

Multimodal Learning in Healthcare

Saaliha Allauddin Khan Ghori

CS 6220 Big Data and Analysis

December 05, 2024

In recent years, the healthcare industry has witnessed a major shift towards data-driven decision making by using artificial intelligence (AI) and machine learning (ML) which played increasingly pivotal roles. Machine learning and Artificial intelligence have led to many advancements such as deep learning, neural networks, reinforcement learning, etc. which are used to analyze large datasets and improve efficiency and accuracy in the decision-making processes. Among these advancements, multimodal learning has become one of the most essential and groundbreaking approaches. Multimodal learning in healthcare is an emerging field that uses different data types such as images, texts like electronic health records (EHRs), physiological signals, and genomic information to develop artificial intelligence (AI) systems that are capable of comprehensive decision-making (Yan et al., 2023).

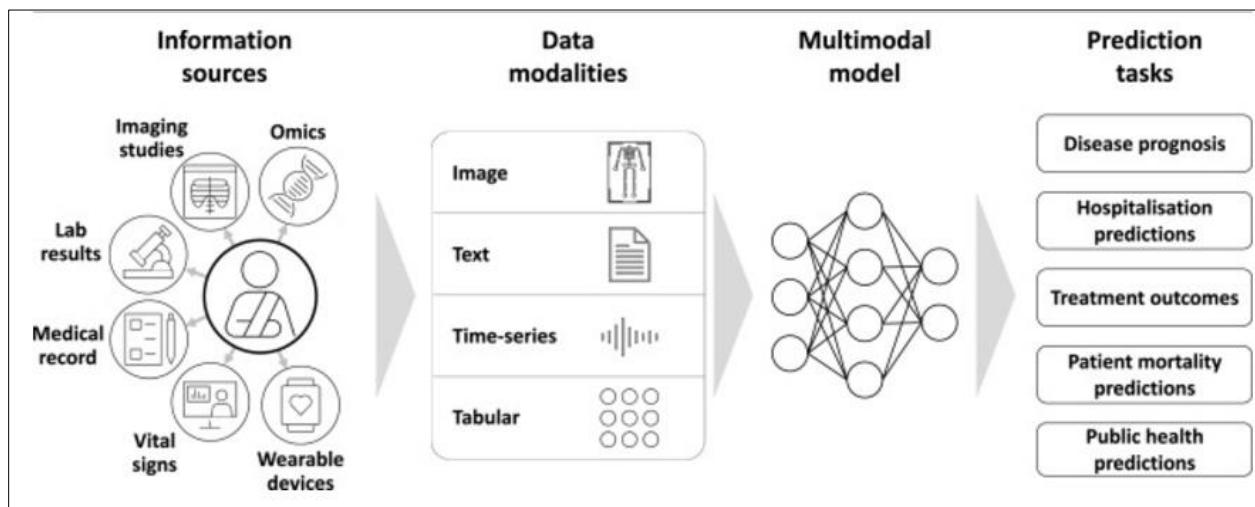


Figure 1: Clinical data modalities and prediction tasks (Krones et al., 2024)

For a long time, traditional medical diagnostics have been constrained by the limitations of single-modal data analysis as they rely on single information sources such as medical imaging, clinical notes, or electronic health records. On the other hand, multimodal learning represents a significant shift, enabling healthcare professionals to benefit from a holistic view of patient data by incorporating multiple information streams simultaneously (Krones et al., 2024). Just as a clinician considers a patient's symptoms, medical history, test results, and imaging studies to provide proper and personalized

diagnosis, AI systems with multimodal learning use various sources and synthesize information and provides more accurate and personalized healthcare solutions. As we can see in figure 1, information sources are transformed to different modalities which are then sent to the multimodal learning model that assists and provides prediction responses to the questions. By integrating different data sources, multimodal learning has the potential to improve diagnostic accuracy, provided personalized treatment strategies, and uncover novel insights into disease mechanisms (Baltrusaitis et al., 2019).

The significance of multimodal learning in healthcare cannot be overstated. Recent studies have shown its potential to dramatically improve the field by providing the facilities and outcomes like healthcare professionals. For instance, researchers have developed machine learning models that combine medical imaging, patient history, and genetic information to achieve unexpected levels of diagnostic precisions for complex conditions like cancer, neurological disorders, and rare genetic diseases (Wang et al., 2022). Also, this multimodal approach has the potential to detect diseases and enable more precise interventions which will help patients' outcomes because medical practitioners will be able to predict the disease and treat accordingly. These advancements bridge the gap between AI research and resulting models and its practical implementation in clinical environments.

This technical review aims to explore and discuss the current state of multimodal learning in healthcare, examining its technological foundations, current research landscape and developments, potential applications, challenges, and future implications. We will be critically analyzing state-of-the-art methodologies and techniques, and emerging trends and how they advance and assist the healthcare field. This review is particularly valuable to any readers who are interested in learning about the intersection of AI, specifically multimodal learning and healthcare, whether they are computer scientists, healthcare practitioners, students, or policy makers. This technical review is aimed at serving as a comprehensive resource.

Background and History

Before diving into the state of the arts research and developments, let's look at the background, existing healthcare technologies, and learn what multimodal learning means. To begin with, the integration of artificial intelligence in healthcare has a long and rich history dating back to the 1950s and the concept of "machine learning" was researched on. The journey of AI/ML in medicine has been marked by both progress and setbacks. The efforts to integrate AI into healthcare began in the 1960s and 1970s.

