A Survey on Multimodal Representation Learning and Modality Fusion in Medical Data Analysis

Abstract

Multimodal representation learning and modality fusion have emerged as pivotal approaches in enhancing medical data analysis and healthcare outcomes. By integrating diverse data types such as structured Electronic Health Records (EHRs), clinical notes, and imaging data, these methodologies offer a comprehensive understanding of patient health, thereby improving predictive accuracy and clinical decision-making. Advanced computational techniques, including deep learning and graph-based models, facilitate the effective synthesis of complex datasets, enhancing scalability and adaptability in healthcare analytics. The integration of information technology in healthcare underscores the potential of these approaches, addressing challenges of data heterogeneity and high dimensionality. This survey highlights the transformative impact of multimodal learning on disease prediction, patient management, and diagnostic accuracy, while recognizing the need for ongoing research to address data quality, interpretability, and privacy concerns. By advancing fusion methods and integrating diverse data modalities, future research can enhance the robustness and generalizability of models, ultimately leading to improved patient outcomes and healthcare delivery. The survey provides a roadmap for researchers and practitioners, emphasizing the importance of interpretability, trustworthiness, and scalability in the development of multimodal models for healthcare applications.

1 Introduction

1.1 Importance of Data Integration

Data integration is crucial in multimodal representation learning, particularly in healthcare analytics, as it synthesizes various data types, including Electronic Health Records (EHRs), medical imaging, and clinical notes. This integration mitigates fragmentation and high dimensionality in healthcare data, fostering a holistic understanding of patient conditions and enhancing diagnostic accuracy. The amalgamation of heterogeneous clinical modalities, such as medical imaging and structured EHR data, significantly improves diagnostic precision and treatment decisions [1].

The challenge of merging structured and unstructured data within EHR systems often leads to fragmented patient data, hindering effective healthcare delivery [2]. Multimodal representation learning addresses this issue by utilizing multiple modalities to extract valuable information, requiring models to learn complex intra-modal and cross-modal interactions for effective predictions. This approach is particularly advantageous in electronic phenotyping, where the goal is to identify patients with specific diseases using their electronic medical records [3].

Ineffective data analysis methods that struggle with large datasets result in significant delays and inaccuracies, highlighting the necessity for efficient data integration strategies [4]. Early multimodal fusion techniques have demonstrated improvements in representation learning by integrating diverse data types at the initial stage, enhancing overall predictive performance [5]. However, many existing

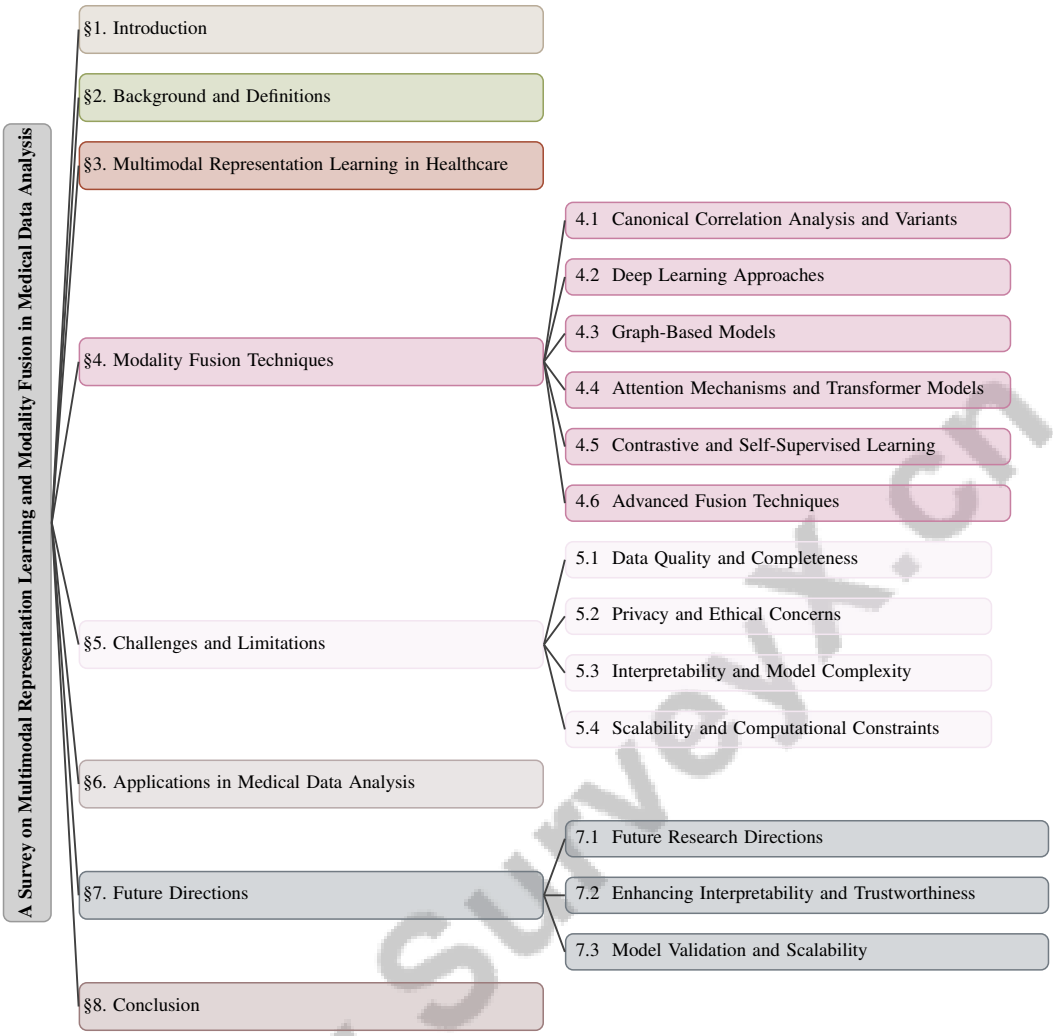


Figure 1: chapter structure

methods depend on complex tensor representations, leading to high computational costs and scalability limitations [6].

Despite advancements, challenges such as data quality and the lack of inter-modal correspondence persist, necessitating robust integration strategies to fully exploit EHRs in healthcare analytics. Addressing these challenges is vital for advancing healthcare analytics, enabling accurate diagnostics, improved patient management, and enhanced predictive models, ultimately supporting the development of machine learning models adept at navigating real-world healthcare complexities. Ruiz et al. [7] illustrate the critical role of data integration in analyzing lung cancer patient profiles, while Shickel et al. [8] point out limitations in previous methods lacking flexibility and comprehensiveness, further emphasizing the importance of data integration in multimodal representation learning.

1.2 Relevance to Medical Data Analysis

Multimodal representation learning significantly enhances medical data analysis by integrating diverse data sources, including structured EHRs, unstructured clinical notes, and medical imaging data, to create a comprehensive view of patient health. This approach capitalizes on the unique strengths of each data modality, thereby improving predictive accuracy and healthcare outcomes. Automated diagnostic methods, as noted in [6], are essential for the early screening of at-risk individuals, which is crucial for enhancing healthcare outcomes.

Advanced frameworks such as the CORE framework improve EHR data analysis by constructing fine-grained cohorts, thereby enhancing the representation of EHR data and contributing to better healthcare outcomes [9]. The integration of Natural Language Processing (NLP) into the data mining pipeline, as demonstrated by Ruiz et al., facilitates the extraction of detailed patient profiles, significantly contributing to improved healthcare outcomes [7].

Innovative methodologies, such as the Low-rank Multimodal Fusion (LMF) method, utilize low-rank tensors for efficient multimodal fusion, overcoming the limitations of existing methods [10]. Additionally, the incorporation of pre-trained vision-and-language models has shown superior performance in learning joint image-text embeddings from chest X-ray images and reports, significantly outperforming traditional CNN-RNN methods [11].

The Heterogeneous Information Network (HIN) approach enhances disease diagnosis accuracy by capturing the semantics of clinical relationships, showcasing the potential of multimodal representation learning to improve clinical outcomes [12]. Furthermore, the Causal Healthcare Embedding (CHE) method advances medical data analysis by improving prediction stability across different data distributions, addressing the challenges of unstable representation learning [3].

The integration of cross-modal sentiment information, as explored in [5], highlights the significance of effective multimodal fusion in enhancing sentiment analysis outcomes, contributing to more nuanced medical data analysis. Collectively, these advancements underscore the transformative potential of multimodal representation learning in medical data analysis, driving improvements in disease prediction models, diagnostics, and patient management, ultimately leading to enhanced healthcare delivery and outcomes.

1.3 Structure of the Survey

This survey provides a comprehensive examination of multimodal representation learning and modality fusion in medical data analysis, emphasizing the integration of diverse medical imaging modalities and corresponding textual reports. It explores the potential benefits of joint embeddings for clinical applications, including cross-domain retrieval and conditional report generation, while reviewing various learning methods and architectures that facilitate effective combinations of visual and textual information. The survey synthesizes recent advancements in the field, offering insights into key concepts, methodologies, and applications that are essential for enhancing multimodal intelligence in medical contexts [13, 14].

The paper begins with an introduction that establishes the significance of integrating diverse data types, such as text, images, and structured data from EHRs, to enhance decision-making in healthcare. Section 2 delves into background and definitions, providing an overview of key concepts such as electronic health records, data integration, and healthcare analytics, foundational to subsequent discussions.

Section 3 focuses on the application of multimodal representation learning in healthcare, examining how different modalities are combined to create comprehensive patient profiles. This section also discusses methodologies for integrating diverse data types, highlighting recent advancements in the field. Section 4 provides an in-depth examination of various modality fusion techniques, including canonical correlation analysis, deep learning approaches, graph-based models, attention mechanisms, and advanced fusion techniques, underscoring the technological innovations driving multimodal data integration and analysis.

Section 5 addresses the challenges and limitations associated with multimodal representation learning and modality fusion in medical data analysis, discussing issues related to data quality, privacy, interpretability, and scalability, which are critical for the successful implementation of these techniques in real-world healthcare settings. Section 6 elaborates on the practical applications of multimodal techniques in medical data analysis, emphasizing how these methods enhance disease prediction, patient management, and diagnostic accuracy. It presents case studies and real-world applications that illustrate the effectiveness of joint embeddings between diverse medical imaging modalities and their corresponding reports, as well as the benefits of image fusion techniques that integrate anatomical and physiological data. The review also discusses the challenges of selecting high-quality multimodal medical databases and evaluates various fusion methodologies that contribute to improved diagnostic outcomes [15, 14].

