data sharing: Challenges and opportunities for improving the diagnostic process. *Diagnosis (Berlin, Germany)*, *6*(3), 241–248. https://doi.org/10.1515/dx-2018-0036

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Razumilava, N., & Gores, G. J. (2014). Cholangiocarcinoma. *Lancet (London, England)*, *383*(9935), 2168–2179. https://doi.org/10.1016/S0140-6736(13)61903-0

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Rizvi, S., Khan, S. A., Hallemeier, C. L., Kelley, R. K., & Gores, G. J. (2018). Cholangiocarcinoma—Evolving concepts and therapeutic strategies. *Nature Reviews. Clinical Oncology*, *15*(2), 95–111. https://doi.org/10.1038/nrclinonc.2017.157

Rj, C., My, L., J, W., Dfk, W., Sj, R., Ni, L., & F, M. (2022). Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Transactions on Medical Imaging*, *41*(4). https://doi.org/10.1109/TMI.2020.3021387

Rowe, S. P., & Pomper, M. G. (2022). Molecular Imaging in Oncology: Current Impact and Future Directions. *CA: A Cancer Journal for Clinicians*, *72*(4), 333–352. https://doi.org/10.3322/caac.21713

Saha, S. K., Parachoniak, C. A., Ghanta, K. S., Fitamant, J., Ross, K. N., Najem, M. S., Gurumurthy, S., Akbay, E. A., Sia, D., Cornella, H., Miltiadous, O., Walesky, C., Deshpande, V., Zhu, A. X., Hezel, A. F., Yen, K. E., Straley, K. S., Travins, J., Popovici-Muller, J., … Bardeesy, N. (2014). Mutant IDH inhibits HNF-4α to block hepatocyte differentiation and promote biliary cancer. *Nature*, *513*(7516), 110–114. https://doi.org/10.1038/nature13441

Shchur, O., Turkmen, C., Erickson, N., Shen, H., Shirkov, A., Hu, T., & Wang, Y. (2023). *AutoGluon-TimeSeries: AutoML for Probabilistic Time Series Forecasting* (arXiv:2308.05566). arXiv. https://doi.org/10.48550/arXiv.2308.05566

Shi, J., Wang, R., Zheng, Y., Jiang, Z., Zhang, H., & Yu, L. (2021). Cervical cell classification with graph convolutional network. *Computer Methods and Programs in Biomedicine*, *198*, 105807. https://doi.org/10.1016/j.cmpb.2020.105807

Singh, G., Manjila, S., Sakla, N., True, A., Wardeh, A. H., Beig, N., Vaysberg, A., Matthews, J., Prasanna, P., & Spektor, V. (2021). Radiomics and radiogenomics in gliomas: A contemporary update. *British Journal of Cancer*, *125*(5), 641–657. https://doi.org/10.1038/s41416-021-01387-w

Sleeman IV, W. C., Kapoor, R., & Ghosh, P. (2021). *Multimodal Classification: Current Landscape, Taxonomy and Future Directions* (arXiv:2109.09020). arXiv. https://doi.org/10.48550/arXiv.2109.09020

Sun, Z., Cristea, A., Yu, J., Shi, L., & Al Moubayed, N. (2022, July 23). *Contrastive Learning with Heterogeneous Graph Attention Networks on Short Text Classification*. https://doi.org/10.1109/IJCNN55064.2022.9892257

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). *Measuring Robustness to Natural Distribution Shifts in Image Classification* (arXiv:2007.00644). arXiv. https://doi.org/10.48550/arXiv.2007.00644

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). *Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms* (arXiv:1208.3719). arXiv. https://doi.org/10.48550/arXiv.1208.3719

Tian, R., Zhao, Z., Liu, W., Liu, H., Mao, W., Zhao, Z., & Yan, K. (2022). *SAMP: A Toolkit for Model Inference with Self-Adaptive Mixed-Precision*. https://doi.org/10.48550/arXiv.2209.09130

Tian, Z., Li, X., Zheng, Y., Chen, Z., Shi, Z., Liu, L., & Fei, B. (2020). Graph-convolutional-network-based interactive prostate segmentation in MR images. *Medical Physics*, *47*(9), 4164–4176. https://doi.org/10.1002/mp.14327

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Waikhom, L., & Patgiri, R. (2022). A survey of graph neural networks in various learning paradigms: Methods, applications, and challenges. *Artif. Intell. Rev.*, *56*(7), 6295–6364. https://doi.org/10.1007/s10462-022-10321-2

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., & Huang, K. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, *12*(1), 3445. https://doi.org/10.1038/s41467-021-23774-w

Waqas, A., Bui, M. M., Glassy, E. F., El Naqa, I., Borkowski, P., Borkowski, A. A., & Rasool, G. (2023). Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *103*(11), 100255. https://doi.org/10.1016/j.labinv.2023.100255

Waqas, A., Tripathi, A., Ramachandran, R., Stewart, P., & Rasool, G. (2024). Multimodal data integration for oncology in the era of deep neural networks: A review. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1408843

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.2978386

Xu, P., Zhu, X., & Clifton, D. (2023). Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*. https://doi.org/10.1109/TPAMI.2023.3275156

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., & Shum, H.-Y. (2022). *DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection* (arXiv:2203.03605). arXiv. https://doi.org/10.48550/arXiv.2203.03605

Zhu, Z., Sun, M., Qi, G., Li, Y., Gao, X., & Liu, Y. (2024). Sparse Dynamic Volume TransUNet with multi-level edge fusion for brain tumor segmentation. *Comput. Biol. Med.*, *172*(C). https://doi.org/10.1016/j.compbiomed.2024.108284

Zöller, M.-A., & Huber, M. F. (2021). *Benchmark and Survey of Automated Machine Learning Frameworks* (arXiv:1904.12054). arXiv. https://doi.org/10.48550/arXiv.1904.12054

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). *Object Detection in 20 Years: A Survey* (arXiv:1905.05055). arXiv. https://doi.org/10.48550/arXiv.1905.05055