One of the significant milestones was the development of MYCIN during the early 1970s. MYCIN is a rule-based system which used rule-based learning to identify bacterial infections and recommend appropriate antibiotics depending on the patient's information (Kaul et al., 2020). It asked a series of questions to the physician or health practitioner and reach a conclusion based on the inputs. Even though it was a promising technology, MYCIN and other similar models were never implemented in clinical settings due to ethical considerations and concerns about the computer-based recommendations and liability issues (Practice Accelerator, 2023). The next major developments began to take place in the early 2000s with the help of the deep learning models (Keragon, 2024). Widespread adoption of electronic health records (EHRs) begun within the healthcare facilities which captured the large clinical datasets essential for improving AI's accuracy in healthcare decision-making.

Today, AI in healthcare has expanded beyond its origins to influence various medical specialties, including radiology, psychiatry, primary care, and telemedicine (Xsolis). These technologies have shown improving accuracy, enhancing workflow efficiency, and developing sophisticated risk assessment models (Kaul et al., 2020). We stand on the cusp of the new era in medicine as multimodal learning emerges and serves as a powerful approach to integrate diverse data types such as EHRs, medical imaging, genomic data, and wearable sensor information—to provide comprehensive insights about patient health. This approach addresses the limitations of traditional unimodal methods by having the ability to capture different data types and full complexity of human health and disease, mimicking the decision-making processes of healthcare professionals.

What is Multimodal Learning?

Multimodal learning is an approach in machine learning that integrates data from multiple sources and modalities to enhance decision-making and prediction accuracy. As we have seen in the previous paragraphs, each modality represents a distinct type of data such as textual descriptions, medical images, genomic sequences, or physiological signals, that contribute unique insights into a task. The core idea is to combine information from these diverse modalities and provide a more comprehensive understanding of complex problems, specifically in domains like healthcare where data richness and heterogeneity are crucial (Baltrusaitis et al., 2019).

In healthcare, multimodal learning integrates the different modalities mentioned above to reach a conclusion. For instance, Alzheimer's disease diagnosis and management benefit from combining brain imaging, cognitive test results, and genetic information. This integration helps health practitioners/professionals reach a more robust prediction of the disease onset and progression, and potentially guiding them to intervene

at the earlier stages. These applications highlight the importance of multimodal learning in tackling the most important issues facing healthcare, such as preventive care and tailored medication (Esteva et al., 2021).

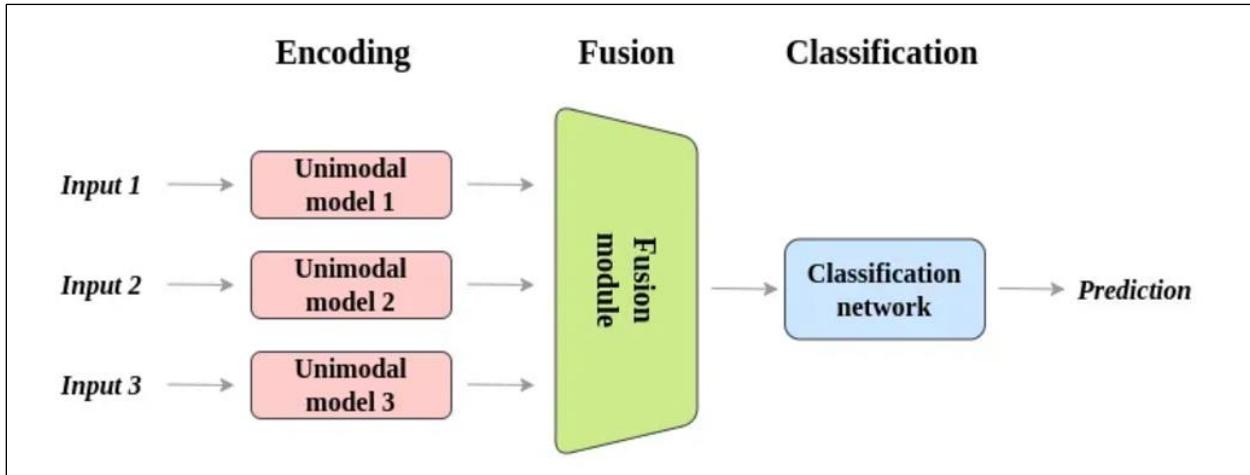


Figure 2: A generated multimodal workflow (Potrimba, 2023).

Now, let's understand how multimodal learning models work. These models are typically composed of “three unimodal neural networks, which process each input modality separately” (Potrimba, 2023). For instance, one of these unimodal networks will process medical images. This individual processing of each modality is known as encoding. Once unimodal encoding is completed, the extracted information from each modality needs to be fused or integrated (Potrimba, 2023). As we can note from figure 2, multimodal architecture consists of three parts:

1. Encoding: for each input modality, a unimodal encoder is assigned to encode them.
2. Fusion: a fusion network that, during the encoding stage, integrates the characteristics taken from each input modality.
3. Classification: A classifier that generates predictions after receiving the fused data.

Multimodal Learning Applications

We have learned how multimodal learning works and investigated the AI integrated healthcare technologies. Moving forward, we will look at some of the applications of multimodal learning evolving from early studies in early studies in audio-visual speech recognition (AVSR) to contemporary advancements in language and vision tasks, and how these technologies assist in healthcare. AVSR, one of the earliest applications of multimodal learning, was inspired by McGurk effect, where hearing and visual inputs interact during speech perception. Being one of the earliest applications, early AVSR models were based on hidden Markov models, aimed to enhance speech recognition in

noisy environments by supplementing auditory data with visual cues did not have significant performance improvement in noiseless conditions (Baltrušaitis, 2017).