The survey concludes with Section 7, which discusses future research directions and opportunities in the field of EHRs, particularly focusing on advancements in clinical summarization, the integration of artificial intelligence in healthcare management, and the potential of multimodal representation learning to enhance diagnostic processes and patient care [16, 17, 18, 19]. It suggests areas for further exploration to enhance the effectiveness and applicability of multimodal techniques, focusing on improving interpretability, trustworthiness, and scalability of models. The survey aims to provide a roadmap for researchers and practitioners in the field, facilitating advancements in multimodal representation learning and modality fusion in healthcare.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Electronic Health Records (EHRs)

Electronic Health Records (EHRs) are comprehensive digital systems that store a wide range of patient health data, including structured elements like clinical codes and lab measurements, as well as unstructured components such as clinical narratives and reports. EHRs streamline the synthesis of critical patient information, which is essential for rapid decision-making in high-pressure environments like Emergency Departments. Tools like MedKnowts aid this process by automating data capture and facilitating natural language input, thereby reducing documentation burdens and enhancing data representation for machine learning applications in clinical settings [20, 21]. The high dimensionality and complexity of EHRs make them crucial for multimodal data integration, which is vital for constructing patient representations that enhance predictive modeling and clinical decision-making.

EHRs are foundational for integrating diverse clinical data modalities, including medical imaging and time-series data, which improves diagnostic accuracy and healthcare outcomes. However, their complexity, characterized by data sparsity, varied recording frequencies, and temporal irregularities, presents both challenges and opportunities for effective integration [2]. Deep learning models are employed to transform raw EHR data into meaningful patient representations, thereby improving diagnostic predictions and decision-making processes [22].

Integrating EHRs with other data modalities leverages the strengths of each, providing a nuanced understanding of patient health and disease trajectories. This integration is particularly beneficial for tasks like classifying clinical reports from CT imaging [23] and profiling lung cancer patients [7]. Nonetheless, the unstructured nature of EHRs and inconsistent recording practices complicate data harmonization efforts [8].

Despite these challenges, EHRs are indispensable for multimodal data integration, enabling the synthesis of rich medical contexts from various data types, which enhances clinical prediction accuracy and the development of robust machine learning models. Advancements in learning vector representations for medical concepts from clinical knowledge graphs further underscore the utility of EHRs in enhancing healthcare analytics [12].

2.2 Healthcare Analytics

Healthcare analytics leverages computational techniques to derive insights from complex multimodal datasets, significantly enhancing the processing of diverse data types, including EHRs, medical imaging, and genomic data [24]. This integration facilitates predictive model development that improves patient outcomes and optimizes clinical workflows.

Artificial intelligence (AI) in healthcare analytics offers opportunities and challenges, particularly in multimodal data analysis. While AI techniques like deep learning are adept at extracting patterns from extensive data, they face barriers related to data quality, interpretability, and scalability [25]. These challenges necessitate robust frameworks to manage healthcare data's complexity and heterogeneity.

Performance evaluation in healthcare analytics often uses metrics like the F1-score, which provides a balanced measure of precision and recall, especially in imbalanced datasets [26]. Additionally, the realism of synthetic data versus real data is assessed using privacy risk metrics, emphasizing data integrity and confidentiality [27].

The effectiveness of multitask learning in healthcare analytics is demonstrated through datasets containing binary indicators for ICD-9, CPT, and RxNorm codes, along with demographic features such as age, gender, race, and ethnicity [28]. These features are crucial for building models that accurately reflect patient health states and support clinical decision-making.

Graph-based embedding methods are evaluated through benchmarks focusing on healthcare-relevant tasks like node classification, link prediction, and patient state prediction [29]. These methods enhance the ability of healthcare analytics to model complex relationships within multimodal data, thereby improving predictive model accuracy and reliability.

3 Multimodal Representation Learning in Healthcare

The integration of multiple data modalities in healthcare is essential for harnessing the comprehensive information within healthcare datasets. This integration is not only a technical challenge but a necessity for effective multimodal representation learning, crucial for enhancing clinical decision-making and predictive analytics. As illustrated in Figure 2, the hierarchical structure of multimodal representation learning in healthcare emphasizes the integration of diverse data modalities and advanced methodologies in multimodal learning. This figure highlights key fusion strategies and innovative methodologies that enhance data integration and predictive analytics, showcasing the transformative potential of these approaches in improving healthcare outcomes. The following subsection explores the integration of these diverse data modalities, focusing on innovative approaches that enhance multimodal representation learning in clinical settings.

3.1 Integration of Diverse Data Modalities

Integrating diverse data modalities, such as structured EHRs, clinical notes, and imaging data, is pivotal for advancing multimodal representation learning in healthcare. This integration allows for comprehensive models that exploit the strengths of each data source, thereby improving clinical decision-making and predictive analytics. The complexity of health informatics, characterized by high dimensionality, heterogeneity, and sparsity, demands sophisticated methodologies for effective synthesis. Temporal cross-attention transformers, for instance, extract features from medical time series and clinical notes to address integration challenges [2].

Fusion strategies for integrating multiple modalities are categorized into early, late, and joint fusion approaches. Early fusion combines data at initial stages for a unified representation, while late fusion integrates individual analyses to leverage each modality's strengths. Joint fusion processes multiple modalities simultaneously, often using advanced deep learning techniques to enhance clinical outcomes. The hierarchical network by CORE, for example, constructs fine-grained cohorts based on medical codes, modeling patient relevance through the Jaccard similarity of diagnosis codes [9].

Innovative methodologies like the Clinical Knowledge Extraction System (C-liKES) normalize and harmonize EHRs to extract structured clinical information from free-text records, facilitating diverse data type integration [7]. Similarly, a flexible Transformer-based EHR embedding pipeline integrates diverse EHR data with minimal preprocessing, enabling simultaneous predictions of multiple clinical outcomes [8].

Efforts to harmonize structured and unstructured data, as demonstrated by temporal cross-attention transformers, further advance healthcare analytics by addressing cross-modal interactions and data irregularities [2]. This integration not only improves predictive model accuracy but also supports robust frameworks capable of navigating the complexities of real-world healthcare settings.

3.2 Advanced Methodologies in Multimodal Learning

Recent advancements in multimodal representation learning have introduced innovative methodologies that significantly enhance the integration and analysis of diverse healthcare data, leading to improved predictive accuracy and clinical decision-making. An architecture combining matrix factorization and convolutional autoencoders effectively extracts representations from both matrix and tensor data, addressing challenges such as high dimensionality and class imbalance. This approach leverages multimodal AI techniques for simultaneous feature learning from diverse data types,

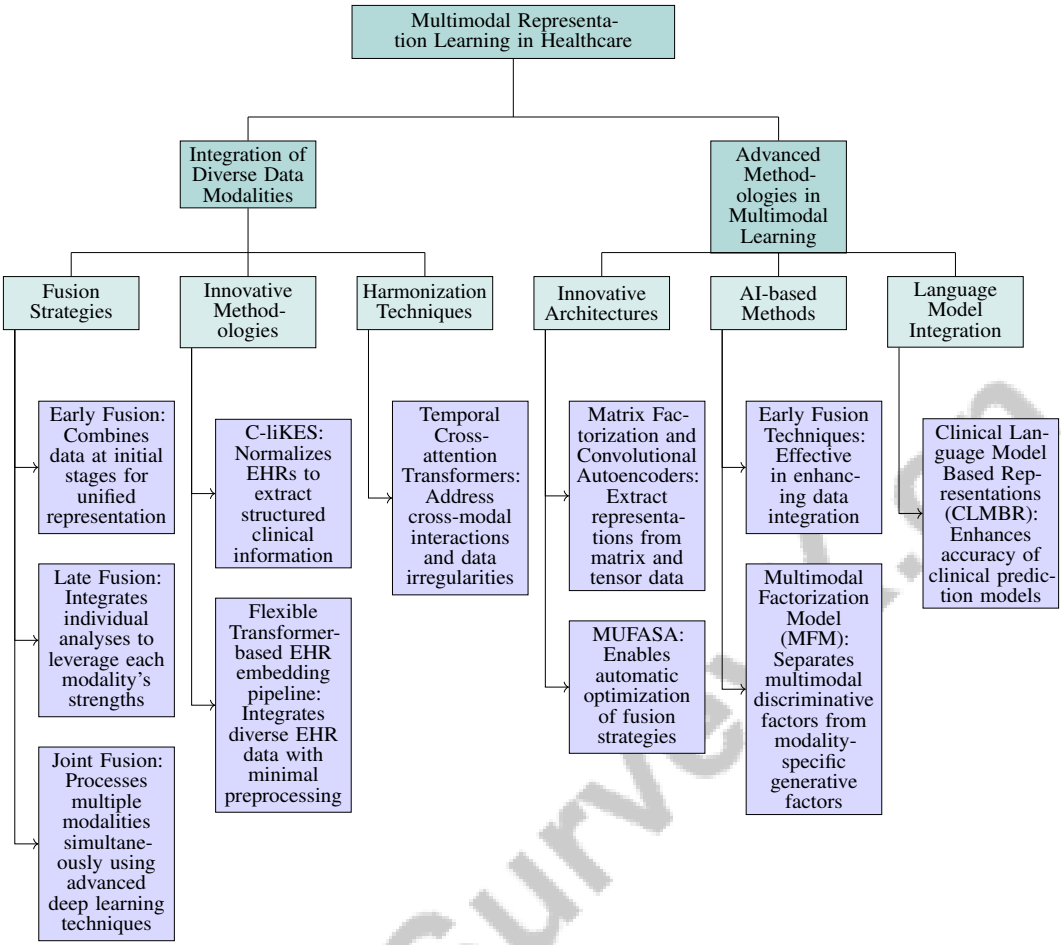


Figure 2: This figure illustrates the hierarchical structure of multimodal representation learning in healthcare, focusing on the integration of diverse data modalities and advanced methodologies in multimodal learning. It highlights key fusion strategies and innovative methodologies that enhance data integration and predictive analytics, showcasing the transformative potential of these approaches in improving healthcare outcomes.

including text and images, while managing intra-modal and cross-modal interactions, demonstrating robust performance even with missing or noisy modalities [30, 17].

MUFASA represents a significant advancement by enabling automatic optimization of fusion strategies tailored to each modality, improving model adaptability to diverse data sources [31]. This flexibility is crucial for accommodating variability in clinical data and enhancing multimodal system performance.

AI-based methods have been pivotal in advancing the fusion of EHR and medical imaging data. Early fusion techniques have proven effective, underscoring the growing trend of using AI to enhance data integration and improve clinical outcomes [32]. These methodologies leverage the unique strengths of each modality, resulting in more accurate patient representations.

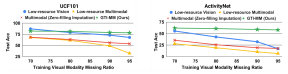
The Multimodal Factorization Model (MFM) separates multimodal discriminative factors from modality-specific generative factors, improving interpretability and performance in multimodal tasks [30]. This approach provides nuanced insights into the interactions between different data modalities, facilitating reliable healthcare predictions.

