Multimodal Machine Learning in Healthcare: A Tutorial and Review

Muntaqim Ahmed Raju *, Priyanka Siddappa, Md Shifat Haider Al Amin, Ruizhe Ma

Posted Date: 16 December 2025

doi: 10.20944/preprints202512.1445.v1

Keywords: multimodal machine learning; healthcare AI; data fusion; heterogeneous data; clinical decision support; medical imaging; electronic health records; time series analysis; fusion strategies; deep learning in medicine

 doi:10.20944/preprints202512.1445.v1

Article

Multimodal Machine Learning in Healthcare: A Tutorial and Review

Muntaqim Ahmed Raju *, Priyanka Siddappa , Md Shifat Haider Al Amin and Ruizhe Ma

Miner School of Computer & Information Sciences, University of Massachusetts Lowell, Lowell, MA 01854, USA
* Correspondence: MuntaqimAhmed_Raju@student.uml.edu

Abstract

Integration of deep learning in healthcare has revolutionized the analysis of complex, high-dimensional, and heterogeneous data. However, traditional single-modal approaches often fail to grasp the multi-faceted nature of human health, in which genetic, environmental, lifestyle, and physiological factors interact in complex ways. The rapid development of multimodal machine learning (MML) has been a transformational paradigm that allows seamless integration of these heterogeneous data sources toward a better understanding of health and disease. This review goes in-depth with the methodologies of MML, with special emphasis on the main strategies of fusion and advanced techniques. We also discuss the wide applications of MML in different health domains, such as brain disorders, cancer prediction, chest-related conditions, skin diseases, and other medical challenges. We illustrate, through detailed case studies, how MML provides better diagnostic accuracy, and personalized treatment strategies. While it has seen huge progress, MML is confronted with a few major challenges around data heterogeneity, alignment complexities, and the subtleties of effective fusion strategies. The review concludes with a discussion on the future directions calling for robust data integration techniques, efficient and scalable architectures, and fairness and bias mitigation. MML is still an evolving field, and it has the potential to revolutionize healthcare delivery and drive innovations in the direction of more personalized, equitable, and effective patient care globally.

Keywords: multimodal machine learning; healthcare AI; data fusion; heterogeneous data; clinical decision support; medical imaging; electronic health records; time series analysis; fusion strategies; deep learning in medicine

1. Introduction

Deep learning has revolutionized many fields, most notably healthcare, by enabling analysis of complex and high-dimensional data [1,2]. Deep learning models, especially multi-layered neural networks, have been performing with great excellence in tasks of image recognition, natural language processing, and predictive analytics [3]. They have been indispensable in the advance of medical imaging analysis, genomics, and mining of electronic health records (EHRs) among others in healthcare [1,3]. However, most of the traditional deep learning approaches have been designed with single-modal data in mind, which may never truly represent the essence of human health [2]. Health and disease are modulated by a myriad of influences, including genetic makeup, environmental exposures, lifestyle choices, and physiological states. Single-modal models might overlook these important interactions between such factors, resulting in incomplete or less accurate outcomes. To cope with this limitation, researchers have turned to multimodal deep learning, which allows the integration of heterogeneous sources of data to more fully understand health and disease [1,3].

MML is a next-generation computational approach wherein multi-modal data are integrated and analyzed to result in more profound insights and more robust decision-making within healthcare [4]. Healthcare data is inherently diverse, encompassing imaging data, structured clinical records, genomics, unstructured text, and sensor data [5–7]. This multi-modality provides different insights into

 doi:10.20944/preprints202512.1445.v1

the health of a patient. MML brings this diversity into unity and gives insights not possible from the single source. The holistic approach, therefore, can potentially result in better diagnostic accuracy, treatment strategies, and improvement in personalized medicine [1,3,5,7]. For example, the integration of imaging data with genomic profiles has driven precision oncology forward, enabling clinicians to define molecular subtypes of cancers and, therefore, tailor treatments to those subtypes [8]. Warner et al. [9] underline the role of MML in medical image analysis and clinical decision support systems for disease diagnosis and treatment planning. Similarly, the integration of real-time sensor data with longitudinal clinical records has improved early warning systems for patient deterioration in intensive care units, enabling timely interventions and improving outcomes. Bertsimas and Ma [10] introduced M3H, a framework for multimodal multitask machine learning, showing improved performance over many medical tasks by fusing data from multiple sources. Integration of multiple sources of data in clinical practice is discussed by Krones et al. [11] in a review focusing on how MML may provide improved decision-making. Beyond the technical advantages, MML extends to address practical challenges faced in healthcare—it allows the development of systems using structured and unstructured data to bridge the gap between complex datasets and clinical applicability. This adaptability has put MML at the very core of a cornerstone technology that will define the future of healthcare and drive innovations in population health, disease prediction, and patient monitoring. There is growing interest in MML, since it uncovers relationships across modalities, supports a more holistic approach to patient care, and ensures that no critical information goes unnoticed. The way that MML keeps being developed keeps revolutionizing healthcare and keeps pushing the boundaries of what can be achieved using data-driven approaches. It enhances accuracy and scalability, perfectly in line with an increased focus on personalized, preventive, and precision medicine. This makes MML an indispensable tool in the solution of complex challenges in modern healthcare and improvement of patient outcomes globally.

The paper is structured into various interrelated sections that provide a holistic exploration of MML in healthcare. It first provides an introduction that focuses on the pivotal role of MML in revolutionizing healthcare through the integration of diverse data sources. It is then followed by the review of the important data modalities which includes various medical imaging techniques, EHRs, unstructured clinical text, time-series data originating from vital signs and wearable sensors, and tabular data comprising demographic, clinical, and pharmaceutical information. Afterwards, it discusses the intrinsic challenges of MML, including data heterogeneity, alignment complexities, and the subtleties of effective fusion strategies. It covers a wide-ranging survey on modality-level fusion approaches, including early, intermediate, late, and hybrid fusion, and feature-level fusion techniques, including concatenation, operation-based, and learning-based approaches. These building blocks then give a good starting point for more advanced MML methodologies, which include attention mechanisms, cross-modal embeddings, generative models, and graph neural networks. The core of the paper describes the applications in practice of MML, using real-world use cases in different health domains, such as brain disorders, cancer prediction, chest-related conditions, skin diseases, and other medical challenges. Notable case studies for each application are then presented, demonstrating the real and tangible benefits of MML. The final sections of the review then look at future directions for MML in healthcare, including the development of better data acquisition and integration pipelines, interpretability and explainability, and fairness and bias. These forward-looking perspectives underline MML's evolving role in shaping a more personalized, equitable, and efficient healthcare landscape. The paper provides an overview of the transformation potential of MML within modern healthcare by systematically navigating methodological aspects, applications, and future prospects.

2. Fundamentals of Multimodal Machine Learning

This section provides a detailed review of the basics of MML, delving into the various data modalities commonly found in healthcare, the intrinsic challenges in integrating such complex data, and sophisticated fusion strategies applied at both modality and feature levels.

 doi:10.20944/preprints202512.1445.v1

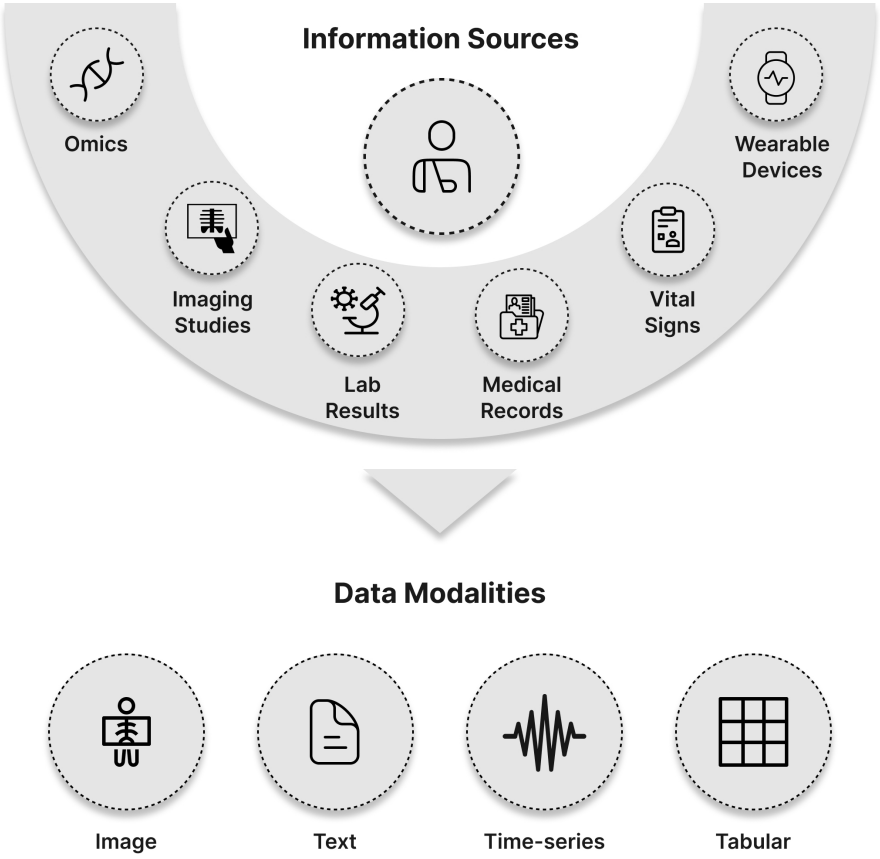

Figure 1. Data Modalities.