Figure 3: Example of Visual question answering (VQA) (Goyal et al., 2017)

Research shifted from keyword-based approaches to analyzing visual and multimodal content directly with the help of innovations like automatic shot-boundary detection and video summarization supported by initiatives such as TrecVid. The early 2000s marked as rise of the multimodal interaction studies where researchers focused on understanding human behavior during social interactions (Baltrušaitis, 2017). With the use of AMI Meeting Corpus and SEMAINE corpus, researchers were able to work on emotional recognition and affective computing, which led to advances in facial recognition technologies and development of depression and anxiety assessment as healthcare applications. In recent time we have image captioning, which helps visually impaired users by providing textual descriptions for images. Also, we have been introduced to Visual question answering (VQA) to address challenges in evaluating the quality of generated descriptions (Baltrušaitis, 2017). In figure 3, we can see how this model works and correctly identifies the answer based on the images. This highlights the ongoing evolution and significance of multimodal learning.

The Importance of Multimodal learning in the healthcare sector, AVSR can be applied to enhance speech recognition in a noisy environment, thus forming a powerful communication tool for people with hearing impairments or speech disorders. For example, multimodal affect recognition, using data from facial expressions, voice and physiological signals can help in diagnosing and monitoring of mental illnesses like depression or anxiety. Newer developments in recognizing emotion could specifically be

useful in building empathic virtual health assistants and applications for emotional well-being management.

Image captioning or visual question answering (VQA), for example, are media description tasks that are transforming accessibility and assisting blind patients with their context such as moving around their surroundings or understanding a medical imaging scene. Besides, the classification and extraction methods for these multimedia contents will make accessing large healthcare databases easier to manage so this can extend to files such as patient records, video diagnostics, and training materials for doctors.

Key Technological Aspects

As we have seen in the above sections, multimodal learning in healthcare involves the integration of diverse data types to build robust models capable of addressing the complexity of medical decision-making. In this section, several key technological aspects will be discussed including data fusion techniques, explainability, low-resource learning, and the emergence of novel paradigms such as self-supervised multimodal learning. It is necessary to learn about these topics to better understand multimodal learning. These advancements have the potential to transform healthcare by improving diagnostic accuracy, enhancing patient care, and enabling innovative applications.

1. Data Fusion Techniques

Data fusion is central to multimodal learning, as it assists in determining how the information from different modalities will be combined. As we learned in the workflow of the multimodal learning models, “Fusion” is the second step after “Encoding.” So, it is essential to learn about the three main approaches: early fusion, intermediate fusion, and late fusion.

- **Early fusion** involves the concatenation of raw data or feature representations from multiple input modalities before training a single machine learning model (see figure 4(a)). The data may need to go through several feature extraction procedures or can be used in its unprocessed state. This can include using distinct models or straightforward aggregation techniques. The particular modalities and models used also influence the feature combination technique (Krones et al., 2024). For instance, image data requires them to be stacked as channels in a Convolutional Neural Network (CNN) framework (Taleb et al., 2021) while time-series data needs to be aggregated with models like XGBoost before use (Krones et al., 2022).

Disadvantages of early fusion include challenges in balancing “data richness” of different modalities, as vision data often requires more processing than language

data, which leads to disproportionate attention to one modality. In addition, modality-specific feature encoding may be required because low-level features from one modality—such word embeddings in language—may not be semantically compatible with features from another, like edges in images. Early fusion approaches are inflexible as they require complete model retraining if new data sources or modalities are incorporated (Krones et al., 2024).

Early fusion has been used to combine MRI scans and patient demographics at the input stage to create richer feature spaces (Ramachandran et al., 2017). Moreover, as mentioned by Krones et al., early fusion is “implemented in clinical machine learning across multiple settings, such as for predictive tasks in cardiological, oncological and neurological domains” and it is used to combine CT and PET scan features with demographic information to diagnose lung cancer and combine MRI and PET pictures with genetic and demographic information to predict Alzheimer’s disease.

- **Intermediate fusion** integrates information at hidden layers of a neural network allowing the model to learn modality-specific features before merging them. To elaborate, prior to the extracted features being integrated and supplied into a final prediction model, various data modalities are initially analyzed by individual models (see Fig. 4(b)). Here, in contrast to early fusion, the loss function is backpropagated via the feature extraction model to produce better feature representations with each training cycle (Krones et al., 2024).

Intermediate fusion has multiple advantages, which include flexibility in the model architecture, which helps preserve modality-specific information by enabling individual processing pathways for different data types before fusion. Also, the benefit of intermediate fusion is that since each modality is first transformed into a machine-understandable representation, the model may learn complex interconnections between them. The downside of this approach is the increase in processing time which impacts inference speed because of individual processing of each modality before fusion (Pulapakura, 2024).

This approach is common in disease diagnosis models, such as combining text-based pathology reports with imaging data to detect cancer (Baltrusaitis et al., 2019). Intermediate fusion is predominantly used for disease predictions. For instance, pathological images and demographic and genomics data are combined for cancer prediction, X-ray images, demographic information, biomarkers and clinical measurements are combined to predict cardiology related diseases, and

finally MRI images and demographic information are combined to predict brain disorders (Krones at al., 2024).

- **Late fusion** acts like an ensemble model. It aggregates predictions made independently by unimodal models. Different models are applied to different data modalities (raw or extracted features), and an auxiliary function or aggregation function is used to combine the predictions that are produced (see fig. 4(c)). This approach has its own advantages and disadvantages. One advantage is that it can deal with missing data for patients which helps train better and output predictions. On the other hand, the disadvantage is that with late fusion it is impossible to predict the linkages and interactions between various modalities with late fusion, which could result in information loss (Krones at al., 2024).