The Clinical Language Model Based Representations (CLMBR) have notably enhanced the accuracy of clinical prediction models, particularly with smaller sample sizes, by effectively capturing clinical

language semantics [33]. This advancement highlights the importance of integrating language models into the multimodal learning pipeline to bolster predictive capabilities.

Furthermore, employing low-rank weight tensors and modality-specific factors in innovative methodologies allows for scalable solutions that reduce computational complexity while maintaining performance [10]. This is particularly advantageous in healthcare settings where computational resources may be limited, yet the demand for accurate predictions is high.

These advancements underscore the transformative potential of recent methodologies in multimodal representation learning. The integration of advanced multimodal EHR data, exemplified by the EMERGE framework, significantly enhances disease prediction, diagnostics, and patient management by incorporating nuanced medical contexts and leveraging AI for clinical summarization. This leads to more accurate clinical predictions, streamlined communications, and improved patient-centered care, ultimately resulting in better healthcare delivery and outcomes [16, 34].



(a) Comparison of Low-resource Vision and Low-resource Multimodal Methods on UCF101 and ActivityNet[35]

(b) Joint-Optimal Intermediates for Unimodal and Coordinated Learning[36]

(c) Generative Adversarial Network for Multimodal Image Synthesis[30]

Figure 3: Examples of Advanced Methodologies in Multimodal Learning

As illustrated in Figure 3, "Multimodal Representation Learning in Healthcare: Advanced Methodologies in Multimodal Learning" is depicted through three distinct approaches, each highlighting the versatility and complexity of multimodal learning techniques. The first approach focuses on the "Comparison of Low-resource Vision and Low-resource Multimodal Methods on UCF101 and ActivityNet," analyzing the effectiveness of three different methods in scenarios with varying degrees of missing visual modality data. This comparison underscores the potential of multimodal methods to maintain higher test accuracy even when significant portions of visual data are unavailable. The second approach, "Joint-Optimal Intermediates for Unimodal and Coordinated Learning," presents a schematic of a learning process utilizing joint and optional intermediate representations. This dual representation strategy enhances learning efficiency in multimodal contexts. Lastly, the "Generative Adversarial Network for Multimodal Image Synthesis" exemplifies the use of GANs to generate synthetic images from multiple latent variables, showcasing the capability of these networks to create diverse and realistic multimodal outputs. Together, these methodologies highlight the advanced strategies employed in multimodal learning, with significant implications for improving healthcare technologies through robust and resource-efficient data integration [35, 36, 30].

4 Modality Fusion Techniques

In the context of modality fusion techniques, understanding the statistical relationships between diverse data sources is essential for effective integration and analysis. Table 1 presents a detailed classification of modality fusion techniques, illustrating the diverse approaches utilized to integrate and analyze multimodal healthcare data. Additionally, Table 5 offers a comprehensive comparison of different modality fusion techniques, illustrating their integration strategies and unique contributions to healthcare data analysis. This section explores the foundational methodologies that underpin modality fusion, beginning with Canonical Correlation Analysis (CCA) and its variants. These approaches provide a robust framework for quantifying the relationships between high-dimensional datasets, thereby facilitating the synthesis of multimodal information in healthcare applications. The subsequent subsection delves into the intricacies of CCA, highlighting its significance and the innovations that enhance its application in the integration of healthcare data modalities.

4.1 Canonical Correlation Analysis and Variants

Canonical Correlation Analysis (CCA) and its variants are integral to modality fusion, particularly in the context of integrating diverse healthcare data modalities. CCA is a statistical approach designed to

Category	Feature	Method
Canonical Correlation Analysis and Variants	Statistical and Dependency Analysis	FRSE[37], MSN-EHR[38], SCDMF[39]
	Interaction and Fusion Modeling	cHITF[40]
	Self-Supervised and Contrastive Learning	MEDHMP[41], GCLF[2]
Deep Learning Approaches	Multimodal Integration	HF[1], DescEmb[42], MA-MGMU[43], C-LSTM[44], FTEHR[8]
	Temporal and Topic Modeling	PTM[45], TM-Classification[23]
	Contrastive and Embedding Techniques	CGL[46], CLAIME[47]
Attention Mechanisms and Transformer Models	Multimodal Integration	EHR-CP[48], DES[49], IMLP[50], IRENE[51], SSRL[52], QFES[18], HTAD[53]
Contrastive and Self-Supervised Learning	Data Integration	SSPCM[54]
Advanced Fusion Techniques	Multimodal Integration	VKD[55], OCLEAR[56], MICRO[57], RAG[58], MODETL[59]
	Temporal Analysis	CTPD[60]

Table 1: This table provides a comprehensive overview of various methodologies employed in modality fusion, with a focus on healthcare applications. It categorizes the methods into Canonical Correlation Analysis and Variants, Deep Learning Approaches, Attention Mechanisms and Transformer Models, Contrastive and Self-Supervised Learning, and Advanced Fusion Techniques, highlighting key features and representative methods within each category.

Method Name	**Modality Fusion**	**Data Representation**	**Learning Strategies**
SCDMF[39]	Multimodal Data Fusion	Multimodal Representation Learning	Adaptive Weighting
FRSE[37]	-	Semantic Embedding Techniques	-
MSN-EHR[38]	Fuse The Representations	Joint Representations	Self-supervised Pretraining
MEDHMP[41]	Multiple Modalities	Complex Clinical Predictions	Hierarchical Pretraining Strategy
GCLF[2]	Temporal Cross-attention	Dynamic Embedding Scheme	Global Contrastive Loss
cHITF[40]	Hidden Interaction Tensor	Latent Factor Matrices	Tensor Factorization Method

Table 2: Overview of various methods used in modality fusion, data representation, and learning strategies for multimodal healthcare data analysis. The table compares different approaches, including multimodal data fusion, semantic embedding techniques, and hierarchical pretraining strategies, highlighting their contributions to advancing electronic phenotyping and multimodal representation learning.

identify and quantify the relationships between two high-dimensional datasets, thereby facilitating the understanding of complex inter-modal interactions. The application of CCA in healthcare analytics is further enhanced by the Statistical Correlation-Driven Multimodal Fusion (SCDMF) method, which integrates statistical analysis with human-centered insights, aligning well with CCA techniques [39].

Table 2 presents a comprehensive comparison of methods employed in the fusion of diverse healthcare data modalities, detailing their strategies for data representation and learning, which are crucial for enhancing multimodal healthcare analytics.

The use of CCA extends beyond simple correlation analysis, playing a crucial role in the semantic integration of heterogeneous data sources such as Electronic Health Records (EHRs) and medical imaging. For instance, the Full-Record Semantic Embedding (FRSE) method captures complex dependencies within clinical data, enabling the deployment of computationally efficient models that maintain high performance [37]. This capability is crucial in clinical settings where computational resources are often limited.

The integration of EHR data with other modalities, such as Chest X-ray (CXR) images, is further advanced by innovations like the EHR encoder and projection module within the Multimodal Masked Siamese Network (MSN) architecture. This framework allows for the learning of joint representations from both CXR images and EHR data, enhancing the fusion process [38].

Moreover, the Hierarchical Pretraining strategy employed by MEDHMP distinguishes itself by utilizing level-specific self-supervised learning tasks, addressing the unique challenges posed by the diverse and hierarchical nature of EHR data [41]. This approach complements CCA by enhancing the alignment of multimodal feature representations, as demonstrated by the global contrastive loss technique that aligns features with corresponding discharge summaries [2].

In the realm of unsupervised learning, methods such as those introduced in the benchmark for joint embedding allow for learning from unannotated data, which is crucial for advancing modality fusion techniques [14]. Additionally, the cHITF framework models hidden interactions between multiple modalities in EHR data, further improving modality fusion [40].

Canonical correlation analysis and its variants, as outlined in existing research methodologies for electronic phenotyping, provide a structured framework for integrating diverse data sources, ultimately enhancing the interpretability and clinical applicability of multimodal healthcare analytics [61]. These techniques continue to evolve, playing a critical role in advancing the capabilities of multimodal representation learning in medical data analysis.

4.2 Deep Learning Approaches

Method Name	Methodologies Used	Data Types	Application Scenarios
TM-Classification[23]	Topic Modeling	Clinical Reports	Clinical Report Analysis
C-LSTM[44]	Convolutional Lstm	Audio And Visual	Emotion Recognition
MA-MGMU[43]	Attention Mechanisms	Medical Imaging	Clinical Decision-making
CLAIME[47]	Contrastive Learning	Multimodal Ehr	Patient Data Analysis
DescEmb[42]	Neural Text Encoder	Ehr Data	Predictive Modeling
HF[1]	Hypernetwork	Medical Imaging	Predictive Modeling
PTM[45]	Recurrent Architectures	Clinical Text	Clinical Decision-making
CGL[46]	Attention Mechanism	Ehr Data	Predictive Modeling
FTEHR[8]	Transformer Architectures	Ehr Data	Clinical Outcome Predictions

Table 3: This table provides a comprehensive overview of various deep learning methods utilized in the fusion of multimodal healthcare data. It details the methodologies employed, types of data integrated, and the specific application scenarios in which these methods are applied, highlighting their role in enhancing predictive accuracy and clinical decision-making.

Deep learning approaches have become pivotal in modality fusion, offering advanced techniques that enhance the integration of diverse healthcare data modalities. These methodologies significantly improve predictive accuracy and clinical decision-making. The survey evaluates deep learning approaches among current methods of multi-modality image fusion, highlighting their effectiveness and suitability for different medical applications [62]. A prominent innovation is the introduction of a two-step masking process in masked language models, which enhances the model's ability to handle interactions between diseases and interventions. This technique exemplifies the adaptability of deep learning models in managing the complexities of healthcare datasets.

Hybrid architectures, such as those combining convolutional neural networks for text features with multi-layer perceptrons for structured data, demonstrate the effectiveness of deep learning in capturing nuanced inter-modal interactions [23]. These models leverage the strengths of each component to construct comprehensive patient representations, thereby enhancing the fusion process. The C-LSTM method demonstrates a deep learning approach for modality fusion, integrating audio and visual inputs simultaneously to enhance performance [44].

Attention mechanisms and gated multimodal units, as employed in methods that fuse medical imaging and Electronic Health Records (EHR) data, further illustrate the transformative impact of deep learning in modality fusion [43]. These techniques facilitate the alignment of multimodal features, improving the accuracy and interpretability of predictive models. CLAIME is a multimodal feature embedding generative model that employs a contrastive loss to learn representations from multimodal EHR data [47].

The integration of shared embedding spaces, achieved through leveraging text descriptions, overcomes the limitations of code-based methods and enhances the fusion of heterogeneous EHR data. This approach underscores the importance of creating unified representations that capture the semantics of clinical data [42]. HyperFusion is a deep learning framework that integrates imaging and tabular data by utilizing a hypernetwork to condition the image analysis on the tabular data [1].