Medical imaging is an integral part of modern health care, providing visual representations of internal body structures and allowing for precise diagnoses and treatment planning. The following describes the main modalities of medical imaging, their characteristics, and applications in health care.

a) X-ray Imaging:

X-ray imaging is one of the most commonly used and available medical imaging modalities [12]. This modality relies on the principle of differential attenuation of X-rays as they pass through different tissues in the body [13]. The resultant images are 2D grayscale radiographs where air appears darkest, and denser materials, such as bone and metal, appear white due to the higher absorption of X-rays. [14]. Applications of X-rays are many, from the diagnosis of bone fractures to the detection of pneumonia and cardiomegaly assessment. More recent developments have integrated X-rays into machine learning pipelines for more complex tasks, such as lung cancer screening and cardiological assessments [15,16]. The preprocessing commonly adopted in machine learning includes resizing images to a standard set of dimensions, such as 224×224 pixels, and, where applicable, converting grayscale images to three-channel RGB formats [15,17].

b) Computed Tomography (CT):

CT produces detailed 3D volumetric images by reconstructing multiple 2D slices obtained through radiographic projections. The imaging intensity, measured in Hounsfield Units (HU), reflects tissue density, allowing for the identification of structural abnormalities and pathological changes with a high degree of precision [18]. CT provides a flexible imaging modality that has come to be used for a fairly broad spectrum of applications in detecting structural abnormalities, tumor identification, and neuroimaging for the diagnostics of neurological diseases [19–21], among many others. CT studies

 doi:10.20944/preprints202512.1445.v1

are also very instrumental in respiratory [22] and cardiological assessment [23–25] and treatment planning [26,27] within oncology [28].

c) Magnetic Resonance Imaging (MRI):

MRI is a non-invasive imaging modality based on the generation of detailed images of soft tissues by using magnetic fields and radiofrequency pulses [29]. It offers high signal-to-noise ratios and excellent contrast for internal structures without ionizing radiation [30,31]. MRI has been found to be especially effective for neurological, musculoskeletal, and cardiovascular imaging. It has been widely used in studying various brain disorders like Alzheimer's [32,33], Parkinson's disease [34], and multiple sclerosis [35]. Most of the machine learning models applied to the MRI data include preprocessing steps like anti-aliasing filters and normalization to remove the image's inherent artifacts from the imaging process [36].

d) Ultrasound Imaging:

Ultrasound imaging uses high-frequency sound waves to create images of body structures [37], enabling 2D, 3D, or even 4D images in real-time [38,39]. It is much valued for being non-invasive, inexpensive, and free of ionizing radiation [39]. Applications include obstetric, cardiology, and abdominal imaging, with the possibility of including Doppler techniques to visualize blood flows [38,40]. Pre-processing for machine learning involves ROI segmentation and noise reduction since speckling artifacts are inherent in ultrasound images [41].

e) Dermoscopic Imaging:

Dermoscopy is a non-invasive imaging technique that takes high-resolution pictures of the skin surface, which is helpful in diagnosing different dermatological diseases, especially melanoma [42]. Using polarized or non-polarized light, dermoscopy shows subsurface skin structures that are invisible to the naked eye, thereby improving the accuracy of clinical assessment [43]. The Dermoscopic images can be analyzed by the computer-aided diagnostic tools, powered by machine-learning algorithms [44], which help the dermatologists in making more accurate and effective diagnoses [11]. Before feeding images into a machine learning model, regions of interest are normally segmented in order to remove the redundant information [45]. Images are also cleaned to remove unwanted artifacts such as ink marks, gel bubbles, and hairs [46]. Dermoscopy in dermatology is mainly used for the visualization of subsurface structures of the epidermis and dermis to improve diagnosis and monitoring of skin lesions and tumors [47,48]. This method has been very useful in differentiating benign from malignant lesions, therefore increasing diagnostic precision, reducing the need for invasive procedures such as biopsies [49].

2.0.1. Text Data

Textual data represents a core of medical information and contains most patient-specific documents integral to clinical practice [50,51]. This may include procedure notes, which contain detailed accounts of medical interventions; comprehensive clinical records that document patient histories, progress reports, and consultation notes; and prescription notes outlining medication regimens with specific dosages and administration guidelines [52]. The medical discharge summary is important during care transitions and summarizes the hospitalization with diagnoses, treatments, and follow-up plans [53]. Referral letters are used to communicate between health providers about the patient's important information and reasons for the referral. Radiology reports detail the findings from imaging studies such as X-rays, CTs, and MRIs [16,54,55]. Besides the patient-specific records, medical text data also includes general medical knowledge from sources such as journals, literature, medical websites, and pharmaceutical labels [56,57]. The nature of medical text is unstructured, which in itself implies challenges in data analysis and integration. Natural Language Processing (NLP) techniques have been applied to the extraction of relevant information from such records and to converting the unstructured text into structured data usable by applications of machine learning [58,59]. The traditional approach

 doi:10.20944/preprints202512.1445.v1

to NLP normally brings along domain experts who are going to annotate features of interest—time and resource consuming. To reduce this bottleneck, active learning techniques have been introduced to select text segments for annotation, hence optimizing the labeling process [60]. Transfer learning and data augmentation techniques, including synthetic clinical note generation, are also being explored to improve efficiency and effectiveness in NLP for medical contexts [61].

2.0.2. Time Series Data

Time series data, in the form of sequential observations collected with time, are very important in healthcare for monitoring patient health and guiding clinical decisions. They contain vital signs, such as heart rate and blood pressure; laboratory results; medication administration records; and sensor outputs from wearable devices [62]. The temporal nature of such data supports trend analysis, anomaly detection, and forecasting of future health events [63]. Time series analysis in health care faces a range of unique challenges, including but not limited to, irregular intervals, missing values, and high dimensional patient information data [64,65]. In time series modeling and in-sample forecasting, classical statistical models like autoregressive integrated moving average (ARIMA) have been an acceptable choice and were widely applied to various purposes in data [66–69]. However, the complexity of the data in health care usually requires more advanced approaches [70]. Machine learning techniques, especially deep learning models such as Long Short-Term Memory (LSTM) networks, have shown great promise in capturing the complex temporal patterns within healthcare time series data [71]. Models like these, which are inherently able to deal with variable lengths of sequences and learn long-term dependencies, are directly applicable to the prediction of disease progression or patient deterioration [66,72–74].

2.0.3. Tabular Data

Tabular data forms the basis of healthcare information systems, clinical research, and patient care. In contrast to time-series data, which captures the continuum of observations with respect to time, tabular data gives static snapshots of a variety of attributes for a patient, rendering it highly versatile for a variety of applications [75,76]. This modality is heavily used in representing demographic information, such as age, sex, ethnicity, and socioeconomic factors, which are critical for tailoring individualized care and understanding population health trends [16,23]. Clinical evaluations, including standardized scoring systems such as the Acute Physiology and Chronic Health Evaluation II (APACHE II) [77], pain scales [78], and quality-of-life indices [79], are mostly kept in tabular form for consistent assessment and to facilitate comparisons of outcomes across groups. Laboratory findings, including blood chemistry, hematology, and microbiology, are also systematized into tables to enhance diagnostic precision and prognostic decision-making [23]. Pharmaceutical data, including drug dosages, routes of administration, and pharmacological properties, are also organized in tables to help medication management, decrease the possibility of drug interactions, and conduct pharmacovigilance studies more easily [25]. The structured nature of tabular data makes it highly amenable to computational analysis. These predictive models, which could be developed to perform disease diagnosis, risk stratification, and outcome forecasting based on traditional supervised machine learning approaches such as logistic regression [44,80], decision trees, and ensemble methods like random forests [44,81], depend on feature engineering, one of the most crucial aspects in tabular data analysis, where methods like correlation-based feature selection and dimensionality reduction techniques, including principal component analysis (PCA), help improve model interpretability and its performance [82]. Recent developments have also brought deep learning solutions specifically designed for tabular data, such as TabNet [83], which uses attention mechanisms to model complex interactions between features while preserving interpretability. Besides, tabular data is often a base for multi-modal integration, enriching machine learning models when integrated with other types of data. For example, the integration of tabular demographic variables with imaging data has improved diagnostic accuracy in prostate cancer detection [84], and a combination of clinical and genomic data with imaging modalities has achieved a robust prediction of treatment response in lung cancer [85].

 doi:10.20944/preprints202512.1445.v1

2.1. Challenges in Multimodal Machine Learning

MML in healthcare encompasses the integration of data from different sources to facilitate better patient care. However, several challenges arise due to this integration, which have to be resolved to create truly effective and dependable models.

2.1.1. Heterogeneity of Modalities

Healthcare data takes many forms, from structured data in lab results to unstructured text in clinical notes, images like X-rays and MRIs, genomic sequences, and continuous data from wearable sensors. Each modality possesses its unique characteristics, which need to be handled with preprocessing and analysis techniques developed around the special characteristics of that modality. For instance, natural language processing is required to extract relevant information from clinical notes [1,86], whereas computer vision methods are needed for the interpretation of medical images [87]. Similarly, improvements in next-generation sequencing have made it possible to process complex genomic data for clinical applications [88], while wearable sensors produce continuous streams of data with their unique set of preprocessing challenges [89]. The heterogeneity of scales, units, and statistical properties in the different modalities makes them hard to be combined into a coherent model. Similarly, such heterogeneity could lead to some issues regarding preprocessing, feature extraction, and normalization of data that are foundational steps toward the development of robust models in machine learning [90]. Integration of multimodal healthcare data requires good data fusion techniques for handling the inherent diversity and complexity in the datasets [91].

2.1.2. Alignment

Another important challenge is alignment, more specifically, the temporal and spatial synchronization of data from different modalities. Data collected from different sources are usually not temporally or spatially aligned [92]. Wearable devices can provide a consecutive stream of physiological data, while imaging studies and lab tests are usually performed intermittently [93–95]. The clinical notes can be recorded hours or days after the patient interactions, hence causing temporal discrepancies [96]. Spatial misalignment can also arise when combining imaging data that capture different anatomical views or resolutions, or when the anatomical differences between patients introduces variability [97]. This type of misalignment will result in the inability to correlate events across modalities and can make the model design even more complex, requiring advanced architectures able to handle missing or asynchronous data points [98]. It becomes very important to deal with such issues of alignment to ensure the multimodal data integrates effectively and the development of strong machine-learning models that are reliable.

2.1.3. Fusion Strategies

Effective fusion strategies are indispensable to combine information from multiple modalities but are challenging. Early fusion may result in high-dimensional feature spaces, thus increasing computational complexity and the risk of overfitting, especially for small datasets [99]. Moreover, the alignment and normalization of features across modalities with different scales, units, and statistical properties is a non-trivial task. Joint fusion, while it does solve some of the problems with early fusion, introduces its own in determining what is the right stage to combine and ensuring balanced contributions from all modalities. Poor design will make some modalities dominate others, resulting in biased learning outcomes [100]. On the other hand, late fusion is less complex in dealing with individual modalities but faces challenges while capturing cross-modal interactions, which may be valuable in some tasks. It's also sensitive to inconsistencies in predictions from individual modalities, which may propagate errors through decision-based aggregation [100]. Hybrid fusion combines early, joint, and late fusion, which further complicates the design and training of models. It requires delicate architectural tuning of the balance between the strengths of different fusion strategies and mitigation of their weaknesses, which can significantly increase computational costs [101]. Moreover, the tendency of multi-modal models toward overfitting is an inherent problem, which is aggravated by the limited

 doi:10.20944/preprints202512.1445.v1

availability of medical datasets due to privacy issues and the high costs associated with data collection. Complex fusion models with a great number of parameters may fit the training data too closely and poorly generalize to new data [102]. Usually, regularization techniques, data augmentation, and strong cross-validation methods have to be employed in order to counter these issues.Finally, training multi-modal models, especially hybrid fusion-based models, is computationally expensive. There is a need for both high-performance computing infrastructure and efficient optimization algorithms in order to optimize the parameters for such resource-demanding processes [103].