Within the healthcare field, this approach is useful when modalities are asynchronous or have different data availability, such as wearable sensor data and periodic clinical lab results. Also, they are used in cancer predictions by combining MRI images and biomarkers, COVID predictions by combining CT scans, clinical measurements and demographic information, and finally cognitive impairment by combining MRI image predictions and cognitive assessment scores ().

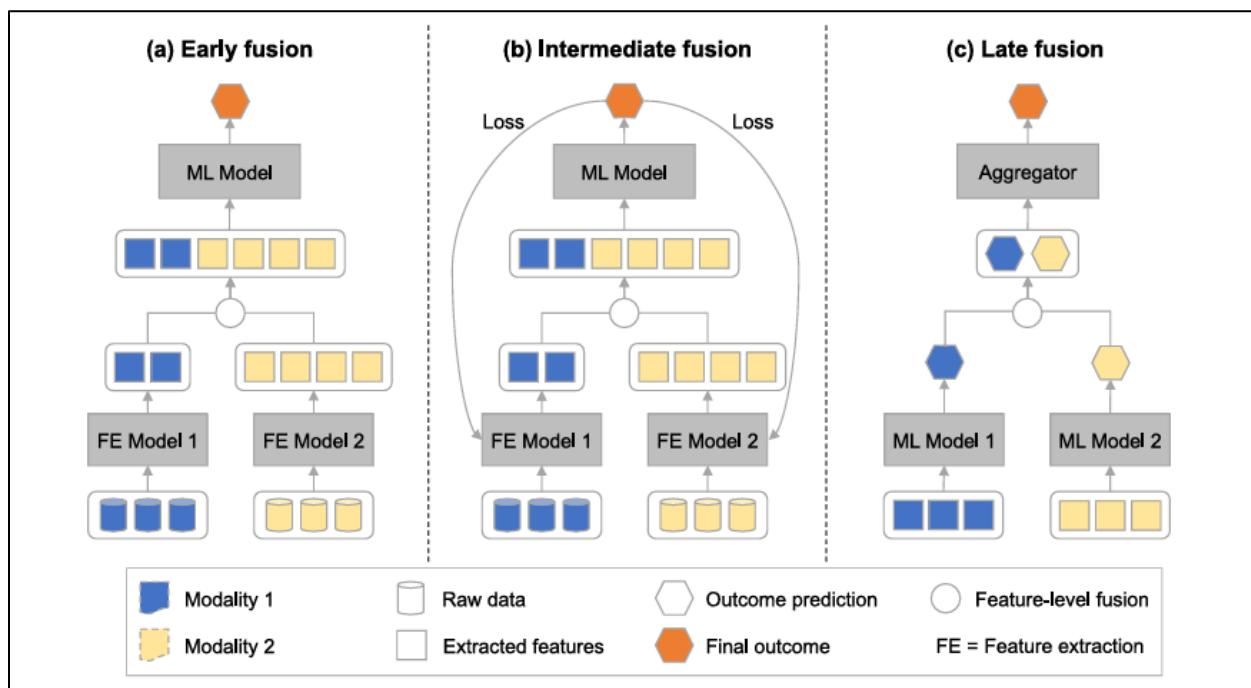


Figure 4: The three different fusion approaches (Krones at al., 2024).

2. Explainability in Multimodal Models

Explainability is the concept of being able to explain a machine learning model's output in a way that is easier for human beings to understand. It is crucial in healthcare, where clinicians must be able to understand the AI model's insights and how it reached the decisions. Models developed using multimodal learning face unique challenges because of its complex interactions between modalities. Techniques such as attention mechanisms and saliency maps have been adapted to highlight which modality or features contribute most to a prediction (Rajkomar et al., 2018). For example, an explainable model might demonstrate how fundus images and patient history work together to diagnose retinal illnesses, promoting usability and trust among medical professionals.

3. Low-Resource Learning

Healthcare datasets often suffer from limited labeled samples due to privacy concerns, the cost of annotation, and data scarcity in rare diseases. This restriction is overcome by low-resource learning techniques like few-shot and zero-shot learning, which make use of auxiliary data or previously trained models. For instance, MedCLIP, a specialized adaptation of the CLIP (Contrastive Language–Image Pretraining) model for medical applications, designed to bridge the gap between visual and textual modalities in healthcare. As mentioned above, MedCLIP uses zero-shot learning, cross-modal retrieval, and self-supervised pretraining, which helped disease diagnosis, clinical decision support, research and development (Wang et al., 2022). These techniques greatly lessen reliance on sizable, labeled datasets, increasing the viability of multimodal learning in practical contexts.

4. Emerging Paradigms: Self-Supervised Learning

A promising paradigm for multimodal healthcare applications is self-supervised learning (SSL). By predicting missing or masked information, SSL uses vast amounts of unlabeled data to learn meaningful representations. For example, by training models to identify pathology reports with associated radiological scans, multimodal SSL frameworks can align imaging and text modalities (Radford et al., 2021). This method lessens the need for labeled data, which is frequently a bottleneck in the development of medical AI, while simultaneously enhancing model performance.

State of the Art Research and Development

Recent advancements in multimodal learning have demonstrated transformative potential in healthcare. We can note that major research and development have occurred

in the last decade. Researchers have explored various frameworks and methodologies to integrate multimodal data in variety of healthcare fields like oncology, Alzheimer's disease, mental health, etc. to enhance diagnostic accuracy, treatment planning, and patient care. Below, we will discuss three key state-of-the-art developments in this domain.