Recurrent architectures, such as those used in modeling patient trajectories, highlight the importance of temporal dynamics in deep learning models. These models are particularly effective in capturing the progression of patient health over time, thereby improving the predictive capabilities of multimodal systems [45]. The use of dual-VAE frameworks and sequentially coupled generators, as seen in EHR-M-GAN, captures temporal dynamics and learns shared latent representations, facilitating the integration of mixed-type longitudinal data.

Furthermore, the collaborative graph learning model, which combines hierarchical disease embeddings with attention-regulated unstructured text features, demonstrates the potential of deep learning in integrating complex healthcare data [46]. The method described in the paper utilizes a modified Transformer encoder architecture, which represents a deep learning approach for modality fusion in

medical data analysis [8]. This approach enhances the interpretability and performance of predictive models by effectively modeling the interactions between different data modalities.

The integration of advanced neural architectures through deep learning approaches has significantly transformed modality fusion, particularly in the context of Electronic Health Records (EHR) and multimodal data analysis. This evolution has led to substantial improvements in predictive modeling and healthcare analytics, as evidenced by the development of automated frameworks like AutoFM, which optimally encodes diverse input modalities and enhances model performance. Furthermore, the application of transformers in healthcare facilitates the analysis of various data types, including medical imaging and structured EHR, thereby advancing clinical diagnosis and data-driven decision-making processes. Table 3 presents a detailed summary of deep learning approaches for modality fusion in healthcare, demonstrating the diverse methodologies and application scenarios that contribute to improved clinical outcomes. Overall, these innovations underscore the critical role of multimodal intelligence in enhancing healthcare outcomes and predictive accuracy. [63, 64, 13, 65, 66]. The integration of hybrid and ensemble methods further enhances the robustness and adaptability of multimodal systems, paving the way for more precise and personalized healthcare interventions.

4.3 Graph-Based Models

Graph-based models play a crucial role in the integration and analysis of multimodal data within healthcare analytics, offering a sophisticated framework to capture the intricate relationships between diverse data types. These models leverage the inherent interconnectedness of medical data, facilitating a more comprehensive understanding of patient health and disease progression. A notable advancement in this domain is HetMed, a heterogeneous graph learning framework that captures complex patient relationships, significantly enhancing the capabilities of traditional single-modality approaches [67].

The introduction of novel graph-based embedding methods, such as node2vec, metapath2vec, and Poincaré embeddings, represents a significant improvement in leveraging the hierarchical structure of medical concepts. These methods, as outlined in the benchmark, effectively utilize the hierarchical nature of medical datasets, offering more nuanced embeddings that improve the accuracy of predictive models [29]. By capturing the semantic relationships between different medical entities, these embeddings facilitate the integration of heterogeneous data sources, thereby enhancing the interpretability and performance of healthcare analytics.

Graph-based models are particularly effective in modeling the complex interactions between structured Electronic Health Records (EHRs) and unstructured data, such as clinical notes and imaging data. These models utilize graph convolutional networks and attention mechanisms to effectively integrate and align multimodal features from diverse sources such as electronic health records (EHR), medical images, and clinical notes. This integration enhances the representation of patient data, leading to improved predictive capabilities in healthcare systems, as evidenced by performance improvements in tasks like hospital readmission prediction and clinical diagnostics across various studies. [68, 11, 69, 70, 51]. The use of graph structures allows for the incorporation of domain knowledge, such as clinical guidelines and disease ontologies, which further enriches the analysis and supports more informed clinical decision-making.

The flexibility and scalability of graph-based models make them well-suited for handling the high dimensionality and heterogeneity inherent in multimodal healthcare data. By establishing a comprehensive framework for the integration of various data modalities, these models enhance the ability to develop effective and scalable solutions that tackle the intricate challenges present in real-world healthcare settings. This is achieved through the utilization of advanced techniques such as cohort representation learning, which captures the nuanced relationships among patients, and Retrieval-Augmented Generation (RAG) methods that incorporate external medical knowledge to enrich multimodal Electronic Health Records (EHR). As a result, these frameworks not only improve predictive capabilities but also ensure the contextual relevance necessary for informed clinical decision-making. [34, 58, 9]. Graph-based approaches continue to drive advancements in multimodal representation learning, offering promising avenues for improving disease prediction, patient management, and healthcare delivery.

Method Name	Modality Fusion	Attention Mechanisms	Model Interpretability
EHR-CP[48]	Fuse Data Modalities	Attention Mechanisms Fuse	Improving Interpretability
IMLP[50]	Multiple Feature Fusion	Graph Attention Network	Linear Fusion Strategies
DES[49]	-	Spotlight Important Events	Enhancing Interpretability
QFES[18]	-	Attention Mechanisms	Relevant Text Snippets
HTAD[53]	Additional Data Types	Hierarchical Attention Mechanism	Enhanced Interpretability
SSRL[52]	Multimodal Data Integration	-	-
IRENE[51]	Multimodal Attention Mechanisms	Multimodal Attention Blocks	Attention Mechanisms

Table 4: Comparison of Various Models Utilizing Modality Fusion and Attention Mechanisms in Healthcare Analytics. This table outlines the different methods employed by several models to integrate and interpret diverse healthcare data, highlighting their approach to modality fusion, attention mechanisms, and model interpretability.

4.4 Attention Mechanisms and Transformer Models

Attention mechanisms and transformer models have become pivotal in advancing modality fusion within healthcare analytics, offering sophisticated techniques to enhance the integration and interpretation of diverse healthcare data. These models excel in focusing on the most relevant features across different modalities, thereby improving predictive accuracy and clinical decision-making. The application of attention mechanisms in fusing data from clinical notes and blood test results exemplifies their effectiveness in synthesizing heterogeneous data sources [48].

The use of graph attention networks for EHR feature selection, as demonstrated in cardiovascular studies, highlights the role of attention mechanisms in identifying critical features that contribute to improved healthcare outcomes [50]. These networks leverage the relational structure of EHR data, allowing for more nuanced feature selection and better model interpretability.

Innovative models such as the deep spotlight framework utilize attention mechanisms to enhance the interpretability of predictions by highlighting significant events within a 2D pathway representation [49]. This approach underscores the importance of attention mechanisms in improving the transparency and reliability of healthcare predictions.

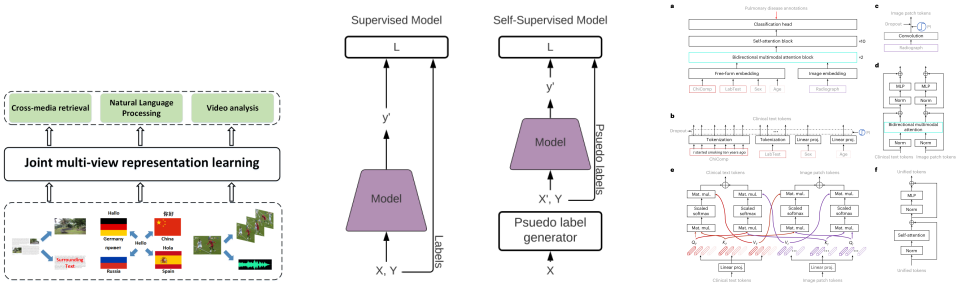
Attention mechanisms are also employed in query-focused EHR summarization, where they enhance the efficiency and accuracy of diagnosis support by concentrating on the most relevant text snippets [18]. This capability is crucial in clinical settings where timely and accurate information retrieval is paramount.

Hierarchical attention mechanisms, as utilized in target-attentive diagnosis prediction models, further demonstrate the versatility of attention-based approaches in modality fusion [53]. These mechanisms enable the alignment of multimodal features, facilitating more accurate and interpretable healthcare analytics.

Attention mechanisms and transformer models are pivotal in advancing modality fusion, as they facilitate the effective integration of diverse data types—such as images, text, and clinical information—by employing sophisticated embedding layers and attention mechanisms. This approach not only enhances model interpretability through the generation of holistic representations but also significantly improves clinical outcomes, evidenced by superior performance in diagnosing pulmonary diseases and predicting adverse outcomes in patients with COVID-19 compared to traditional models. By streamlining the triaging process and aiding clinical decision-making, these models represent a significant leap forward in multimodal representation learning and its applications in healthcare [51, 13, 64, 66]. Their continued development and application in healthcare analytics promise to drive significant advancements in predictive modeling and personalized healthcare delivery.

Table 4 provides a comprehensive overview of several models that leverage modality fusion and attention mechanisms to enhance healthcare analytics, elucidating their methodologies and contributions to model interpretability.

As shown in Figure 4, In the realm of advanced machine learning and artificial intelligence, the integration of modality fusion techniques, attention mechanisms, and transformer models has become pivotal in enhancing system performance across various applications. The provided examples illustrate the diverse applications and effectiveness of these techniques. The first example, "Joint Multi-View Representation Learning for Cross-Media Retrieval and Natural Language Processing," showcases a framework that harmonizes cross-media retrieval, natural language processing, and video



(a) Joint Multi-View Representation Learning for Cross-Media Retrieval and Natural Language Processing[71]

(b) Comparison between Supervised and Self-Supervised Learning Models[52]

(c) Multimodal Attention Network for Pulmonary Disease Classification[51]

Figure 4: Examples of Attention Mechanisms and Transformer Models

analysis to bolster system capabilities through a unified multi-view representation learning approach. The second example offers a comparative analysis between supervised and self-supervised learning models, highlighting the distinctions in training methodologies and the implications on model performance. The third example, "Multimodal Attention Network for Pulmonary Disease Classification," demonstrates the application of a sophisticated multimodal attention network architecture, which leverages layers such as self-attention and bidirectional multimodal attention to effectively classify pulmonary diseases. Together, these examples underscore the transformative potential of integrating modality fusion, attention mechanisms, and transformer models in addressing complex challenges in machine learning and artificial intelligence [71, 52, 51].

4.5 Contrastive and Self-Supervised Learning

Recent advancements in contrastive and self-supervised learning techniques have positioned them as influential methodologies in the field of modality fusion, particularly for improving the integration and analysis of diverse medical data types, such as structured electronic health records (EHR) and unstructured clinical notes. These approaches leverage the complementary nature of different data modalities to create a more comprehensive representation of patient health information. For instance, multimodal contrastive learning has shown promise in capturing the intricate relationships between modalities, enhancing the ability to analyze complex medical datasets. Additionally, self-supervised learning methods, which derive supervisory signals from the data itself, have gained traction for their efficiency in handling large-scale datasets without the need for extensive annotation. Together, these techniques facilitate more nuanced insights into patient health by effectively merging various types of clinical data and improving the accuracy of medical image segmentation and other analytical tasks. [47, 72, 73]. These approaches leverage the inherent structure of multimodal data to learn robust representations without the need for extensive labeled datasets, making them highly valuable in healthcare settings where annotated data can be scarce.