2.2. Techniques for Multimodal Machine Learning

Fusion strategies are the core of MML in order to combine information across different data modalities for better model performance and capture complementary information that may not be caught by a single modality. Generally, these fusion strategies can be divided into two classes: modality level fusion and feature level fusion. Below, we will discuss each class and their subtypes with thorough explanations.

2.2.1. Modality Level Fusion

Modality level fusion refers to the level at which data from different modalities are fused within a machine learning pipeline. The main approaches are: Early Fusion, Intermediate (Joint) Fusion, Late Fusion, and Hybrid (Mixed) Fusion.

a) Early Fusion:

Early fusion [104–106], also called feature-level fusion [107–109], is one of the core approaches in MML and involves directly fusing the raw data or extracted features from each of the modalities at an early learning stage. It combines extracted features of each modality into a single representation by which a machine learning model learns joint patterns and correlations across the modalities right from the very beginning.

In early fusion, feature extraction from each modality is generally performed, followed by feature concatenation, and then model training (see Figure 2(a)). For each modality m in the set of modalities $\mathcal{M}$, relevant features are extracted, denoted as $\mathbf{x}^{(m)} \in \mathbb{R}^{d_m}$, where d_m is the dimensionality of the modality's feature space. The extracted feature vectors from all modalities are concatenated into a single feature vector:

$$\mathbf{x} = [\mathbf{x}^{(1)}; \mathbf{x}^{(2)}; \ldots; \mathbf{x}^{(M)}] \in \mathbb{R}^{d},$$

where

$$d = \sum_{m=1}^{M} d_m,$$

and $[\cdot\,;\cdot]$ denotes the concatenation operation.

A machine learning model f is trained on the concatenated feature vector $\mathbf{x}$ to predict the target variable y:

$$y = f(\mathbf{x}) = f\Big([\mathbf{x}^{(1)}; \mathbf{x}^{(2)}; \ldots; \mathbf{x}^{(M)}]\Big).$$

In a supervised learning scenario, the objective is to learn a function f that minimizes a loss function L over the training data

$$\{(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(M)}, y_i)\}_{i=1}^{N},$$

where N is the number of training samples. The optimization problem is defined as:

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} L\Big(y_i, f\Big([\mathbf{x}_i^{(1)}; \mathbf{x}_i^{(2)}; \ldots; \mathbf{x}_i^{(M)}]\Big)\Big).$$

 doi:10.20944/preprints202512.1445.v1

For instance: In regression tasks, L could be the mean squared error (MSE):

$$L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2.$$

In classification tasks, L could be the cross-entropy loss.

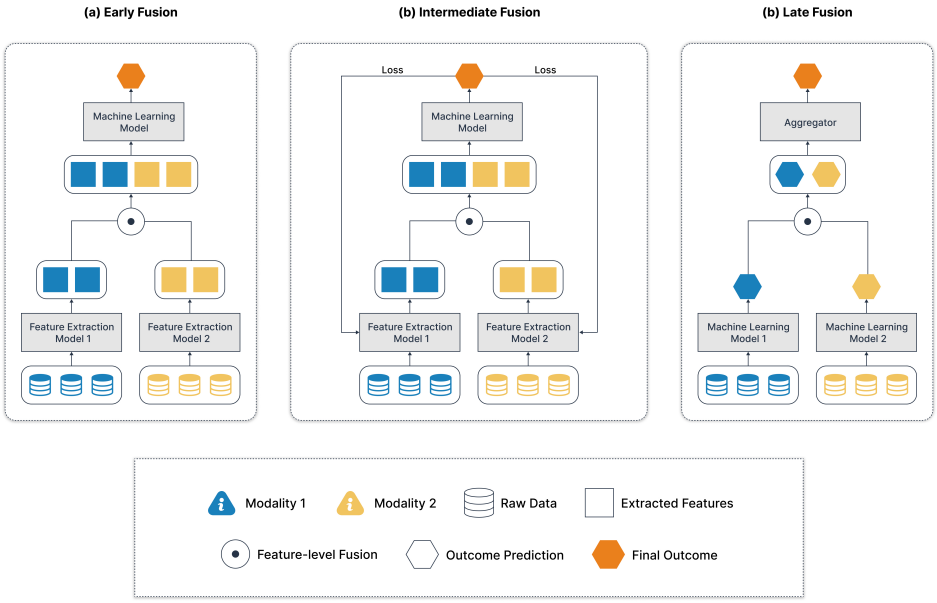

Figure 2. Comparison of fusion strategies: (a) Early fusion integrates raw features or extracted features (feature extraction is optional) before feeding them into the final model, (b) Intermediate fusion combines extracted features with backpropagation through a unified model, and (c) Late fusion aggregates outputs from independently processed modalities.

The key benefit of early fusion lies in the fact that it allows the model to learn relationships and dependencies across different modalities from the very beginning. Processing the combined features simultaneously, the model can learn complex patterns that might not be evident when the modalities are considered separately [100]. This unified representation also brings about a simplification in the architecture and potentially lessens the need for complicated synchronization mechanisms. Moreover, learning can be improved by using all data available in a combined manner at least for cases where the modalities convey complementary information. However, it also suffers from several drawbacks. For instance, concatenating feature vectors of many modalities leads to very high-dimensional feature space that may cause increased computational complexity and overfitting in case of limited training data available [110]. Another disadvantage lies in requiring synchronized data since the early fusion assumes that data in different modalities are aligned. This assumption could be difficult to fulfill in many real-world applications. Furthermore, if one modality has much more features or stronger predictive power, it may dominate the learning and somehow shadow the valuable information in other modalities [111]. The complexity of the model and computational intensity increase with the number of modalities and feature dimensions, which may not scale well.

b) Intermediate Fusion or Joint Fusion:

Intermediate Fusion [106,112] also known as Joint Fusion [104,105] is a technique in MML where multiple modalities are fused at one or more intermediate layers in a model, rather than at the input or output stages alone (see Figure 2(b)). This approach is in contrast to Early Fusion, where different modalities are combined at the input level, with the loss of modality-specific nuances before higher-level feature extraction. In Intermediate Fusion, the loss function is back-propagated through

 doi:10.20944/preprints202512.1445.v1

each modality's feature extractor in order to fine-tune those modality-specific representations during training [104].

In this framework, each modality $m \in \{1, 2, \ldots, M\}$ provides input data $x^{(m)}$, which is first processed independently by a modality-specific model or layer. This model—parameterized by $\theta^{(m)}$—transforms the raw input into a higher-level feature representation:

$$h^{(m)} = f^{(m)}(x^{(m)}; \theta^{(m)}).$$

For example:

- Visual data may be processed by convolutional neural networks (CNNs).
- Text data could be processed by recurrent neural networks (RNNs) or transformers.

These $h^{(m)}$ vectors capture the salient features of each modality separately.

The fusion occurs at one or more intermediate layers where these modality-specific representations are combined to form a joint representation $\mathbf{h}^{(F)}$. The fusion function Fusion can be a simple operation like concatenation, addition, or a more complex mechanism like attention or gating:

$$\mathbf{h}^{(F)} = \text{Fusion}\left(\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \ldots, \mathbf{h}^{(M)}\right).$$

For instance, if concatenation is used:

$$\mathbf{h}^{(F)} = \left[\mathbf{h}^{(1)}; \mathbf{h}^{(2)}; \ldots; \mathbf{h}^{(M)}\right].$$

Alternatively, if an attention mechanism is employed, the fusion might involve learned weights that modulate the influence of each modality:

$$\mathbf{h}^{(F)} = \sum_{m=1}^{M} \alpha^{(m)} \mathbf{h}^{(m)},$$

where

$$\alpha^{(m)} = \text{softmax}\left(\mathbf{w}^{\top} \mathbf{h}^{(m)}\right).$$

Here, $\mathbf{w}$ is a parameter vector learned during training, and $\alpha^{(m)}$ represents the attention weight for modality m.

After fusion, the joint representation $\mathbf{h}^{(F)}$ is processed through additional layers $f^{(F)}$ with parameters $\theta^{(F)}$ to produce the final output y:

$$y = f^{(F)}\left(\mathbf{h}^{(F)}; \theta^{(F)}\right).$$

Intermediate fusion has quite a few advantages and disadvantages, with diverse applications across fields. One major advantage is balanced processing: it allows deep, modality-specific feature extraction before fusion, hence guaranteeing that the unique characteristics of each modality are well captured. The second major benefit is that it can model complex inter-modal relationships; the fusion of modalities at intermediate layers makes the model learn intricate interactions that may not be captured with early or late fusion approaches. Intermediate fusion, however, has a few more challenges: it increases model complexity with separate networks for each modality and extra fusion layers, hence increasing computational requirements and the risk of overfitting. Further, there are challenges in the architectural design, since much thought has to be given to deciding the best fusion methods and where in the network the modalities should be combined most effectively.

Applications of intermediate fusion span a variety of fields. In multimodal sentiment analysis, Zadeh et al. [113] introduced the Memory Fusion Network (MFN), which processes language, visual, and acoustic modalities through separate LSTM networks and fuse their output at intermediate layers by using a multi-view gated memory mechanism. In medical image analysis, Suk et al. [102] introduced a deep learning framework for the diagnosis of Alzheimer's disease that processes MRI and PET

 doi:10.20944/preprints202512.1445.v1

images via separate convolutional layers and fuses modality-specific features at intermediate layers to form a joint representation for classification. In multimodal machine translation, Calixto et al. [114] have introduced a neural machine translation model, which separately processes the textual and visual information and fuse them at some intermediate layers so that the translation quality could be improved. These examples draw the flexibility and effectiveness of intermediate fusion in exploiting multimodal data for solving difficult problems.

c) Late Fusion:

In MML, late fusion [94,104,106], also known as decision-level fusion [82,108,115], is an approach where each modality is processed independently using separate models or pipelines; the outputs from these models are combined at the end of the learning process to make a final prediction [99,116]. This approach lends simplicity and modularity, as each modality can be handled using the most appropriate methods without regard for compatibility at the feature level. It is especially useful when the modalities are heterogeneous or when it is infeasible to combine raw data or features directly. In late fusion, each modality m in the set of modalities $\mathcal{M}$ is processed through its own machine learning model $f^{(m)}$, parameterized by $\theta^{(m)}$. The input data $\mathbf{x}^{(m)}$ for each modality is used to produce an output $\mathbf{y}^{(m)}$:

$$\mathbf{y}^{(m)} = f^{(m)}\left(\mathbf{x}^{(m)}; \theta^{(m)}\right), \quad \text{for } m = 1, 2, \ldots, M.$$

These outputs $\mathbf{y}^{(m)}$ can represent class probabilities, regression estimates, or any other form of predictions relevant to the task. The individual outputs are then combined using a fusion function Combine to produce a final output $\mathbf{y}$:

$$\mathbf{y} = \text{Combine}\left(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(M)}\right).$$

The fusion function can be implemented using various methods, such as averaging, majority voting, weighted summation, or training a meta-classifier on the outputs of the individual models. For instance, in the case of weighted summation, weights $w^{(m)}$ are assigned to each modality's output based on their reliability or importance:

$$\mathbf{y} = \sum_{m=1}^{M} w^{(m)} \mathbf{y}^{(m)}, \quad \text{with } \sum_{m=1}^{M} w^{(m)} = 1.$$

After combining the outputs, the final prediction is made based on the aggregated result (see Figure 2(c)). For classification tasks, this may involve selecting the class with the highest combined probability or applying a threshold to determine the class label.