1. Multimodal learning approach for Precision Oncology

The researchers have provided a thorough review of the integration of multimodal data in precision oncology in the research paper "Multimodal Data Integration for Precision Oncology: Challenges and Future Directions," addressing established areas, current challenges, and upcoming opportunities. The goal of precision medicine is to provide individualized care based on the distinct features of every tumor. The diversity of tumor data necessitates the integration of several information types, including imaging, clinical, and omics data. By improving clinical procedures and tailored treatments, this integration contributes to a more accurate understanding of tumor dynamics. The study highlights applications in biomarker creation, prognosis prediction, and diagnostics by analyzing 300 research articles published over a ten-year period.

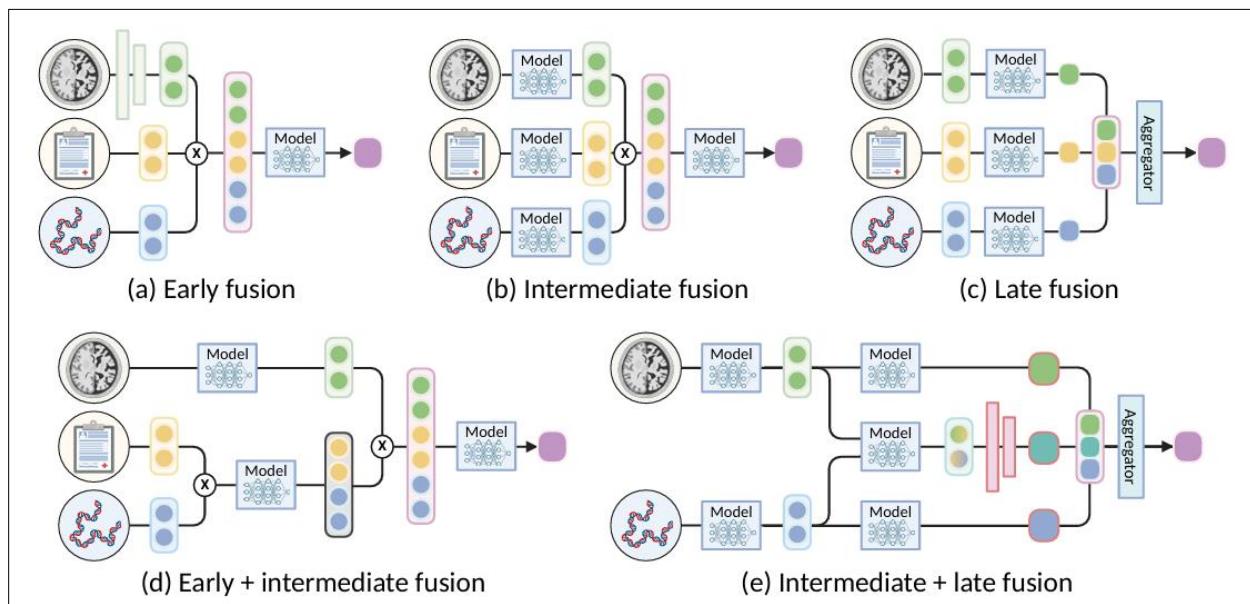


Figure 5. Fusion strategies for complete data (Zhou et al., 2024).

The researchers have discussed the benefits and drawbacks of the multimodal data integration approaches and divided them into strategies for complete and incomplete data. When dealing with complete data, they investigate techniques like early, intermediate, late, and multi-level fusion (see figure 5), each of which offers unique approaches to modeling the connections between various modalities while addressing problems like modality collapse. The study examines two approaches for incomplete data: imputation-free

approaches that focus on robustness and knowledge distillation without replacing the missing data, and imputation-based approaches that either generate or source the missing modalities (see figure 6). The advantages and disadvantages of both approaches are assessed in clinical contexts, emphasizing the need for tailored solutions depending on the information and tasks at hand (Zhou et al., 2024).