Self-supervised learning, in particular, has shown significant promise in improving prediction accuracy by utilizing unlabeled data to pre-train models that can then be fine-tuned for specific tasks. The integration of multimodal data with self-supervised learning frameworks has been demonstrated to enhance predictive accuracy, especially for forecasting far-future events, which is critical in the context of long-term patient management and outcome prediction [54]. This approach allows for the extraction of meaningful patterns from complex datasets, thereby improving the overall performance of predictive models.

Contrastive learning, on the other hand, focuses on learning representations by contrasting positive pairs (similar data points) against negative pairs (dissimilar data points). This method is particularly effective in distinguishing subtle differences between data modalities, thereby facilitating more precise integration and analysis. The use of contrastive learning in modality fusion is exemplified by its application in extracting medical concepts and temporal reasoning from unstructured clinical narratives [74]. By leveraging state-of-the-art natural language processing models, contrastive learning enhances the ability to derive meaningful insights from complex clinical data.

Overall, the application of contrastive and self-supervised learning techniques in modality fusion represents a significant advancement in healthcare analytics. These methodologies enhance the accuracy and robustness of predictive models by effectively integrating and utilizing diverse and unstructured data sources, such as clinical narratives and examination reports, alongside structured electronic health records (EHR). This comprehensive approach not only addresses the limitations of relying solely on structured data—demonstrated by its inadequacy in meeting eligibility criteria for clinical trials—but also facilitates richer medical context through advanced representation learning and multimodal fusion techniques. Ultimately, these improvements contribute to more informed clinical decision-making and the delivery of personalized healthcare by leveraging the full spectrum of patient information. [34, 21, 74]

4.6 Advanced Fusion Techniques

Advanced fusion techniques play a crucial role in the integration of multimodal data, offering innovative solutions that enhance predictive accuracy and clinical decision-making in healthcare. A significant advancement in this domain is the use of Variational Knowledge Distillation (VKD), which utilizes Electronic Health Records (EHRs) during the training phase to enable disease classification based solely on images at inference time. This method enhances the flexibility and efficiency of multimodal data utilization, allowing for more effective integration of diverse data types [55].

Another notable approach is the Dynamic Resource Allocation Algorithm (DRAA), which dynamically allocates resources based on real-time workload assessments. This capability enhances the efficiency of computational resource utilization and streamlines the processing and analysis of multimodal data, leading to significant performance improvements compared to traditional methods. It addresses key challenges in multimodal retrieval, such as managing skewed data and noisy labels, integrating diverse information from different modalities, and facilitating effective training in large-scale industrial contexts, as demonstrated by the successful deployment of advanced frameworks like OCLEAR in real-world applications. [10, 13, 56, 66]

The Category-Oriented Learning with Exemplar-based Adaptive Representation (OCLEAR) addresses challenges such as data skewness and noise through a novel data governance scheme and a large-scale classification-based learning paradigm. This method effectively manages the complexity of multimodal data, improving the robustness and accuracy of predictive models [56].

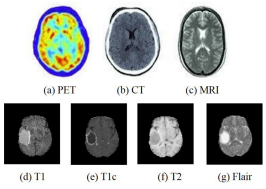
MICRO introduces a multimodal contrastive fusion technique that learns modality-aware item relationships, significantly enhancing the interpretability and performance of multimodal models by effectively aligning features across different data modalities [57].

The Cross-modal Temporal Pattern Discovery (CTPD) module represents another advancement by identifying corresponding temporal patterns across modalities. This capability is crucial for capturing dynamic interactions in healthcare data, facilitating more accurate and timely clinical predictions [60].

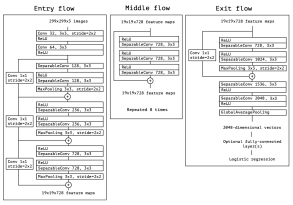
REALM exemplifies the advancements in fusion techniques by leveraging both structured and unstructured data to capture complex medical insights, thereby improving the accuracy of clinical predictions. This underscores the importance of integrating diverse data types in healthcare analytics [58].

The advanced fusion techniques discussed in recent research underscore the significant potential of innovative methodologies to enhance the integration of multimodal data, particularly in applications such as sentiment analysis, where combining visual, auditory, and textual information leads to more accurate sentiment inference and improved model performance across various tasks, including emotion recognition and public opinion analysis. [5, 13, 75, 10, 66]. By addressing the unique challenges posed by healthcare data, these approaches drive significant improvements in predictive modeling, ultimately leading to more informed clinical decision-making and personalized healthcare delivery.

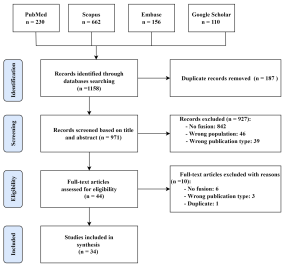
As shown in Figure 5, In the exploration of advanced modality fusion techniques, the integration of various imaging and data processing methods plays a crucial role in enhancing the precision and efficacy of medical imaging and analysis. The provided examples illustrate this integration through different approaches. The first example, "Comparison of Different Imaging Techniques for Brain Imaging," visually contrasts various imaging modalities such as PET, CT, and MRI,



(a) Comparison of Different Imaging Techniques for Brain Imaging[72]



(b) A flowchart illustrating the process of feature extraction and classification in a neural network[59]



(c) Flowchart of Study Selection Process[32]

Figure 5: Examples of Advanced Fusion Techniques

highlighting their unique contributions to brain imaging. Each modality is depicted with its distinctive image type, underscoring the importance of selecting appropriate techniques for specific diagnostic needs. The second example, "A flowchart illustrating the process of feature extraction and classification in a neural network," demonstrates the structured approach of neural networks in processing imaging data, divided into entry, middle, and exit flows, to extract meaningful features and perform accurate classifications. Lastly, the "Flowchart of Study Selection Process" provides a systematic overview of how studies are selected for synthesis, emphasizing the meticulous process of screening and eligibility assessment that filters through a substantial number of records to ensure the inclusion of relevant studies. Together, these examples underscore the complexity and sophistication of advanced fusion techniques in medical imaging and data analysis, highlighting their potential to significantly enhance diagnostic processes and outcomes. [**?**]zhou2020reviewdeeplearningmedical,kaur2021multi,mohsen2022artificialintelligencebasedmethodsfusion)

Feature	Canonical Correlation Analysis and Variants	Deep Learning Approaches	Graph-Based Models
Integration Technique	Statistical Correlation	Neural Networks	Graph Embeddings
Data Types	High-dimensional Datasets	Images And Text	Ehr And Images
Unique Feature	Semantic Integration	Predictive Accuracy	Patient Relationships

Table 5: This table provides a comparative analysis of various modality fusion techniques used in healthcare analytics, focusing on Canonical Correlation Analysis and its variants, Deep Learning Approaches, and Graph-Based Models. Each method is evaluated based on its integration technique, applicable data types, and unique features, highlighting their respective strengths in semantic integration, predictive accuracy, and patient relationship modeling.

5 Challenges and Limitations

Multimodal representation learning in healthcare is beset by several challenges that affect the efficacy of analytical models. Chief among these are data quality and completeness, which are crucial for successful predictive algorithms. This section investigates issues related to data integrity, particularly within Electronic Health Records (EHRs), and their impact on model performance and healthcare outcomes.

5.1 Data Quality and Completeness

In healthcare analytics, the quality and completeness of data are critical, as the complexity and heterogeneity of medical data can significantly impact model performance. Variability and inconsistency in EHRs often result in suboptimal feature representation and predictive accuracy. The CORE framework's effectiveness, for instance, heavily depends on the quality and granularity of EHR data, which influences cohort construction [9]. Poor data quality is a frequent limitation across methodologies, hindering model performance even with advanced fusion strategies.

Challenges arise with EHR data, particularly in free-text formats, complicating the accurate capture of patient risk levels and the development of robust models [7]. Reliance on a limited set of predefined clinical descriptors can lead to suboptimal predictions, missing valuable patient data available in EHR systems [8]. Furthermore, datasets that lack generalizability across different healthcare systems limit the applicability of predictive models. Spurious correlations within EHRs undermine data quality and completeness, complicating the integration and analysis of multimodal datasets.

The complex nature of EHRs, encompassing both structured data (clinical codes) and unstructured data (clinical notes), necessitates advanced methodologies for accurate data extraction and interpretation. Often, the synergy between these modalities is overlooked, leading to incomplete clinical insights that hinder deep learning applications [47, 76, 18]. Additionally, existing methods struggle to capture the temporal and hierarchical structure of EHR data, especially with small datasets, resulting in suboptimal model performance.

Implementing robust data governance and preprocessing techniques is essential to mitigate these challenges and enhance data quality and completeness. By improving the consistency and reliability of multimodal datasets through the integration of structured EHR data and unstructured clinical narratives, researchers can leverage advanced machine learning techniques for more accurate healthcare analytics. Enhanced analytical capabilities can facilitate early predictions of critical clinical outcomes, such as mortality risk and hospital stay length, ultimately leading to informed clinical decision-making and improved patient outcomes. Studies indicate that multimodal approaches, which effectively fuse diverse data sources, significantly outperform traditional single-modality models, underscoring the importance of comprehensive data integration in healthcare analytics [47, 32, 69, 77, 78]. Addressing data quality challenges is crucial for advancing multimodal representation learning in healthcare.

5.2 Privacy and Ethical Concerns

The integration and analysis of multimodal medical data raise significant privacy and ethical concerns due to the sensitive nature of healthcare information. The deployment of artificial intelligence (AI) in healthcare analytics introduces substantial privacy challenges, necessitating robust frameworks to safeguard patient data and ensure ethical compliance [25]. The use of large language models (LLMs) in healthcare further emphasizes the need for stringent privacy measures and bias mitigation strategies to prevent misuse of sensitive data [79].

EHRs present unique ethical challenges, as inherent biases can affect the accuracy and generalizability of predictive models. The lack of interpretability in many AI models, often functioning as black boxes, raises concerns about fairness and the generalizability of AI-driven healthcare solutions [21]. Although de-identified EHR data is commonly used to minimize privacy risks, potential breaches remain a concern if data is not managed carefully [34].

Federated learning (FL) has been proposed as a solution to privacy concerns, enabling collaborative model training without sharing raw data. However, the clinical benefits of FL compared to single-site analyses remain insufficiently evaluated, raising further privacy and ethical issues [80]. Additionally, while synthesizing high-quality synthetic data can preserve data utility, it may inadvertently increase re-identification risks, presenting a significant privacy trade-off [81].

The ethical implications of AI integration into clinical workflows extend beyond privacy, encompassing fairness in model predictions and potential biases in training data. These concerns necessitate comprehensive ethical frameworks to guide the responsible deployment of AI systems in healthcare settings [82]. The variability and sparsity of EHR data further complicate these issues, as biases can compromise the generalizability and fairness of predictive models [83].