One of the most obvious advantages of late fusion is simplicity and flexibility. Since each modality can be processed independently, implementation and debugging are much easier. Models can be developed and updated independently without affecting the whole system; also, new modalities can be added or old ones removed without retraining the whole system. Moreover, late fusion is robust to modality-specific noise or failure, in the sense that if one modality provides poor-quality data, the others can still contribute to the final decision. On the downside, late fusion suffers from a number of disadvantages due to its limited ability to capture interactions between modalities during learning. Processing each modality independently may prevent the model from fully exploiting the rich correlations and complementary information that exist between modalities. Such limitations might translate to suboptimal performance with regard to those methods that are integrated earlier in the learning process by modalities.

A good example of late fusion can be found in the analysis of multimedia content. Snoek et al. [110] compared early and late fusion methods for semantic video analysis. In their late fusion approach, they trained separate classifiers on visual and textual features extracted from videos and combined the output of the classifiers using weighted summation. They showed that late fusion, while

 doi:10.20944/preprints202512.1445.v1

simpler to implement and more flexible, might not model synergies between modalities as well as early fusion does. Their work exposes the trade-offs entailed by the choice of fusion strategy.

d) Hybrid fusion or Mixed Fusion:

Hybrid Fusion [11,99,112], also known as Mixed Fusion, is an MML method that combines early and late fusion in order to harvest their benefits. This means that data from different modalities are fused at multiple levels in the model, allowing modality-specific processing and cross-modal interactions at various points of the learning process. By doing so, it decreases the modality imbalances sometimes introduced in early fusion while still modeling inter-modality dependencies that late fusion alone cannot capture [99,112,117].

Key Idea:

- Early Fusion can cause imbalance if one modality dominates at the raw feature level.
- Late Fusion might miss subtle inter-modality interactions.
- Hybrid Fusion tries to find a tradeoff by processing each modality to some optimal extent before and after their combination [118–121].

While hybrid fusion has the advantage of dealing with modality imbalances, designing such networks is challenging because deciding where in the processing pipeline the modalities should be combined needs to be done very carefully [122].

An example of hybrid fusion in practice is the Tensor Fusion Network (TFN) introduced by Zadeh et al. [117] for multimodal sentiment analysis. The TFN performs hybrid fusion by capturing both individual modality features and their interactions. Modality-specific features from text, audio, and video are first extracted and then fused using a tensor outer product to capture high-order interactions:

$$\mathbf{h}_{\text{fusion}} = \mathbf{h}_{\text{text}} \otimes \mathbf{h}_{\text{audio}} \otimes \mathbf{h}_{\text{video}}.$$

This fused representation is then used for sentiment prediction. The TFN also considers modality-specific predictions, combining them with the joint prediction at the decision level to enhance performance [117].

2.2.2. Feature Level Fusion

Feature level fusion describes how features from the different modalities are fused. The most adopted approaches are concatenation, operation-based fusion, and learning-based fusion. For early and intermediate fusion, different strategies to integrate the features can be used. Whereas we only give a brief overview in Figure 3, a more elaborated discussion can be found in the literature [99,123–125].

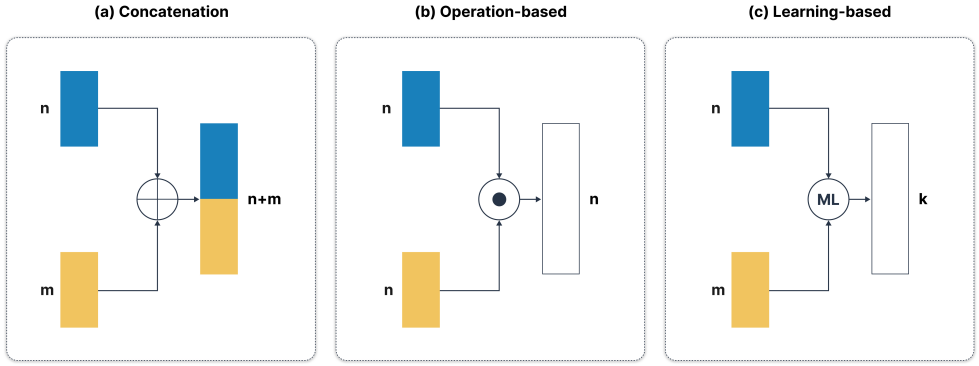

Figure 3. Feature-level fusion: (a) End-to-end concatenation of feature vectors, (b) Element-wise operations or attention mechanisms for same-shaped vectors, and (c) Learning-based methods for reconstructing features in a shared space.

 doi:10.20944/preprints202512.1445.v1

a) Concatenation:

Concatenation is one of the simplest methods for feature-level fusion in MML. It involves combining feature vectors extracted from different modalities into a single, unified feature vector by horizontally stacking them (see Figure 3(a)) [123]. For each modality m in the set of modalities $\mathcal{M}$, relevant features are extracted, resulting in feature vectors $\mathbf{x}^{(m)} \in \mathbb{R}^{d_m}$, where d_m is the dimensionality of that modality's feature space. These extracted feature vectors are concatenated to form a single feature vector:

$$\mathbf{x} = \left[\mathbf{x}^{(1)}; \mathbf{x}^{(2)}; \ldots; \mathbf{x}^{(M)}\right] \in \mathbb{R}^{d},$$

where

$$d = \sum_{m=1}^{M} d_m,$$

and $[\cdot\,;\cdot]$ denotes the concatenation operation.

Optionally, normalization techniques such as min-max scaling or z-score normalization [44] can be applied to each feature vector, before or after concatenation, to ensure compatibility between the features from the different modalities.

A machine learning model f is then trained on the combined feature vector $\mathbf{x}$ to predict the target variable y, expressed as:

$$y = f(\mathbf{x}; \theta),$$

where θ represents the model parameters.

This method is straightforward and easy to implement, hence attractive for initial experiments or in cases where computational resources are limited [126]. By concatenating all features, it ensures that no information from any modality is thrown away, potentially providing a very rich set of features that the model can learn from. However, The resulting feature vector can be of high dimension, especially for multiple modalities with large feature spaces. This will not only increase computational complexity but may also lead to the "curse of dimensionality," in which the volume of the feature space becomes so large that the available data becomes sparse. In addition, concatenation may include redundant or irrelevant features that do not contribute much to the model's predictive power but can hurt the performance. With high-dimensional input, models are more prone to overfitting, especially if the amount of training data is limited relative to the number of features [16,66,127]. In order to reduce the problems of high dimensionality, dimensionality reduction techniques like Principal Component Analysis (PCA) [128] or Linear Discriminant Analysis (LDA) [129] can be applied before or after concatenation to reduce the feature space. Feature selection [44] methods can also be used to select the most relevant features from each modality before concatenation, which helps to eliminate redundant or irrelevant information and improves the model performance. Features should be normalized to a common scale so that no single modality dominates the learning process due to differences in feature value ranges.

b) Operation-Based Fusion:

It is a feature-level fusion method where features of different modalities are combined using element-wise mathematical operations such as addition, multiplication, or averaging. The operations are performed on corresponding elements in two or more feature vectors of the same dimensions [11] (see Figure 3(b)). Operation-based fusion captures patterns of similarity, interaction, or synergy between modalities by emphasizing relationships between aligned features and producing a composite feature vector integrating the fused information. Contrary to concatenation, which involves a simple stacking of feature vectors and thus increase in dimensionality, operation-based fusion requires its feature vectors to be of the same shape [11] and applies element-wise or channel-wise operations

 doi:10.20944/preprints202512.1445.v1

directly, often leading to a much more compact representation. This approach pays attention to shared or complementary information in modalities while keeping computational efficiency.

In operation-based fusion, feature vectors $\mathbf{x}^{(m)} \in \mathbb{R}^d$ are extracted from each modality m in the set of modalities $\mathcal{M}$. The extracted features are combined using an element-wise operation. For two modalities m_1 and m_2, the combined feature vector $\mathbf{z}$ can be computed as:

$$\mathbf{z} = \mathbf{x}^{(m_1)} \odot \mathbf{x}^{(m_2)},$$

where $\odot$ represents an element-wise operation such as:

- **Addition:**
$$z_i = x_i^{(m_1)} + x_i^{(m_2)},$$
- **Multiplication:**
$$z_i = x_i^{(m_1)} \times x_i^{(m_2)},$$
- **Averaging:**
$$z_i = \frac{x_i^{(m_1)} + x_i^{(m_2)}}{2}.$$

For multiple modalities, these operations can be generalized to combine all feature vectors element-wise. For instance, an average-based fusion for M modalities can be expressed as:

$$\mathbf{z} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}^{(m)}.$$

The resulting composite feature vector $\mathbf{z}$ is then used as input to a machine learning model f to predict the target variable y:

$$y = f(\mathbf{z}; \theta),$$

where θ represents the model parameters.

In more advanced forms, operation-based fusion extends to channel-wise multiplication, where entire channels are multiplied (treating each channel as a single entity, or specific dimensions in a multi-dimensional array). For example, the technique has proven effective in CNNs for combining multi-channel feature maps in tasks such as image classification and medical image analysis [130, 131]. Furthermore, tensor-based fusion applies outer products between feature vectors in order to encode higher-order interactions, providing a richer representation but with increased computational complexity [8,123,132]. Examples include [8] where the authors used tensor-based fusion to correlate pathological images with genomics data to improve diagnostic accuracy. Another variation involves attention-based fusion, where one feature vector is used as attention weights for another. This allows the model to focus on important features of one modality, as indicated by knowledge from the other, thus strengthening the representational power of the fused vector [33,123,133,134]. Common applications of these attention mechanisms include tasks like image-caption generation [135] and audio-visual emotion recognition [136].

c) Learning-Based Fusion:

Learning-based fusion is a more advanced approach to the integration of multi-modal data, where machine learning models [44] are used to learn an optimal way of combining features from different modalities (see Figure (3c)). This approach goes beyond simple methods like concatenation or operation-based fusion, since algorithms and architectures are designed to automatically identify relationships and interactions between modalities. These models are especially good at capturing complex nonlinear relationships and generalizing across different types of data and tasks.

 doi:10.20944/preprints202512.1445.v1

In learning-based fusion, the process begins with feature extraction for each modality. For a set of modalities $\mathcal{M}$, feature vectors $\mathbf{x}^{(m)} \in \mathbb{R}^{d_m}$ are extracted from each modality m, where d_m is the dimensionality of the feature space for modality m. These features are then passed through a fusion model, trained to learn joint representation. The fusion model may take a number of forms, depending on the task at hand and the nature of the data.