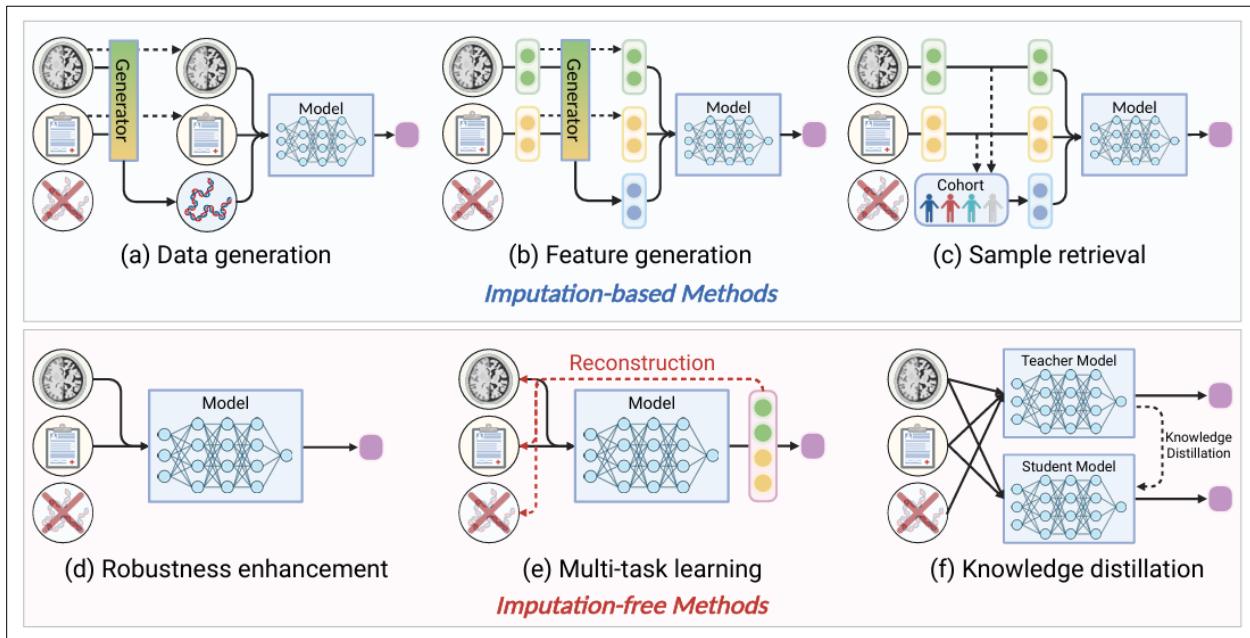


Figure 6. Fusion strategies for incomplete data, including imputation-based methods and imputation-free methods (Zhou et al., 2024).

Using these different techniques, the paper discusses and emphasizes the important role of multimodal data integration in various clinical applications like early assessment, diagnosis, prognosis, and biomarker discovery. In early assessment, by using these techniques the researchers have noted that it enhances cancer risk stratification by leveraging data from diverse sources such as medical history and imaging. This outperforms traditional methods in predictive accuracy. In diagnosis, these multimodal learning techniques significantly improved segmentation, subtyping, and grading of tumors by combining data like MRI scans, pathology images, and omics information, which enables precise characterization of cancer and personalized treatment planning.

Under prognosis, survival analysis and treatment response prediction are included. With these new techniques, the integrated models help predict outcomes, optimize therapeutic strategies, and monitor disease progression. Finally, biomarker discovery is enriched through these multimodal learning approaches as it helps identify critical features across imaging, genetic, and clinical data, aiding in early detection, prognosis, and therapeutic decision-making (Zhou et al., 2024).

2. Multimodal Learning Framework for Alzheimer's Disease Diagnosis

The next state-of-the-art development is the use of multimodal learning to develop a framework for Alzheimer's disease diagnosis. This framework uses data from multiple modalities including:

1. Neuropsychological test data: data that captured cognitive performance across domains like memory, attention, and language, presented in forms such as speech, text, and video.
2. Neuroimaging data: using MRI and other imaging techniques to identify structural changes in the brain.
3. Other modalities: incorporating demographic information, clinical history, and cerebrospinal fluid (CSF) biomarkers.

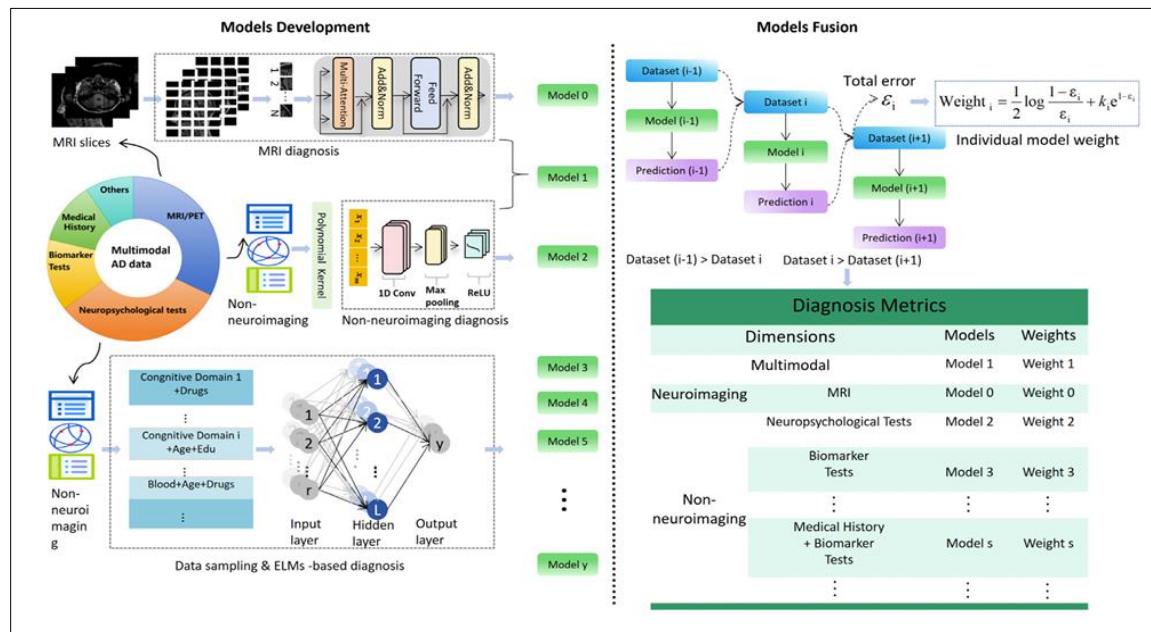


Figure 7. The overall structure of the framework (Zhang et al., 2024).

This article highlights a significant development covered in the study: the possibility of using a polynomial dimension expansion function to improve and retrieve multimodal clinical data. This approach guarantees the preservation and efficient integration of neuropsychological evaluations' diverse and intricate characteristics with neuroimaging and other forms of data. A significant advancement in this paradigm is also the use of Extreme Learning Machines (ELMs) that are cognitive in nature for the domain-specific interpretation of neuropsychological test data. The model extracts the most relevant information for diagnosis by reducing noise and redundancy by breaking tests into smaller groups. Finally, by applying dynamic weights to different models and modalities, our

method improves interpretability and performance. Figure 7 shows the entire framework in its entirety (Zhang et al., 2024).