Addressing privacy and ethical challenges requires a multifaceted approach, including advanced data governance policies, robust privacy-preserving technologies, and ethical guidelines to ensure responsible use of multimodal medical data in healthcare analytics. Limitations, such as reliance on specific imaging protocols and the need for validation across diverse populations, highlight the necessity for ongoing evaluation and adaptation of these frameworks [6]. By prioritizing privacy and ethical considerations, researchers can enhance the trustworthiness and efficacy of multimodal representation learning and modality fusion techniques in medical data analysis.

5.3 Interpretability and Model Complexity

Interpretability and model complexity pose significant challenges in multimodal learning applications within healthcare analytics, affecting the adoption and trust of these models in clinical settings. The integration of diverse data types often results in complex models that are difficult to interpret, as seen in the processing of EHRs [49]. The complexity inherent in deep learning architectures, such as those used in autoencoder-based methods, can lead to computational expense and difficulties in providing clear interpretations of model outputs [84].

Proposed multimodal ensemble approaches aim to enhance interpretability by providing meaningful insights while addressing the complexity of integrating diverse data types [85]. However, existing studies frequently fall short in fully representing healthcare phenomena, particularly in capturing temporal relationships and integrating diverse data sources [86]. The difficulty in effectively capturing temporal information may lead to ambiguous representations in certain cases [87].

The lack of interpretability in existing deep learning models is a significant challenge that frameworks like ConvLSTM1d seek to overcome by offering interpretable results [88]. However, the computational resources required for training models like CLMBR introduce additional challenges related to model complexity [33]. Furthermore, reliance on limited datasets and the lack of complexity in non-imaging data within EHRs restrict the full potential of multimodal fusion [32].

Developing interpretable data-driven models that balance complexity with transparency is essential for facilitating the integration of multimodal learning in clinical environments. Enhancing the transparency and explainability of AI models used in clinical settings can significantly bolster user trust and facilitate technology integration into clinical practice. Such improvements streamline the management of unstructured patient data, crucial for accurate diagnosis and patient care, and aid in the effective recruitment of patients for clinical trials by ensuring appropriate utilization of both structured and unstructured data. Ultimately, these advancements can lead to better patient outcomes by reducing administrative burdens on clinicians, improving patient-centered care, and ensuring critical contextual information is readily accessible for informed decision-making [81, 18, 74, 16].

5.4 Scalability and Computational Constraints

Scalability and computational constraints are vital challenges in processing multimodal data within healthcare analytics. The inherent complexity and high dimensionality of multimodal datasets often result in significant computational overhead, impeding efficient model scaling. A notable limitation is the computational complexity associated with processing large multimodal datasets, which can strain existing resources and hinder scalability [89]. The integration of diverse data modalities, including EHRs, clinical notes, and imaging data, exacerbates these challenges, necessitating sophisticated methodologies for effective computational management.

The Omni-tomography approach illustrates the computational overhead and resource limitations of current methods, posing significant obstacles to scaling with increasing data sizes [90]. This issue is compounded by the complexity of the search space in automated fusion methods, impacting computational efficiency and scalability [65]. The exponential increase in computational complexity and memory requirements associated with tensor representations presents additional challenges, particularly when scaling to multiple modalities [10].

Innovative methodologies, such as MUFASA, aim to address these constraints by reducing the computational resources required compared to traditional neural architecture search (NAS) methods, yet they still necessitate substantial computational power [31]. The collaborative graph learning method also encounters scalability challenges, particularly when clinical notes are unavailable or model complexity increases, underscoring the need for efficient data integration strategies [46].

To effectively tackle scalability and computational limitations in processing multimodal datasets, developing advanced algorithms and architectural frameworks capable of efficiently handling the complexities of high dimensionality and diverse data types—such as text, images, and other modalities—is essential, while ensuring robust performance in real-world applications [13, 17, 56, 36, 91]. By optimizing computational resources and enhancing the scalability of multimodal models, researchers can improve the efficiency and applicability of healthcare analytics, ultimately leading to more effective and scalable solutions in real-world healthcare settings.

6 Applications in Medical Data Analysis

6.1 Improving Disease Prediction and Diagnostic Accuracy

Multimodal learning techniques significantly enhance disease prediction and diagnostic accuracy by integrating diverse data sources, including structured Electronic Health Records (EHRs), clinical notes, and imaging data. These approaches capture complex relationships and patterns often overlooked by single-modal methods, thereby improving predictive performance in healthcare settings. Systematic reviews highlight that the integration of multi-source data from EHRs markedly enhances predictive outcomes in disease prediction and diagnostic accuracy [61]. Ma et al. established a state-of-the-art framework for predicting postoperative complications, exemplifying the impact of multimodal techniques on diagnostic accuracy [2]. The incorporation of Natural Language Processing (NLP) techniques enables comprehensive extraction of patient features, leading to improved diagnostic insights, as evidenced in lung cancer patient profiling [7]. The CORE framework outperforms existing methods by leveraging cohort information for enhanced EHR data representation and analysis, thus validating its efficacy in improving disease prediction [9]. Additionally, a flexible Transformer architecture demonstrated superior mean area under the receiver operating characteristic curve (AUROC) across various clinical outcome predictions, further solidifying the role of multimodal learning in enhancing diagnostic accuracy [8]. These advancements underscore the transformative impact of multimodal learning techniques on disease prediction and diagnostic accuracy, significantly improving clinical decision-making and patient outcomes. By employing advanced artificial intelligence techniques to integrate and analyze diverse modalities—including EHRs, medical imaging, and unstructured clinical narratives—these methodologies foster a deeper understanding of patient health. This comprehensive approach mitigates challenges posed by information overload and facilitates targeted insights, leading to timely and accurate diagnoses, ultimately enhancing patient-centered care [74, 16, 32, 18, 92].

6.2 Patient Management and Outcome Prediction

Multimodal learning enhances patient management and outcome prediction by integrating various data modalities, including EHRs, clinical notes, and imaging data, to create comprehensive patient profiles. This approach utilizes advanced techniques such as vector space modeling and topic modeling to extract interpretable features from complex unstructured clinical notes. By incorporating features derived from medical entities within clinical notes alongside time-series data, multimodal architectures improve predictive accuracy for critical outcomes like readmission and mortality, while providing insights validated by healthcare professionals. Recent studies demonstrate that these models outperform traditional deep learning methods, particularly in data-limited scenarios, thereby facilitating better resource management and patient care [69, 85]. The integration of diverse data sources through advanced multimodal learning techniques, such as those exemplified in the CORE framework, enhances patient cohort representation, leading to more precise clinical outcome predictions [9]. This framework improves EHR data analysis granularity, contributing to effective patient management strategies. Furthermore, transformer-based architectures have shown superior performance in predicting multiple clinical outcomes, reinforcing the potential of multimodal learning in enhancing patient management and outcome prediction [8]. NLP techniques enrich the multimodal data landscape by extracting detailed patient features from clinical notes, enabling more nuanced management and outcome predictions [7]. By integrating textual data with structured EHRs and imaging data, multimodal learning frameworks provide a holistic view of patient health, allowing healthcare providers to tailor interventions and monitor patient progress effectively. Innovative methodologies, such as the low-rank multimodal fusion method, address data integration challenges by efficiently combining diverse data types, thereby enhancing predictive accuracy in patient management models [10]. These advancements highlight the transformative potential of multimodal learning in optimizing patient management and improving outcome predictions, ultimately leading to enhanced healthcare delivery and patient satisfaction.

6.3 Case Studies and Real-World Applications

The integration of multimodal techniques in real-world applications has significantly enhanced clinical decision-making and improved patient outcomes. The HetMed framework, for instance, substantially improves the integration of multimodal medical data, thereby refining clinical decision-

making processes. Its robustness and explainability have been validated in practical applications, showcasing effectiveness in managing complex healthcare data [67]. In large-scale data processing, the Dynamic Resource Allocation Algorithm (DRAA) has been shown to reduce processing time by up to 40% compared to baseline methods, highlighting its potential as a solution for large-scale data challenges in healthcare environments [90]. The practical deployment of such algorithms underscores their capability to handle multimodal data complexities in clinical settings. The MMEDA-II architecture exemplifies the practical applications of multimodal representation learning in medical data analysis, showcasing its potential for improving healthcare outcomes through enhanced data integration and analysis [17]. The LLM-PTM method has been employed in comprehensive case studies to enhance the patient-trial matching process, improving efficiency and accuracy in clinical trial recruitment [81]. In medical imaging, non-conventional methods utilizing transform domains have demonstrated superior performance in image quality and diagnostic efficacy compared to traditional spatial domain methods [62]. Ongoing collaborations with industry partners to evaluate predictive algorithms implemented in Python and available as open-source packages further illustrate the real-world applicability of multimodal techniques in healthcare [4]. Collectively, these case studies and real-world applications underscore the transformative potential of multimodal techniques in healthcare, driving advancements in clinical decision-making, patient management, and diagnostic accuracy. By effectively integrating and analyzing diverse data modalities—such as structured EHRs, unstructured clinical narratives, medical imaging, and multi-omics data—these advanced techniques yield comprehensive insights into patient health. This multifaceted approach enhances the understanding of clinical protocols and patient pathways while addressing challenges like information overload and data fusion complexities, ultimately improving healthcare delivery and outcomes through accurate diagnoses, optimized clinical management, and personalized treatment strategies [74, 16, 32, 21, 19].

7 Future Directions

7.1 Future Research Directions

Advancing multimodal representation learning and modality fusion requires the development of sophisticated fusion techniques and the exploration of new data modalities to enhance model robustness and generalization. Integrating a variety of data types, such as combining Electronic Health Records (EHR) with imaging data, can expand the hypernetwork framework beyond traditional imaging and tabular data applications [1]. Enhancing classifier performance with complex datasets through additional features or hybrid models can significantly improve predictive accuracy in clinical settings [23].

The application of global contrastive learning frameworks to EHR modalities, including medical images and lab results, offers a promising direction for advancing multimodal representation learning [2]. Research should focus on refining fusion methods, addressing data quality challenges, and improving model generalization [5]. Moreover, incorporating mild cognitive impairment in classification tasks and applying benchmarks in longitudinal studies can provide deeper insights into neurodegenerative diseases [6].

Refining cohort construction methods and adapting frameworks like CORE for diverse healthcare tasks beyond readmission prediction could enhance healthcare analytics [9]. Enhancing Natural Language Processing (NLP) techniques and exploring additional data sources are crucial for improving patient analysis and telehealth interactions [7].

Exploring the Causal Healthcare Embedding (CHE) method in multimodal healthcare data indicates potential for improving predictive model stability and accuracy [3]. Future research should also focus on hyperparameter tuning, examining additional EHR data modalities, and enhancing clinical interpretability through self-attention distribution analyses [8].