One common approach is the use of autoencoders, which are neural networks trained to reconstruct input features. In multimodal learning, autoencoders are extended to process multiple modalities simultaneously [137]. Let $\mathbf{z}$ represent the joint representation learned by the autoencoder, obtained as:

$$\mathbf{z} = f_{\text{encoder}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}; \theta\right),$$

where f_{encoder} is the encoder function parameterized by θ. The goal is to minimize the reconstruction error:

$$\min_{\theta} \left\| \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)} - f_{\text{decoder}}(\mathbf{z}; \theta) \right\|_2^2,$$

where f_{decoder} is the decoder function.

Another method is Canonical Correlation Analysis (CCA) [138,139], which finds linear projections of the feature vectors from each modality into a shared latent space such that the correlations between the projected features are maximized. For two modalities $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, the projections $\mathbf{u} = \mathbf{W}_1^\top \mathbf{x}^{(1)}$ and $\mathbf{v} = \mathbf{W}_2^\top \mathbf{x}^{(2)}$ are learned by maximizing:

$$\text{corr}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}.$$

CCA and its non-linear extensions, such as Deep CCA [139], are widely used for learning joint representations in multimodal data.

Neural networks, and particularly those with attention mechanisms, constitute another strong learning-based fusion tool. This will learn to highlight the feature in each modality that most helps the fusion by attributing the highest weights to those. For example, the joint representation $\mathbf{z}$ can be computed as:

$$\mathbf{z} = \sum_{m=1}^{M} \alpha^{(m)} \mathbf{x}^{(m)},$$

where $\alpha^{(m)}$ are attention weights learned during training, often computed as:

$$\alpha^{(m)} = \text{softmax}\left(\mathbf{W}^\top \mathbf{x}^{(m)}\right),$$

with $\mathbf{W}$ as learnable parameters.

Once the fusion model has learned the joint representation $\mathbf{z}$, this representation is used as input for downstream tasks such as classification, regression, or clustering. The combined feature representation enables the model to leverage complementary information from multiple modalities effectively.

The key advantage of learning-based fusion is that it can learn the best fusion strategy directly from data without pre-defined operations. That imbues adaptability into the model, enabling the learning of complex nonlinear relations between the modalities, hence suitable for a large number of tasks and types of data. Moreover, it is capable of dynamic adaptation to the modalities with different levels of importance or reliability, like in the case of attention-based approaches. Nevertheless, there are some challenges associated with learning-based fusion. The increased model complexity often requires more data and computational resources for training and bears a higher risk of overfitting, especially for small or imbalanced training datasets. It also takes more time to design and tune such models

 doi:10.20944/preprints202512.1445.v1

compared with simpler fusion methods. It has to be noted that the data requirements and overfitting problems common in multimodal fusion are generally overcome using data augmentation, transfer learning, and regularization. For example, pretraining on large datasets followed by fine-tuning on multimodal tasks will help alleviate overfitting and result in better generalization. Besides, hybrid approaches, by which learning-based fusion is combined with simpler methods like concatenation, can reduce computational costs while maintaining performance.

3. Advanced Multimodal Machine Learning

Over the past decade, advances in representation learning, neural architectures, and large-scale training methodologies have led to incredible progress in MML. This has supported applications ranging from vision-language understanding and medical diagnostics to multimedia retrieval, human-computer interaction, and robotics. This section covers state-of-the-art approaches in MML, with special emphasis on attention mechanisms, cross-modal embeddings, generative models, and graph neural networks. We review their recent advances, point to representative works, and outline significant challenges and future directions.

3.1. Attention Mechanisms for Multimodal Integration

Attention mechanisms have been a powerful tool for selectively focusing on the most relevant parts of input data, hence providing dynamic and contextual integration across multiple modalities. Based on the Transformer architecture [140], attention enables models to go beyond the uniform encoding strategies and assign higher weights to salient features, tokens, or regions. In this way, the attention-based multimodal models are better equipped to represent complex relationships between the modalities compared to the traditional fusion methods.

The Transformer [140] fundamentally changed the paradigm in sequence modeling from recurrent to an attention-centric one. Although developed for text initially, the conceptual framework of self-attention and cross-attention proved to be readily adaptable to multi-modal data. Early work in multi-modal attention began with image captioning tasks, where models like "Show, Attend and Tell" [141] used spatial attention over image regions as the model generated descriptive captions. The model learned to "attend" to specific parts of an image at each word generation step, aligning visual features with linguistic concepts.

In Visual Question Answering (VQA), attention is used to align the relevant parts of an image with the respective words in a question. The work of Bottom-Up and Top-Down Attention [142] introduced object-level attention by extracting region features (detected objects) and a top-down mechanism for selecting the most informative visual elements conditioned on the question. Follow-up work leveraged Transformers to process visual and textual input jointly. For example, LXMERT [121] and ViLBERT [120] proposed to use separate streams for vision and language and then applied co-attention layers to enable cross-modal interaction. These methods achieved significant improvements on benchmarks such as VQA v2 and GQA.

More recent models have integrated modalities more tightly by directly applying transformers to both text and image patches, eschewing dependence on external object detection frameworks. ViLT (Vision-and-Language Transformer) [143] eliminated region-based features by directly processing raw image patches along with textual tokens in a single transformer. This simplification reduced computational overhead and improved efficiency. ALBEF (Align Before Fuse) [144] combined contrastive learning with attention-based fusion and aligned textual and visual space before fusing them, which achieved strong performance on the downstream tasks of image-text retrieval and VQA. Likewise, UNITER [145] pre-trained a single Transformer model on multiple vision-language tasks simultaneously and utilized attention to associate textual tokens with their corresponding visual regions based on learned alignments. Attention mechanisms have also proven invaluable in more dynamic multimodal tasks. In vision-language navigation [146], an agent follows language instructions to navigate through a simulated 3D environment. Attention allows the agent to associate particular phrases in the instruction with visual information in its current field of view, so that it can perform the

 doi:10.20944/preprints202512.1445.v1

appropriate actions given the context. Attention comes in handy for tasks that are audio-visual, like video question answering or audio-visual speech recognition (AVSR), by matching spoken words or environmental sounds with their corresponding visual frames.

Beyond simple token-level or patch-level attention, recent work has also considered hierarchical attention mechanisms that operate over more than one level of granularity. For instance, hierarchical attention can decide first which modality or data source is most helpful at this moment and select relevant features for that modality. For instance, Zou et al. [147] introduces a hierarchical attention network to model the structure in each modality and the correlations across modalities, hence capturing the complicated interactions in multimodal data. Similarly, the HAMLET framework utilizes a hierarchical attention mechanism for disentangling unimodal features in computing multimodal representations for tasks such as human activity recognition [148].

While attention-based multimodal models have been shown to achieve impressive feats, they do not come without challenges. Especially, learning stable and interpretable attention patterns is usually hard, which becomes even more difficult when the number of modalities and data complexity increases. There is a further need for developing stronger evaluation metrics that can tell whether models really learn meaningful cross-modal correspondences or exploit dataset biases. Furthermore, the most promising direction is to combine attention with other paradigms like reinforcement learning for interactive tasks or graph neural networks for structured reasoning.

3.2. Cross-Modal Embeddings and Alignment

At the core of MML lies the problem of aligning heterogeneous data sources—images, text, audio, structured signals—into a common latent space. In this space, cross-modal comparison, retrieval, and joint reasoning are greatly simplified. Historical approaches like CCA [149] and Kernel CCA [150] pioneered foundational methods that gave linear or kernel-based projections to correlate pairs of modalities. However, these traditional approaches usually suffered from the presence of complex, high-dimensional, and nonlinear relationships in real-world multimodal data. Advances in deep learning allowed models to capture significantly more complex correlations and nonlinear relationships between modalities. Early deep multimodal embeddings used either stacked autoencoders or deep canonical correlation variants to learn joint representations of paired inputs [139]. By optimizing objectives that encourage correlated latent features, these methods transformed the problem of cross-modal alignment from shallow, linear projections to deep, expressive embeddings.

Recent breakthroughs have been achieved by large-scale training on gigantic, weakly-labeled datasets of image-text pairs crawled from the web. For instance, CLIP (Contrastive Language-Image Pretraining) [151] and ALIGN [152] are trained on hundreds of millions of image-text pairs. Using a contrastive objective—whereby the model is encouraged to bring closer the embeddings of matching image-text pairs and push apart those of mismatched pairs—these models learn robust, semantic-rich embeddings. Crucially, CLIP and ALIGN enable zero-shot transfer: without explicit fine-tuning, they are found to perform competitively in many downstream tasks, ranging from image classification to retrieval, by simply prompting the text encoder with adequate class descriptions or queries. Such a large-scale pretraining paradigm revolutionized the field of image-text alignment by making models generalize well beyond their training distribution. Furthermore, these embeddings provide backbones for a broad class of applications—image retrieval, caption generation, and visual question answering—that largely obviate the requirement for task-specific labeled data.

Outside of general web data, multimodal embeddings have been promising in specialized domains. In healthcare, the integration of clinical notes, EHRs, and medical imaging presents unique challenges; learned embeddings, however, can bring these diverse data types into a single semantic space, aiding in disease diagnosis, prognosis modeling, patient stratification, and personalized treatment planning [153]. For example, representations that combine chest X-ray images and radiology reports allow for faster and more accurate triaging and decrease the cognitive burden on physicians. Likewise, embeddings of genomic data combined with imaging and textual EHR notes might help to pinpoint biomarkers associated with particular disease subtypes. As multimodal embeddings are

 doi:10.20944/preprints202512.1445.v1

increasingly applied to a lot of applications, ensuring their robustness, interpretability, and fairness has been a major goal of research [99,154]. Another major challenge is interpretability; although multimodal embeddings capture the semantic relationships, it is very difficult to exactly understand why some items are clustered and how the model aligns the features across modalities [155,156]. There is a growing need for further research to make multimodal embeddings understandable to both domain experts and end-users.