The framework achieves a diagnosis accuracy of over 98% accuracy and F1 score on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, showcasing the ability to differentiate between mild cognitive impairment (MCI) and Alzheimer's disease. This research highlights the clinical potential of this multimodal approach. Furthermore, the approach's adaptability to diverse clinical datasets underscores its potential for broader applications in personalized healthcare(Zhang et al., 2024).

3. Multimodal Mental Health Analysis on Social Media

So far, we have discussed the recent state-of-the-art developments for physical disease diagnosis like Oncology and brain disorder, Alzheimer's. Now, we will look at state-of-the-art development using multimodal learning for mental health. The research paper, “Multimodal mental health analysis in social media” introduces a multimodal framework to analyze mental health, specifically depressive behaviors by utilizing data from social media platforms like Twitter.

Model#	Data Source	Ref.	Year	Features					Model	Spec.	Sens.	F-1	Acc.	
				N-grams	LIWC	Sentiment	Topics	Metadata						
I	Content	[110]	2016	X					NB	0.69	0.70	0.69	0.70	
II		[111]	2016	X		X			User Acti.	N/A (LR)	0.73	0.74	0.73	0.74
III		[112]	2015	X	X	X			User Acti.	Log-linear	0.83	0.80	0.81	0.82
IV		[113]	2015	X	X	X	X		LR	0.84	0.83	0.84	0.84	
V		[114]	2015	X	X	X	X	User Acti.	SVM	0.86	0.84	0.85	0.85	
VI		N/A	N/A	X					SVM(Pre. embed.)	0.72	0.72	0.72	0.72	
VII		N/A	N/A	X					SVM(Train w2vec)	0.70	0.70	0.70	0.70	
VIII		Cont., Net.	[10]	2013	X	X	X		SVM, PCA	0.84	0.80	0.83	0.85	
IX	Image	N/A	N/A						LR	0.68	0.67	0.67	0.68	
X		N/A	N/A						SVM	0.69	0.67	0.67	0.69	
XI		N/A	N/A						RF	0.72	0.70	0.69	0.71	
Ours	Cont.,Image,Net.	N/A	X	X	X	X	X	X	N/A	0.87	0.92	0.90	0.90	

Table 1: Model’s performance for depressed user identification in Twitter using different data modalities (Yazdavar et al., 2020).

The framework uses visual data (images), textual data (tweets), and user metadata to uncover patterns that signal depression. The study builds upon the finding that social media users frequently use a variety of modalities, including shared posts, linguistic style, and profile images, to communicate their feelings and mental health issues. Early fusion was utilized in the framework. This strategy outperforms conventional single-modality approaches by utilizing sophisticated machine learning and feature fusion techniques, enhancing the F1-score for diagnosing depression by 5% in comparison to state-of-the-art models. In Table 1, we can clearly see that developed multimodal framework outperforms

the other baseline models in identifying depressed users in terms of average specificity, sensitivity, F-measure and accuracy (Yazdavar et al., 2020).

The researchers created an 8,770 Twitter users' dataset using psycholinguistic analysis and explicit self-reported symptoms that was annotated for depression and non-depressive behaviors. Features including engagement metrics, language patterns, and the aesthetics of profile images were examined using the framework. Users who are depressed, for instance, have been shown to like images that are darker and less vivid, which reflect their bad emotional states. Higher use of first-person pronouns and terms conveying negative sentiment were found in the text, which is in line with clinical indicators of depression. By the inclusion of demographic assumptions such as age and gender, the analysis was in-depth and revealed that younger and female users were more likely to exhibit sad expressions (Yazdavar et al., 2020).

One of the most significant outcomes was the interpretability of the multimodal framework. By visualizing feature contributions, the model illustrated the importance of various factors, such as image color or linguistic authenticity, in predicting depressing moods. The study emphasizes the potential of multimodal techniques to support conventional mental health evaluations, enabling early detection and personalized interventions on a large scale.

Challenges

There are several major challenges in this area. One of the primary challenges is the integration and fusion of heterogeneous data from various sources (Zhou et al., 2024). As we have seen in the above sections, in healthcare, there are several types of data, including clinical text, medical images (e.g., MRI, X-rays), electronic health records (EHRs), genomic data, and even real-time sensor data from wearable devices. Since each of these data have their own structure, format, and scale, it is difficult to combine them in a way that preserves their individual information and maximize potential insights. Moreover, combining modalities is not always straightforward and determining the best way to integrate the data in a meaningful way is an ongoing research challenge. As we discussed in the types of data fusion, the effectiveness of different fusion approaches depends on the quality of data and the algorithms used for integration.

In addition, information redundancy is a challenge for these models as it is difficult to distinguish between task-relevant and irrelevant information. This results in inconsistency and extraction of meaningless insights. Due to this, it is challenging to achieve the goal of gaining the trust of clinicians and patients to trust and accept the multimodal AI model's diagnosis and treatment recommendations (Zhou et al., 2024).

Apart from this, high-quality data is essential for training to result in accurate predictions, but it is difficult to obtain due to limited access to large, annotated datasets with complete patient records and correct labeling. Even if we have access to large healthcare related datasets, we face the issue of lack of significant computational resources to process, integrate, and analyze. Training and deploying multimodal models require substantial computational power, efficient algorithms, and resource-constrained environments.

At the beginning of this review, we learned that successful AI models were developed but none were deployed due to ethical concerns, biased outputs, and fairness. These reasons still one of the most crucial challenges to tackle. The use of multimodal data in healthcare raises important privacy and ethical issues. These patient datasets are highly sensitive and thus integration of diverse data sources from third part platforms intensifies the issue of data privacy and security. Furthermore, as mentioned previously, bias is a significant concern in healthcare, as it can lead to disparities in treatment and outcomes for different patient populations. Multimodal learning models may inadvertently learn biases present in the data, such as gender, age, or racial disparities in healthcare access and treatment.