Focusing on these research areas will drive significant advancements in developing effective, interpretable, and scalable multimodal analytics solutions. Innovations like the Categorical Sequence Encoder (CaSE), which captures the sequential nature of EHRs, demonstrate superior performance in identifying healthcare objectives compared to traditional models. Leveraging comprehensive frameworks such as MultiMed, which integrates diverse biomedical modalities and tasks, will facilitate the creation of robust AI tools that enhance disease prognosis, diagnosis, and management, ultimately leading to improved patient outcomes and healthcare delivery efficiency [86, 19, 91].

7.2 Enhancing Interpretability and Trustworthiness

Improving interpretability and trustworthiness in multimodal models is essential for clinical adoption and reliable decision-making in healthcare analytics. Attention mechanisms can offer insights into model decision processes by highlighting relevant features across modalities, thereby improving transparency and helping clinicians understand model predictions [49]. Hierarchical attention mechanisms in target-attentive diagnosis prediction models further align multimodal features, enhancing interpretability [53].

Developing interpretable data-driven models that balance complexity and transparency is crucial for integrating multimodal learning into clinical settings. Frameworks like ConvLSTM1d enhance trustworthiness by capturing temporal dependencies in patient data, providing interpretable results [88]. Visualization techniques that intuitively represent model outputs can further enhance interpretability, enabling healthcare providers to gain insights into model behavior and predictions.

Trustworthiness can be reinforced through rigorous validation processes that ensure the robustness and reliability of multimodal models. Employing cross-validation techniques and extensive testing on diverse datasets can help identify biases and improve generalizability. Additionally, incorporating ethical guidelines and privacy-preserving technologies, such as federated learning, can address data security and ethical compliance concerns, enhancing trustworthiness [80].

Engaging clinicians in the model development process and incorporating their feedback can improve interpretability and trust. Aligning model outputs with clinical expertise ensures practical applicability and acceptance of multimodal analytics solutions in healthcare. This approach leverages structured data, such as clinical codes, and unstructured data, like clinical narratives, to enhance predictive accuracy for tasks like mortality risk assessment and hospital resource management. Advanced multimodal architectures, such as the Medical Multimodal Pre-trained Language Model (MedM-PLM), effectively capture interactions between these data types, leading to superior performance in clinical tasks like medication recommendation and readmission prediction. Focusing on these integrations fosters a robust and interpretable decision-making framework in clinical settings [78, 69].

Prioritizing interpretability and trustworthiness in multimodal model development is crucial for advancing healthcare analytics. These models must effectively integrate structured and unstructured data from EHRs to enhance clinical workflows. By leveraging interactions between modalities—such as clinical codes and narratives—these models provide comprehensive insights that inform clinical decision-making, improving patient outcomes and healthcare delivery by ensuring that predictions are interpretable and meaningful to healthcare professionals. Fostering trust in these multimodal systems is essential for their successful adoption in clinical settings [78, 85].

7.3 Model Validation and Scalability

Benchmark	Size	Domain	Task Format	Metric
MedBench[26]	400	Clinical Text Classification	Text Classification	F1-score
TabDDPM[27]	46,000	Electronic Health Records	Data Generation	F1-score, MMD
EHR-BM[93]	899,128	Health Informatics	Disease Prediction	AUROC
GLoRIA[94]	1,000	Radiology	Image-Text Alignment	AUROC, Avg. P
DeCode[95]	6,829,064	Healthcare	Disease Outcome Prediction	AUPRC, AUROC
MultiMed[91]	2,560,000	Medical Imaging	Disease Classification	Accuracy, Pearson Correlation Score
EHR-CT[74]	1,000,000	Clinical Trials	Eligibility Criteria Resolution	Match Percentage
HTML[96]	1,100,000	Clinical Prediction	Risk Prediction	AUC

Table 6: Table showcasing a selection of benchmarks utilized for evaluating multimodal models in healthcare, detailing their size, domain, task format, and metrics. These benchmarks represent diverse datasets and tasks, reflecting the complexity and heterogeneity of healthcare data essential for model validation and scalability.

Model validation and scalability are critical for deploying multimodal models in healthcare, ensuring reliability and applicability across diverse clinical environments. Effective validation methods are essential for assessing model performance, identifying biases, and ensuring generalization to unseen data. Cross-validation techniques, such as k-fold cross-validation, are commonly used to evaluate model robustness and mitigate overfitting by providing comprehensive assessments of predictive accuracy across different data subsets [80]. Table 6 presents a comprehensive overview of

representative benchmarks used in the validation and scalability assessment of multimodal models within healthcare settings.

Scalability is equally important, as healthcare data is often high-dimensional and heterogeneous, requiring models to efficiently process and integrate large volumes of data from various sources. Advanced methodologies, such as efficient tensor representations and optimization algorithms, are crucial for managing the computational demands of scaling multimodal models [10]. Implementing distributed computing frameworks and cloud-based solutions can further enhance scalability by leveraging parallel processing capabilities and optimizing resource allocation [90].

Developing scalable architectures based on deep learning frameworks is essential for accommodating the increasing complexity and size of healthcare datasets. These architectures must handle the integration of diverse data modalities, including structured EHRs, unstructured clinical notes, and imaging data, while maintaining computational efficiency and performance [89]. Techniques like the Dynamic Resource Allocation Algorithm (DRAA) have shown significant improvements in processing efficiency, reducing computational overhead, and enhancing the scalability of multimodal systems [90].

Ensuring scalability and validation also involves addressing data quality and completeness challenges, as these factors significantly impact model performance and generalizability. Robust data preprocessing and integration strategies are essential for maintaining data integrity and ensuring that models can effectively leverage the full spectrum of available healthcare data [9].

By emphasizing model validation and scalability, researchers and practitioners can significantly improve the reliability and applicability of multimodal models in healthcare. This approach fosters the integration of diverse data types—such as structured clinical codes and unstructured clinical narratives—and enhances predictive capabilities for critical clinical tasks like medication recommendation and readmission prediction. Frameworks like MedM-PLM and EMERGE demonstrate that effective multimodal models can leverage interactions between various data modalities to provide comprehensive insights, ultimately leading to more effective and scalable solutions that enhance patient outcomes and streamline healthcare delivery [69, 34, 85, 78, 91].

8 Conclusion

The exploration of multimodal representation learning and modality fusion in medical data analysis reveals their substantial impact on advancing healthcare outcomes. By synthesizing structured Electronic Health Records, clinical notes, and imaging data, these methodologies provide a comprehensive view of patient health, thereby enhancing predictive accuracy and clinical decision-making. The application of sophisticated computational techniques, including deep learning and graph-based models, facilitates the integration of complex datasets, improving the scalability and adaptability of healthcare analytics.

The integration of information technology within healthcare underscores the potential of these methodologies to address challenges associated with data heterogeneity and high dimensionality. Continuous research is vital to develop integrated solutions that enhance the clinical applicability and outcomes of deep learning models, particularly with temporal EHR data. Furthermore, the development of innovative methodologies that minimize preprocessing requirements enables the use of simpler models, broadening accessibility for researchers and practitioners working with EHR data.

As the field progresses, addressing challenges related to data quality, interpretability, and privacy is imperative to fully harness the potential of multimodal representation learning in healthcare. Overcoming these issues is essential for achieving significant advancements in disease prediction, patient management, and healthcare delivery, ultimately leading to improved patient outcomes and more robust healthcare systems.

References

[1] Daniel Duenias, Brennan Nichyporuk, Tal Arbel, and Tammy Riklin Raviv. Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling, 2025.

[2] Yingbo Ma, Suraj Kolla, Zhenhong Hu, Dhruv Kaliraman, Victoria Nolan, Ziyuan Guan, Yuanfang Ren, Brooke Armfield, Tezcan Ozrazgat-Baslanti, Jeremy A. Balch, Tyler J. Loftus, Parisa Rashidi, Azra Bihorac, and Benjamin Shickel. Global contrastive training for multimodal electronic health records with language supervision, 2024.

[3] Yingtao Luo, Zhaocheng Liu, and Qiang Liu. Deep stable representation learning on electronic health records, 2022.

[4] Zichang Wang, Haoran Li, Luchen Liu, Haoxian Wu, and Ming Zhang. Predictive multi-level patient representations from electronic health records, 2019.

[5] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023.

[6] Arjun Punjabi, Adam Martersteck, Yanran Wang, Todd B. Parrish, Aggelos K. Katsaggelos, and the Alzheimer's Disease Neuroimaging Initiative. Neuroimaging modality fusion in alzheimer's classification using convolutional neural networks, 2018.

[7] Ernestina Menasalvas Ruiz, Juan Manuel Tuñas, Guzmán Bermejo, Consuelo Gonzalo Martín, Alejandro Rodríguez-González, Massimiliano Zanin, Cristina González de Pedro, Marta Mendez, Olga Zaretskaia, Jesús Rey, Consuelo Parejo, Juan Luis Cruz Bermudez, and Mariano Provencio. Profiling lung cancer patients using electronic health records, 2018.

[8] Benjamin Shickel, Patrick J. Tighe, Azra Bihorac, and Parisa Rashidi. Multi-task prediction of clinical outcomes in the intensive care unit using flexible multimodal transformers, 2021.

[9] Changshuo Liu, Wenqiao Zhang, Beng Chin Ooi, James Wei Luen Yip, Lingze Zeng, and Kaiping Zheng. Toward cohort intelligence: A universal cohort representation learning framework for electronic health record analysis, 2023.

[10] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[11] Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports, 2020.

[12] Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. Heteromed: Heterogeneous information network for medical diagnosis, 2018.

[13] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2020.

[14] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports, 2018.

[15] Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H Gandomi. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144:105253, 2022.

[16] Chanseo Lee, Kimon-Aristotelis Vogt, and Sonu Kumar. Assessing the role of clinical summarization and patient chart review within communications, medical management, and diagnostics, 2024.

[17] Aishwarya Jayagopal, Ankireddy Monica Aiswarya, Ankita Garg, and Srinivasan Kolumam Nandakumar. Multimodal representation learning with text and images, 2022.

[18] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. Query-focused ehr summarization to aid imaging diagnosis, 2020.

[19] Adrian Caruana, Madhushi Bandara, Daniel Catchpoole, and Paul J Kennedy. Beyond topics: Discovering latent healthcare objectives from event sequences, 2021.

[20] Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, and David R. Karger. Medknowts: Unified documentation and information retrieval for electronic health records, 2021.

[21] Wei-Hung Weng and Peter Szolovits. Representation learning for electronic health records, 2019.

[22] Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review, 2020.

[23] Efsun Sarioglu Kayi, Kabir Yadav, James M. Chamberlain, and Hyeong-Ah Choi. Topic modeling for classification of clinical reports, 2017.