3.3. Generative Models for Multimodal Data Synthesis

Generative modeling takes MML beyond purely discriminative tasks and toward creative synthesis and transformation, enabling models to generate new data in multiple modalities. This capability opens doors to a variety of applications, including text-to-image generation, image-to-text captioning, and even video generation conditioned on textual descriptions. It allows the researcher to fill in missing modalities, to augment scarce datasets, and to create more flexible and adaptive multimodal systems using generative frameworks.

Foundational generative models, such as Variational Autoencoders (VAEs) [157] and Generative Adversarial Networks (GANs) [158], laid the foundations for learning latent representations from which new data points could be sampled. The first multimodal extensions of these models looked at basic modality pairs, e.g., images with textual descriptions or the combination of audio and video. Through its ability to encode shared latent factors across modalities, such approaches made possible cross-modal generation, e.g., generating images from text or completing missing audio tracks given visual context. Text-to-image synthesis has been one of the most visible areas of rapid progress. The first approaches were not capable of generating coherent or detailed images; however, with better architectures and training strategies, there are now models like DALL·E [159] and Stable Diffusion [160] that have capabilities never seen before. DALL·E maps natural language prompts into diverse, high-fidelity images, capturing very fine details and even complicated semantic relations. Stable Diffusion helps to redefine generative processing via diffusion models, achieving remarkable clarity and consistency even in challenging prompts. More generally, text-to-image models can be used as basic building blocks for more complex multimodal pipelines, e.g., generating images for downstream tasks such as image captioning or video summarization. The more challenging generation of videos from textual or audio descriptions has also seen progress, enabling the creation of synthetic video clips for training data augmentation. One of the powerful aspects of multimodal generative models is that they can be used to synthesize missing modalities. For example, in such situations when only a subset of modalities is available at inference time, generative models may "impute" or hallucinate the data that is missing in order to create full multimodal input for further processing. Early work by Ngiam et al. [100] showed how multimodal deep learning could reconstruct missing audio from video or vice versa, while more recent methods extend these ideas to more complex modalities.

Multimodal generative models help to alleviate data scarcity through data augmentation. Under conditions where large-scale dataset collections are infeasible—because of privacy constraints, the rarity of some medical conditions, or the cost of specialized equipment—synthetic samples can be generated to enrich the training distribution with the help of generative models. The synthetic samples, if carefully controlled and validated, could help improve model robustness and generalization.

Despite recent rapid progress, there remain several notable challenges that generative multimodal models face. The first is that ensuring fidelity and coherence across all modalities is non-trivial, especially since the systems deal with high-dimensional data. Another pressing challenge is that of quality assurance and validation; in sensitive domains such as healthcare, generated data must be plausible but also preserve critical medical properties so that reliable clinical decisions can be supported. Medical image synthesis and translation approaches will need to embrace domain knowledge and uncertainty quantification for results that are both safe and clinically meaningful. [161]. Moreover, as generative models increasingly interact with diverse populations and data distributions, fairness and bias control must be addressed. This is an active area of research on techniques for detecting,

 doi:10.20944/preprints202512.1445.v1

mitigating, and explaining biases to ensure that generated outputs do not inadvertently perpetuate societal stereotypes or mislead end-users [162].

3.4. Graph Neural Networks for Structured Multimodal Reasoning

While attention mechanisms and Transformers are incredibly effective at modeling sequential and pairwise relationships, many real-world tasks involve complicated relational structures that go beyond simple token-level interactions. Graph Neural Networks (GNNs) [163] provide a natural framework for modeling such relational data, allowing MML systems to represent entities, concepts, and events as nodes in a graph with edges capturing their relationships. Through this structure, GNNs can combine several sources coming from different modalities—images, text, audio, and sensor data—into a single model that explicitly represents relational information.

GNNs generalize message passing and convolutional operations to irregular graph domains. Architectures like Graph Convolutional Networks (GCNs) [164], Graph Attention Networks (GATs) [165], Graph Isomorphism Networks (GINs) [166], and GraphSAGE [167] enable each node to iteratively aggregate information from its neighbors. Stacking multiple layers allows the model to propagate features across the graph, capturing higher-order dependencies. These approaches have demonstrated effectiveness in node classification, link prediction, and graph-level classification tasks in both unimodal and multimodal settings.

In multimodal settings, GNNs can encode complex relational structures by building heterogeneous or multimodal graphs. For example, in vision-language tasks, nodes can represent image regions, objects, textual tokens, or semantic concepts, while the edges represent spatial proximity, co-occurrence, semantic similarity, or syntactic relationships [168]. Messages passed through the edges allow the GNN to learn joint representations that fuse information across modalities and capture intricate interactions that might be missed by sequence-based models alone.

Many studies apply GNNs to fuse visual and textual information for tasks such as visual question answering (VQA) and video reasoning. For example, multimodal graph reasoning methods construct graphs from image objects and question words so that the model can perform reasoning over object-object and object-language relations [145,169]. This approach brings improvement in performance on benchmarks that require detailed relational understanding. Still, GNN-based multimodal models suffer from a set of challenges. Scaling GNNs to massive, dynamic graphs calls for efficient sampling, incremental training, and distributed computation. Noise, incomplete data, and uncertainty management continue to be crucial, as real-world multimodal data are typically messy and changing [170]. Techniques for robust training, adversarial defense, and uncertainty quantification is an active research area. In addition, future work may investigate how to best fuse multiple data types into a single graph, how to adapt graphs when adding new modalities or data sources, and how to learn graph structures end-to-end from raw input. As GNN-based methods mature, their integration in multi-modal foundation models and large-scale pretraining paradigms could lead to the development of yet more expressive and versatile systems capable of structured reasoning in a wide range of applications.

4. Applications of Multimodal Machine Learning in Healthcare

The integration of multi-modal deep learning techniques in healthcare has led to significant advancements across various domains. By leveraging heterogeneous data sources, these models provide more comprehensive insights into patient health, enhance diagnostic accuracy, personalize treatment plans, and support clinical decision-making. This section explores the key applications of multi-modal deep learning in healthcare.

4.1. Multimodality Approaches in Brain Disorder

Recent research in multimodal deep learning for medical image analysis has leveraged a broad spectrum of data types to enhance predictive tasks across various neurological and psychiatric conditions. For instance, Parisot et al. [32] integrated T1-weighted structural MRI and pheno-

 doi:10.20944/preprints202512.1445.v1

typic/demographic features to construct population graphs for classifying both autism spectrum disorder and Alzheimer's disease (AD). Similarly, Huang and Chung [171] introduced an edge-variational GCN framework that uses structural imaging data along with uncertainty-aware graph connections, thereby strengthening disease prediction robustness. In the context of longitudinal modeling, Li and Fan [172] combined baseline hippocampal MRI and 1-year follow-up cognitive assessments in an RNN to forecast early progression to AD dementia.

Moving into more complex multimodal fusions, Dwivedi et al. [125] exploited MRI, PET, and clinical/neuropsychological measures in a deep learning network for AD diagnosis, while Zhou et al. [173] employed a stage-wise deep architecture to systematically fuse structural MRI, PET, and potentially cerebrospinal fluid (CSF) biomarkers for dementia classification. Studies focused on missing or incomplete data—such as Thung et al. [174] demonstrated that multi-task deep learning can effectively handle datasets where some subjects lack certain imaging modalities, improving multi-stage AD diagnosis. Along the same lines, El-Sappagh et al. [175] combined time series data from cognitive tests, clinical evaluations, and imaging biomarkers to detect AD progression stages more accurately.

Early work from Suk et al. [102] showcased how stacked autoencoders, trained on MRI and PET scans, could learn hierarchical features for AD vs. mild cognitive impairment (MCI) classification. Further advances came from Spasov et al. [176], who devised a multimodal CNN to jointly analyze MRI and PET data for more robust AD prediction. To incorporate clinical or demographic attributes in imaging-based models, Pölsterl et al. [33] introduced a dynamic affine feature map transform that fuses 3D MRI with tabular features, adapting the network's spatial transformations based on each patient's unique profile. Venugopalan et al. [177] also combined MRI, PET, and clinical measures (e.g., demographics, neuropsychological scores) in a unified deep learning pipeline to identify early AD stages.

Beyond AD and dementia, Achalia et al. [178] demonstrated that combining neuroimaging (both structural and functional) and neurocognitive measures can yield predictive biomarkers for bipolar disorder. In a related vein, Ceccarelli and Mahmoud [73] used multimodal temporal signals—including behavioral and physiological data—to recognize bipolar disorder and depression states. Qiu et al. [179] showed that fusing structural MRI with standard cognitive evaluations (MMSE and logical memory tests) amplifies the detection of MCI. From the perspective of multiple sclerosis (MS), Yoo et al. [35] focused on user-defined MRI-based lesion features and additional clinical metrics to predict which patients with clinically isolated syndrome (CIS) would convert to full-blown MS. Other researchers, such as Ghosal et al. [180], expanded into imaging-genetics by combining brain scans with genetic data (e.g., single nucleotide polymorphisms) to uncover biologically meaningful disease signatures. Moreover, Zheng et al. [71] applied a multiscale deep neural network to EEG and clinical variables for predicting neurological outcomes in comatose patients after cardiac arrest.

Across these studies, the primary prediction tasks include binary or multi-class disease classification (e.g., AD vs. controls, bipolar vs. healthy), disease staging (e.g., MCI vs. AD), or forecasting transition/conversion risks (e.g., MCI to AD, CIS to MS). Some works explore mood-episode detection in bipolar or depression, while others target short- and long-term prognosis, such as neurological recovery in intensive care settings. By integrating structural/functional imaging, clinical assessments, cognitive test scores, and even genetic factors, these investigations consistently affirm that multimodal approaches enhance predictive accuracy, interpretability, and clinical applicability compared to unimodal models.