Discussion

The outlook for multimodal learning in healthcare is highly promising from the information gathered so far. It is expected to have a big impact on individualized care, treatment planning, and medical diagnostics. We have much ongoing research on this topic and several healthcare fields are being focused. In the coming years, I predict more thorough and precise patient health models will be possible as healthcare continues its digital revolution through the integration of several data modalities, such as genomic data, electronic health records (EHRs), medical imaging, and real-time wearable device monitoring. Because it integrates the advantages of several data kinds, multimodal learning has the potential to completely transform diagnostic accuracy by providing insights that single-modality techniques are unable to provide.

Within the on-going research, there are not many focused on minimizing the bias in healthcare dataset. More research needs to be conducted in this area so that these new models will be implemented in the real world and make clinicians work efficiently. Apart from this data quality and availability remain significant barriers. Addressing these issues requires the standardization of healthcare data, the development of efficient data fusion techniques, and the implementation of robust privacy measures. Also, research on optimizing multimodal is necessary to ensure scalability and developing models that require less resource constrained.

In terms of new applications, the next 10-20 years could see multimodal learning may find new uses in cutting-edge domains like precision medicine, where AI models integrate lifestyle, environmental, and genetic data to provide individualized treatment regimens. By examining speech patterns, social media activity, and physiological data from wearable technology, multimodal learning may also be used in the mental health field to help identify early indicators of anxiety, depression, and other mental health issues. Furthermore, more comprehensive patient care may result from combining clinical data with social determinants of health (SDoH), such as socioeconomic and environmental factors. These developments may result in healthcare delivery that is more egalitarian, lowering inequalities and enhancing health outcomes worldwide.

References

- Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). *Multimodal Machine Learning: A Survey and Taxonomy*. <https://arxiv.org/pdf/1705.09406>
- Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7), 389–403. <https://doi.org/10.1038/s41576-019-0122-6>
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A Guide to Deep Learning in Healthcare. *Nature Medicine*, 25(1), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2018). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 127(4), 398–414. <https://doi.org/10.1007/s11263-018-1116-0>
- Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, 92(4), 807–812. <https://doi.org/10.1016/j.gie.2020.06.040>
- Krones, F. H., Walker, B., Mahdi, A., Kiskin, I., Lyons, T., & Parsons, G. (2022). Dual Bayesian ResNet: A Deep Learning Approach to Heart Murmur Detection. *Computing in Cardiology*. <https://doi.org/10.22489/cinc.2022.355>
- Krones, F., Umar Marikkar, Parsons, G., Szmul, A., & Mahdi, A. (2024). Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114(1566-2535), 102690–102690. <https://doi.org/10.1016/j.inffus.2024.102690>
- Potrimba, P. (2023, May 10). *Multimodal Models and Computer Vision: A Deep Dive*. Roboflow Blog. <https://blog.roboflow.com/multimodal-models/>
- Practice Accelerator. (2023, August 29). *A Historical Look at AI in Health Care | WoundSource*. www.woundsource.com. <https://www.woundsource.com/blog/historical-look-ai-in-health-care>
- Pulapakura, R. (2024, February 20). *Multimodal Models and Fusion - A Complete Guide | Medium*. Medium. <https://medium.com/@raj.pulapakura/multimodal-models-and-fusion-a-complete-guide-225ca91f6861>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv:2103.00020 [Cs]*. <https://arxiv.org/abs/2103.00020>

- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/nejmra1814259>
- Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34(6), 96–108. <https://doi.org/10.1109/msp.2017.2738401>
- Taleb, A., Lippert, C., Klein, T., & Nabi, M. (2021). *Multimodal Self-supervised Learning for Medical Image Analysis*. 661–673. https://doi.org/10.1007/978-3-030-78191-0_51
- Wang, T., Chen, X., Zhang, J., Feng, Q., & Huang, M. (2023). Deep multimodality-disentangled association analysis network for imaging genetics in neurodegenerative diseases. *Medical Image Analysis*, 88, 102842. <https://doi.org/10.1016/j.media.2023.102842>
- Wang, Z., Wu, Z., Agarwal, D. C., & Sun, J. (2022). MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *ArXiv (Cornell University)*, 2022. <https://doi.org/10.18653/v1/2022.emnlp-main.256>
- Xsolis. (2021, February 2). *The Evolution of AI in Healthcare*. Xsolis. <https://www.xsolis.com/blog/the-evolution-of-ai-in-healthcare/>
- Yan, K., Li, T., Marques, J. A. L., Gao, J., & Fong, S. J. (2023). A review on multimodal machine learning in medical diagnostics. *Mathematical Biosciences and Engineering: MBE*, 20(5), 8708–8726. <https://doi.org/10.3934/mbe.2023382>
- Zhang, M., Cui, Q., Lü, Y., Yu, W., & Li, W. (2024). A multimodal learning machine framework for Alzheimer's disease diagnosis based on neuropsychological and neuroimaging data. *Computers & Industrial Engineering*, 206(1053-8119), 110625–110625. <https://doi.org/10.1016/j.cie.2024.110625>
- Zhou, H., Zhou, F., Zhao, C., Xu, Y., Luo, L., & Chen, H. (2024). *Multimodal Data Integration for Precision Oncology: Challenges and Future Directions*. ArXiv.org. <https://arxiv.org/abs/2406.19611>