[24] Shipu Debnath. Integrating information technology in healthcare: Recent developments, challenges, and future prospects for urban and regional health, 2023.

[25] Gloria Hyunjung Kwak and Pan Hui. Deephealth: Review and challenges of artificial intelligence in health informatics, 2020.

[26] Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. Comparative analysis of text classification approaches in electronic health records, 2020.

[27] Taha Ceritli, Ghadeer O. Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P. Creagh, and David A. Clifton. Synthesizing mixed-type electronic health records using diffusion models, 2023.

[28] Daisy Yi Ding, Chloé Simpson, Stephen Pfohl, Dave C. Kale, Kenneth Jung, and Nigam H. Shah. The effectiveness of multitask learning for phenotyping with electronic health records data, 2019.

[29] Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics, 2019.

[30] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[31] Zhen Xu, David R. So, and Andrew M. Dai. Mufasa: Multimodal fusion architecture search for electronic health records, 2021.

[32] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data, 2022.

[33] Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective patient representation learning technique for electronic health record data, 2020.

[34] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation, 2025.

[35] Tiantian Feng, Daniel Yang, Digbalay Bose, and Shrikanth Narayanan. Can text-to-image model assist multi-modal learning for visual recognition with visual modality missing?, 2024.

[36] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[37] Jacek M. Bajor, Diego A. Mesa, Travis J. Osterman, and Thomas A. Lasko. Embedding complexity in the data representation instead of in the model: A case study using heterogeneous medical data, 2025.

[38] Saeed Shurrab, Alejandro Guerra-Manzanares, and Farah E. Shamout. Multi-modal masked siamese network improves chest x-ray representation learning, 2024.

[39] Xingrui Gu, Zhixuan Wang, Irisa Jin, and Zekun Wu. Advancing multimodal data fusion in pain recognition: A strategy leveraging statistical correlation and human-centered perspectives, 2024.

[40] Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. Learning inter-modal correspondence and phenotypes from multi-modal electronic health records, 2020.

[41] Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. Hierarchical pretraining on multimodal electronic health records, 2023.

[42] Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Young-Hak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding, 2022.

[43] Cheng Jiang, Yihao Chen, Jianbo Chang, Ming Feng, Renzhi Wang, and Jianhua Yao. Fusion of medical imaging and electronic health records with attention and multi-head machanisms, 2021.

[44] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning, 2020.

[45] João Figueira Silva and Sérgio Matos. Modelling patient trajectories using multimodal information, 2022.

[46] Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare, 2021.

[47] Tianxi Cai, Feiqing Huang, Ryumei Nakada, Linjun Zhang, and Doudou Zhou. Contrastive learning on multimodal analysis of electronic health records, 2024.

[48] Chun-Chieh Liao, Wei-Ting Kuo, I-Hsuan Hu, Yen-Chen Shih, Jun-En Ding, Feng Liu, and Fang-Ming Hung. Ehr-based mobile and web platform for chronic disease risk prediction using large language multimodal models, 2024.

[49] Thanh Nguyen-Duc, Natasha Mulligan, Gurdeep S. Mannu, and Joao H. Bettencourt-Silva. Deep ehr spotlight: a framework and mechanism to highlight events in electronic health records for explainable predictions, 2021.

[50] Prasun C Tripathi, Sina Tabakhi, Mohammod N I Suvon, Lawrence Schöb, Samer Alabed, Andrew J Swift, Shuo Zhou, and Haiping Lu. Interpretable multimodal learning for cardiovascular hemodynamics assessment, 2024.

[51] Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, and Weimin Li. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature biomedical engineering*, 7(6):743–755, 2023.

[52] Shohreh Deldari, Hao Xue, Aaqib Saeed, Jiayuan He, Daniel V Smith, and Flora D Salim. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint arXiv:2206.02353*, 2022.

[53] Anahita Hosseini, Tyler Davis, and Majid Sarrafzadeh. Hierarchical target-attentive diagnosis prediction in heterogeneous information networks, 2019.

[54] Kwanhyung Lee, John Won, Heejung Hyun, Sangchul Hahn, Edward Choi, and Joohyung Lee. Self-supervised predictive coding with multimodal fusion for patient deterioration prediction in fine-grained time resolution, 2023.

[55] Tom van Sonsbeek, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational knowledge distillation for disease classification in chest x-rays, 2021.

[56] Zida Cheng, Chen Ju, Shuai Xiao, Xu Chen, Zhonghua Zhai, Xiaoyi Zeng, Weilin Huang, and Junchi Yan. Category-oriented representation learning for image to multi-modal retrieval, 2024.

[57] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. Latent structure mining with contrastive modality fusion for multimedia recommendation, 2022.

[58] Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and Chengwei Pan. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models, 2024.

[59] Manjit Kaur and Dilbag Singh. Multi-modality medical image fusion technique using multi-objective differential evolution based deep neural networks. *Journal of Ambient Intelligence and Humanized Computing*, 12(2):2483–2493, 2021.

[60] Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. Ctpd: Cross-modal temporal pattern discovery for enhanced multimodal electronic health records analysis, 2024.

[61] Norman Hiob and Stefan Lessmann. Health analytics: a systematic review of approaches to detect phenotype cohorts using electronic health records, 2017.

[62] Manoj Diwakar, Prabhishek Singh, Vinayakumar Ravi, and Ankur Maurya. A non-conventional review on multi-modality-based medical image fusion. *Diagnostics*, 13(5):820, 2023.

[63] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, Kia Khezeli, and Parisa Rashidi. Transformers in healthcare: A survey, 2023.

[64] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

[65] Suhan Cui, Jiaqi Wang, Yuan Zhong, Han Liu, Ting Wang, and Fenglong Ma. Automated fusion of multimodal electronic health records for better medical predictions, 2024.

[66] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[67] Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. Heterogeneous graph learning for multi-modal medical data analysis, 2023.

[68] Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. Graph-text multi-modal pre-training for medical representation learning, 2022.

[69] Batuhan Bardak and Mehmet Tan. Improving clinical outcome predictions using convolution over medical entities with multimodal learning, 2020.

[70] Yan Miao and Lequan Yu. Must: Multimodal spatiotemporal graph-transformer for hospital readmission prediction, 2023.

[71] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[72] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion, 2020.

[73] Naman Goyal. A survey on self supervised learning approaches for improving multimodal representation learning. *arXiv preprint arXiv:2210.11024*, 2022.

[74] Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?, 2015.

[75] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. Complementary fusion of multi-features and multi-modalities in sentiment analysis, 2019.

[76] Bo Yang and Lijun Wu. How to leverage multimodal ehr data for better medical predictions?, 2021.

[77] Ziyi Liu, Jiaqi Zhang, Yongshuai Hou, Xinran Zhang, Ge Li, and Yang Xiang. Machine learning for multimodal electronic health records-based research: Challenges and perspectives, 2021.

[78] Sicen Liu, Xiaolong Wang, Yongshuai Hou, Ge Li, Hui Wang, Hui Xu, Yang Xiang, and Buzhou Tang. Multimodal data matters: language model pre-training over structured and unstructured electronic health records, 2022.

[79] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L. Assimes, Libby Hemphill, and Siyuan Ma. A scoping review of using large language models (llms) to investigate electronic health records (ehrs), 2024.

[80] Siqi Li, Pinyan Liu, Gustavo G. Nascimento, Xinru Wang, Fabio Renato Manzolli Leite, Bibhas Chakraborty, Chuan Hong, Yilin Ning, Feng Xie, Zhen Ling Teo, Daniel Shu Wei Ting, Hamed Haddadi, Marcus Eng Hock Ong, Marco Aurélio Peres, and Nan Liu. Federated and distributed learning applications for electronic health records and structured medical data: A scoping review, 2023.

[81] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large language models for healthcare data augmentation: An example on patient-trial matching, 2023.

[82] Riyi Qiu, Yugang Jia, Mirsad Hadzikadic, Michael Dulin, Xi Niu, and Xin Wang. Modeling the uncertainty in electronic health records: a bayesian deep learning approach, 2019.

[83] Wei Liao and Joel Voldman. Learning and disentangling patient static information from time-series electronic health record (steer), 2023.

[84] Najibesadat Sadati, Milad Zafar Nezhad, Ratna Babu Chinnam, and Dongxiao Zhu. Representation learning with autoencoders for electronic health records: A comparative study, 2019.

[85] Bonggun Shin, Julien Hogan, Andrew B. Adams, Raymond J. Lynch, Rachel E. Patzer, and Jinho D. Choi. Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission, 2019.

[86] Keith Feldman, Louis Faust, Xian Wu, Chao Huang, and Nitesh V. Chawla. Beyond volume: The impact of complex healthcare data on the machine learning pipeline, 2018.

[87] Hao-Ren Yao, Nairen Cao, Katina Russell, Der-Chen Chang, Ophir Frieder, and Jeremy Fineman. Self-supervised representation learning on electronic health records with graph kernel infomax, 2024.

[88] Fabio Azzalini, Tommaso Dolci, and Marco Vagaggini. An interpretable deep-learning framework for predicting hospital readmissions from electronic health records, 2023.

[89] Yingbo Ma, Suraj Kolla, Dhruv Kaliraman, Victoria Nolan, Zhenhong Hu, Ziyuan Guan, Yuanfang Ren, Brooke Armfield, Tezcan Ozrazgat-Baslanti, Tyler J. Loftus, Parisa Rashidi, Azra Bihorac, and Benjamin Shickel. Temporal cross-attention for dynamic embedding and tokenization of multimodal electronic health records, 2024.

[90] Ge Wang, Jie Zhang, Hao Gao, Victor Weir, Hengyong Yu, Wenxiang Cong, Xiaochen Xu, Haiou Shen, James Bennett, Yue Wang, and Michael Vannier. Omni-tomography/multi-tomography – integrating multiple modalities for simultaneous imaging, 2011.

[91] Shentong Mo and Paul Pu Liang. Multimed: Massively multimodal and multitask medical understanding, 2024.

[92] Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, Ph. D, and Kathryn Turner. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review.

[93] William La Cava, Christopher Bauer, Jason H. Moore, and Sarah A Pendergrass. Interpretation of machine learning predictions for patient outcomes in electronic health records, 2019.

[94] Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C. Wallace. That's the wrong lung! evaluating and improving the interpretability of unsupervised multimodal encoders for medical data, 2022.

[95] Zhichao Yang, Weisong Liu, Dan Berlowitz, and Hong Yu. Enhancing the prediction of disease outcomes using electronic health records and pretrained deep learning models, 2022.

[96] Ross S. Kleiman, Paul S. Bennett, Peggy L. Peissig, Richard L. Berg, Zhaobin Kuang, Scott J. Hebbring, Michael D. Caldwell, and David Page. High-throughput machine learning from electronic health records, 2019.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.