4.2. Multimodality Approaches in Cancer Prediction

Nie et al. [181] used multi-modal neuroimaging data, such as T1- and T2-FLAIR MRI scans, to train a multi-channel 3D deep learning model for predicting survival time in brain tumor patients. In a similar vein, Braman et al. [132] fused radiology, pathology, genomic, and clinical information to discover multimodal prognostic biomarkers for improving cancer outcome predictions. Focusing on breast cancer, Duanmu et al. [182] leveraged integrative imaging (MRI), molecular profiles, and demographic data in a deep learning framework to estimate pathological complete response to

 doi:10.20944/preprints202512.1445.v1

neoadjuvant chemotherapy. Yala et al. [183] constructed a mammography-based deep model for enhanced breast cancer risk assessment, while Yan et al. [127] proposed a richer fusion network that combines multiple imaging and non-imaging features to improve breast cancer classification. Liu and Hu [184] employed denoising autoencoders on genomic data to extract deep genomic features associated with breast cancer subtypes, and Li et al. [185] fused pathological images with genomic data to predict breast cancer survival outcomes. Similarly, Holste et al. [126] demonstrated an end-to-end approach for fusing breast MRI with tabular clinical descriptors to boost classification accuracy. Kharazmi et al. [186] targeted basal cell carcinoma detection by combining dermoscopic images and patient profiles in a feature-fusion system, whereas Hyun et al. [187] applied PET-based radiomics to distinguish histological subtypes in lung cancer. Vanguri et al. [85] expanded the scope of multimodal fusion—integrating radiology images, pathology slides, and genomic features—to predict response to PD-(L)1 blockade therapy in non-small cell lung cancer. Vale Silva and Rohr [133] developed a pan-cancer prognosis model using multimodal deep learning, an approach echoed by Cheerla and Gevaert [188], who combined multi-omic data for pan-cancer survival prediction. Beyond these examples, Rubinstein et al. [189] introduced an unsupervised technique for tumor detection in dynamic PET/CT scans of the prostate, while Reda et al. [84] highlighted how deep learning can aid in the early diagnosis of prostate cancer. Schulz et al. [134] similarly deployed a multimodal deep learning pipeline to forecast prognosis in renal cancer patients. Guo et al. [190] provided a broader perspective on deep learning-based segmentation techniques for multimodal medical imaging, illustrating the central role of robust image analysis in supporting clinical workflows. Lastly, Chen et al. [8] presented "Pathomic Fusion," an integrated framework that unites histopathology images and genomic features for both cancer diagnosis and outcome prediction, underscoring the power of data-driven multimodal strategies across diverse oncology applications.

4.3. Multimodality Approaches in Chest Related Conditions

Palepu and Beam [191] developed TIER (Text-Image Entropy Regularization), a method designed for CLIP-style vision-language models that integrates a learned entropy penalty into the contrastive training objective. Their work demonstrates how carefully controlling feature entropy can enhance the alignment of text and image representations, leading to improved robustness and interpretability in multimodal tasks. Duvieusart et al. [16] addressed cardiomegaly classification by extracting digital biomarkers from chest X-rays—such as heart size indices—and merging them with patient metadata (e.g., vital signs, laboratory values). Their multimodal approach outperformed imaging-only baselines, highlighting the value of combining subtle radiographic cues with non-imaging clinical features for more accurate detection of enlarged hearts. Bagheri et al. [192] tackled cardiovascular risk prediction by building a multimodal model around EHR data, including both structured data (e.g., diagnoses, medication history) and unstructured text. Their system leveraged deep learning architectures to capture complex interactions between demographic variables, comorbidities, and other risk factors, thereby offering more precise predictions for potential cardiovascular events. Similarly, Grant et al. [193] proposed a deep neural network for detecting cardiomegaly in an ICU setting. In addition to analyzing chest radiographs, the model incorporated ICU-specific information such as vital sign trends, ventilator settings, and lab results. This integrated design allowed the authors to identify critical risk patterns that purely image-based methods might overlook, thereby improving classification performance. By contrast, Baltruschat et al. [194] conducted a comprehensive evaluation of multiple deep learning architectures for multi-label pathology classification on chest X-ray datasets (e.g., ChestX-ray14). Their comparison encompassed convolutional networks and transfer learning setups, ultimately providing guidelines on which configurations performed best across different pathological findings, such as cardiomegaly, effusion, and infiltration. In the context of acute ischemic stroke, Brugnara et al. [195] built a multimodal machine-learning framework that incorporated CT imaging, perfusion maps, and clinical factors (e.g., stroke severity scores, time since symptom onset) to predict patient outcomes after endovascular treatment. Their results demonstrated that the fusion of neuroimaging and clinical variables improved prognostic accuracy over single-modality methods. Along similar lines, Samak et

 doi:10.20944/preprints202512.1445.v1

al. [25] combined clinical, radiological, and procedural data to forecast functional outcomes following thrombectomy, showing how integrated models can inform more personalized stroke management strategies. Walker et al. [66] tackled a different cardiac challenge—heart murmur detection—by introducing the Dual Bayesian ResNet. Their system leverages phonocardiogram signals (audio recordings) and Bayesian inference to handle uncertainty, showcasing how deep learning can detect subtle acoustic markers of valvular heart conditions. Meanwhile, Nishimori et al. [196] analyzed ECG signals, electrophysiology lab data, and clinical attributes using a multimodal deep neural network to localize accessory conduction pathways, an important step in treating arrhythmias such as Wolff-Parkinson-White syndrome. Chauhan et al. [197] concentrated on pulmonary edema assessment, jointly modeling chest X-ray images and corresponding radiology reports. By using natural language processing for text and CNN-based feature extraction for images, they learned a shared representation that yields more nuanced severity estimates than visual inspection or text parsing alone. In the realm of infectious diseases, Xu et al. [23] utilized a late fusion strategy that aggregates CT imaging features, clinical lab results, and demographic variables to distinguish COVID-19 patients from other viral pneumonia cases and healthy controls. Fang et al. [24] likewise applied deep learning to chest CT scans—alongside vital sign and lab data—to predict which COVID-19 patients were at higher risk of "malignant" or severe disease progression. Finally, Zhou et al. [198] introduced a cohesive multi-modality fusion network to estimate the severity of COVID-19 infection. Their model integrates CT-based lesion metrics, laboratory markers (e.g., blood oxygen levels), and demographic or clinical data through a carefully designed feature fusion pipeline. This holistic approach demonstrated superior performance in triaging patients according to severity risk, underscoring the continued importance of multimodal integration in critical care settings.

4.4. Multimodality Approaches in Skin Related Conditions and Other Diseases

Taleb et al. [199] presented a multimodal self-supervised learning strategy for medical image analysis, combining different imaging modalities under a shared representation space to reduce reliance on large labeled datasets. In a similar vein, Huang et al. [54] introduced GLoRIA, a global-local representation learning framework that links localized medical image features with corresponding text labels, enabling label-efficient medical image recognition. Addressing hematological disorders, Purwar et al. [200] leveraged CBC parameters and microscopic blood film images, extracting CNN-based features for classifying microcytic hypochromic anemia with various downstream classifiers. By contrast, Jin et al. [201] aimed to improve hospital mortality prediction through a multimodal architecture that fuses EHR data—including medical named entities—with other patient information, ultimately enhancing prediction accuracy. Salekin et al. [74] proposed a spatio-temporal deep learning model that integrates video, audio, and physiological data to assess postoperative pain in neonates, demonstrating the viability of multimodal inputs for more sensitive pain evaluation. Tiulpin et al. [202] merged standard radiographs with clinical variables to predict knee osteoarthritis progression; their machine learning model underscored how radiographic and patient metadata can provide complementary prognostic insights. Rodin et al. [203] introduced a multitask and multimodal neural network for X-ray interpretation, offering explainable outputs across multiple clinical tasks and emphasizing interpretability in medical AI systems. In dermatology, Yap et al. [47] utilized a multimodal deep learning framework that draws on dermoscopic images and metadata to enhance skin lesion classification, while Gessert et al. [48] demonstrated how multi-resolution EfficientNets and auxiliary patient data (e.g., lesion location, demographic information) can be ensembled for robust skin lesion classification. Finally, Kawahara et al. [45] extended a multitask multimodal approach using both clinical and dermoscopic imaging to implement the seven-point checklist for skin lesion analysis, showing how task-specific subnetworks can be trained in parallel to systematically address different diagnostic criteria.

5. Discussion and Future Directions

MML has developed into a strong paradigm, enabling more robust, accurate, and interpretable modeling in several application domains: medical imaging, language processing, robotics, and beyond.

 doi:10.20944/preprints202512.1445.v1

In the health domain, this integration of different modalities, such as MRI, PET, CT scans, EHR data, clinical notes, and genetic information, has led to significant gains in performance in disease diagnosis, prognosis, and patient management. Below, we distill the lessons learned from recent advances and identify open challenges and research directions that are promising for the next generation of MML systems.

5.1. Advantages of Multimodality in Healthcare

The studies reported in the preceding sections demonstrate that using multiple data sources almost invariably results in performance gains over unimodal systems. For example, combining structural imaging with clinical and genetic data leads to a much more holistic view of the patient's disease state, allowing for earlier diagnosis and more precise prognostic predictions. Integration of MRI/PET scans with cognitive tests or demographic variables in a multimodal fashion will reflect the underlying complex pathophysiological processes better than the single-modality approaches in diseases such as Alzheimer's disease or bipolar disorder. Similarly, integration of radiological and pathological information with genomic features in cancer prediction could provide insights about tumor heterogeneity and therapy response, which might not be possible through imaging alone.

5.2. Challenges in Attention-Based and Transformer Models

While attention mechanisms and Transformer-based architectures have revolutionized multimodal integration, allowing for fine-grained alignments at the token, patch, or image-region level, many practical and theoretical challenges persist:

- **Interpretability and Reliability:** Attention scores do not necessarily reflect true causal importance, and high-dimensional attention maps can be difficult to validate clinically. More robust interpretability strategies are needed for transparency.
- **Data Scale and Quality:** Transformers typically require large-scale, high-quality datasets. In health care, data are often siloed, limited in size, noisy, or otherwise difficult to scale in training. A few methods, such as self-supervised learning, efficient pretraining, and model distillation, can help mitigate these data bottlenecks.
- **Modality Balancing:** Differences in information density among modalities—for instance, rich imaging data versus sparse text notes—can skew attention and degrade downstream performance. Balancing the relative contributions of each modality remains a key research question.

5.3. Graph Neural Networks for Structured Reasoning

GNNs enable elegant encoding of structured relationships among entities. However, while successful in tasks such as visual question answering and disease progression modeling, GNN-based approaches also present their own set of challenges:

- **Graph Construction and Heterogeneity:** It is non-trivial to decide how to encode diverse data, be it images, clinical metrics, or genomic markers, as nodes or edges in a graph. Automating the process of graph construction that adapts to the diversity of clinical scenarios remains an active research area.
- **Scalability and Dynamic Graphs:** Large patient cohorts and real-time streams of data call for scalable GNNs, which can efficiently handle dynamic updates, new modalities, or newly acquired data for patients.
- **Uncertainty and Noise:** Real-world clinical data are usually incomplete or noisy. There is a strong need for effective uncertainty modeling and robust training strategies of GNNs to make reliable predictions.

5.4. Generative Models in Healthcare

Many possiblities emerge with VAEs, GANs, and Diffusion Models, like data augmentation, missing-modality completion, or synthetic data generation:

 doi:10.20944/preprints202512.1445.v1

- **Data Augmentation for Rare Conditions:** Generative models can synthesize realistic examples of rare diseases, which may help to mitigate class imbalance and improve the training of discriminative models.
- **Clinical Validity:** It is important that the generated samples retain medically valid features. Small deviations in synthetic medical images can have a huge impact on diagnosis or treatment planning downstream.
- **Ethical and Regulatory Concerns:** Synthetic data has to ensure the privacy of patients and meet regulatory standards. Methods of privacy-preserving generation—for example, through differential privacy—and transparent validation are vital for clinical adoption.

5.5. Multimodal Learning in Specialized Healthcare Domains

a) Neurological and Psychiatric Disorders:

Studies in Alzheimer's disease, multiple sclerosis, and bipolar disorder have demonstrated the need for longitudinal modeling and integration of complex data streams. Future work should focus on:

- **Longitudinal Consistency:** How to capture progressive and temporal features of neurodegenerative diseases using recurrent networks or temporal transformers.
- **Standardized and Open Data Repositories:** Good quality longitudinal datasets are still very limited. The creation of larger, more heterogeneous, and carefully annotated databases is thus important for model development and benchmarking.

b) Oncology and Cancer Prediction:

Recent studies emphasize the strong complementarity of imaging and genomic data in tumor subtyping and treatment response prediction. Next steps include:

- **Explainable AI for Oncology:** Clinicians require transparent explanations of model predictions when managing critical decisions like chemotherapy regimens or immunotherapies.
- **Integration of Liquid Biopsy and Proteomic Data:** Beyond imaging and genomics, molecular profiles (e.g., circulating tumor DNA) and proteomic features may further refine and personalize treatment strategies.

c) Cardiovascular and Pulmonary Applications:

Research in cardiomegaly detection and COVID-19 severity prediction underscores the importance of combining imaging with real-time vitals, lab results, and textual physician notes:

- **Streaming Data Integration:** Continuous patient monitoring devices produce dynamic, high-frequency data. Incorporating these signals into multimodal networks can facilitate early warning systems and preventive care.
- **Generalization to Low-Resource Settings:** Automated methods that perform reliably even where medical data is sparse or of lower quality (e.g., remote regions) can help address global healthcare disparities.

5.6. Interpretability, Fairness, and Ethical Considerations

As multi-modal models become increasingly complex, interpretability and fairness concerns come to the forefront. In healthcare, these considerations are paramount:

- **Human-Centered Interpretability:** Clinicians and patients need to understand the rationale behind a model's prediction, especially for high-stakes decisions. Techniques such as attention visualization, saliency maps, concept-based explanations, and post-hoc analysis can increase trust.
- **Bias and Fairness:** Disparities in dataset demographics can result in biased models that underperform in certain subpopulations. Addressing these issues may involve collecting more diverse datasets, performing bias audits, or adopting fairness-aware training objectives.

 doi:10.20944/preprints202512.1445.v1

- **Robustness and Safety:** Medical data can contain noise, artifacts, or adversarial corruption (e.g., sensor errors, malicious attacks). Ensuring robustness against such distortions is critical, particularly for real-world deployment in critical care environments.

5.7. Path Forward

Looking ahead, several key themes stand out:

1. **Unified Foundation Models in Healthcare:** Inspired by CLIP, ALIGN, and large language models, future research may seek to develop foundation models that can handle imaging, textual EHRs, laboratory data, and genetic information in a single framework. These models, trained on large and diverse datasets, can be adapted to a wide range of downstream healthcare tasks with minimal supervision.
2. **Causality and Counterfactual Reasoning:** Current MML approaches excel at correlational reasoning but often fail to capture causal relationships. Developing causal representation learning methods that can disentangle confounders and more accurately predict intervention outcomes remains a priority.
3. **Multimodal Reinforcement Learning (RL):** Interactive clinical tasks—such as robotic surgeries or therapy optimizations—may benefit from combining RL with multimodal understanding. Systems could learn to perform safe interventions by balancing real-time imaging, vital signs, and textual feedback from clinicians.
4. **Privacy-Preserving and Federated Learning:** As patient data typically reside in multiple institutions with strict privacy regulations, federated and privacy-preserving ML approaches are essential for building large-scale multimodal models without sharing sensitive patient information.

In sum, the future of multimodal learning in healthcare is both exciting and challenging. Continued advances in model architectures, representation learning, generative techniques, and robust evaluation frameworks promise to revolutionize clinical workflows, reduce diagnostic errors, and pave the way toward personalized medicine. By thoughtfully addressing interpretability, fairness, and ethical concerns, researchers can ensure that these powerful techniques truly benefit all patients and medical practitioners in a safe, transparent, and inclusive manner.

6. Conclusion

MML has rapidly evolved into a cornerstone of next-generation healthcare solutions by integrating diverse data types such as medical images, clinical notes, genomics, and sensor streams. From foundational approaches in fusion strategies—early, intermediate, late, and hybrid—to cutting-edge techniques involving attention-based transformers, cross-modal embeddings, generative modeling, and graph neural networks, MML has demonstrated remarkable capabilities in capturing the complexity of patient health. Across neurology, oncology, cardiopulmonary medicine, and beyond, these techniques have yielded notable improvements in diagnostic accuracy, disease staging, prognosis prediction, and clinical decision support. However, realizing MML's full potential demands navigating several critical challenges. Data heterogeneity, alignment, and robust fusion remain pivotal concerns, while interpretability, fairness, and ethical considerations are paramount in clinical environments. Moreover, as the size and diversity of medical datasets continue to grow, scalable architectures that can manage dynamic updates and heterogeneous data streams gain importance. Looking forward, the integration of MML with causal reasoning, reinforcement learning, and privacy-preserving technologies promises to unlock novel avenues for personalized and precision medicine. By proactively addressing these challenges—through interdisciplinary collaboration among clinicians, data scientists, and policymakers—MML stands poised to transform healthcare by improving patient outcomes, streamlining clinical workflows, and broadening equitable access to advanced data-driven insights.

 doi:10.20944/preprints202512.1445.v1

Author Contributions: Conceptualization, M.A.R and R.M.; methodology, M.A.R; software, M.A.R; validation, M.A.R and R.M.; formal analysis, M.A.R and R.M.; investigation, M.A.R and R.M.; resources, M.A.R and R.M.; writing—original draft preparation, M.A.R, P.S, M.S.H.A and R.M.; writing—review and editing, M.A.R, P.S, M.S.H.A and R.M.; visualization, M.A.R; supervision, R.M.; project administration, R.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; Dudley, J.T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* **2018**, *19*, 1236–1246.
2. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
3. Esteva, A.; et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118.
4. AlSaad, R.; Abd-alrazaq, A.A.; Boughorbel, S.; Ahmed, A.; Renault, M.A.; Damseh, R.R.; Sheikh, J. Multimodal large language models in health care: Applications, challenges, and future outlook. *Journal of Medical Internet Research* **2024**, *26*.
5. Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.Z. Big Data for Health. *IEEE Journal of Biomedical and Health Informatics* **2015**, *19*, 1193–1208.
6. de Gomez, M.R.C. A Comprehensive Introduction to Healthcare Data Analytics. *Journal of Biomedical and Sustainable Healthcare Applications* **2024**. n. pag.
7. Seneviratne, M.G.; Kahn, M.G.; Hernandez-Boussard, T. Merging heterogeneous clinical data to enable knowledge discovery. *Pac Symp Biocomput* **2019**, *24*, 439–443.
8. Chen, R.J.; Lu, M.Y.; Wang, J.; Williamson, D.F.; Rodig, S.J.; Lindeman, N.I.; Mahmood, F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging* **2020**, *41*, 757–770.
9. Warner, E.; Lee, J.; Hsu, W.; et al. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. *International Journal of Computer Vision* **2024**, *132*, 3753–3769. https://doi.org/10.1007/s11263-024-02032-8.
10. Bertsimas, D.; Ma, Y. M3H: Multimodal Multitask Machine Learning for Healthcare. *arXiv preprint arXiv:2404.18975* **2024**.
11. Krones, F.; Marikkar, U.; Parsons, G.; Szmul, A.; Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Information Fusion* **2025**, *114*, 102690.
12. England, N.H.S.; Improvement, N.H.S. Diagnostic imaging dataset statistical release. *Department of Health* **2016**, *421*.
13. Dendy, P.P.; Heaton, B. *Physics for diagnostic radiology*; CRC Press, 2011.
14. Brant, W.E.; Helms, C.A., Eds. *Fundamentals of diagnostic radiology*; Lippincott Williams & Wilkins (LWW), 2012.
15. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
16. Duvieusart, B.; Krones, F.; Parsons, G.; Tarassenko, L.; Papież, B.W.; Mahdi, A. Multimodal cardiomegaly classification with image-derived digital biomarkers. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis. Springer International Publishing, 2022, pp. 13–27.
17. Zhang, G.X. Machine learning in X-ray imaging and microscopy applications. *Advanced X-ray Imaging of Electrochemical Energy Materials and Devices* **2021**, pp. 205–221.
18. Goldman, L.W. Principles of CT and CT technology. *Journal of Nuclear Medicine Technology* **2007**, *35*, 115–128.
19. Adam, A.; Dixon, A.K.; Gillard, J.H.; Schaefer-Prokop, C. *Grainger & Allison's Diagnostic Radiology, 2 Volume Set E-Book*; Elsevier Health Sciences, 2020.
20. Bremner, D. *Brain imaging handbook*; WW Norton & Co, 2005.

 doi:10.20944/preprints202512.1445.v1

21. Schoepf, J.; Zwerner, P.; Savino, G.; Herzog, C.; Kerl, J.M.; Costello, P. Coronary CT angiography. *Radiology* **2007**, *244*, 48–63.
22. Doğan, H.; de Roos, A.; Geleijins, J.; Huisman, M.V.; Kroft, L.J.M. The role of computed tomography in the diagnosis of acute and chronic pulmonary embolism. *Diagnostic and Interventional Radiology* **2015**, *21*, 307.
23. Xu, M.; Ouyang, L.; Gao, Y.; Chen, Y.; Yu, T.; Li, Q.; Sun, K.; Bao, F.S.; Safarnejad, L.; Wen, J.; et al. Accurately differentiating COVID-19, other viral infection, and healthy individuals using multimodal features via late fusion learning. *medRxiv* **2020**.
24. Fang, C.; Bai, S.; Chen, Q.; Zhou, Y.; Xia, L.; Qin, L.; Gong, S.; Xie, X.; Zhou, C.; Tu, D.; et al. Deep learning for predicting COVID-19 malignant progression. *Medical Image Analysis* **2021**, *72*, 102096.
25. Samak, Z.A.; Clatworthy, P.; Mirmehdi, M. Prediction of thrombectomy functional outcomes using multimodal data. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis. Springer, 2020, pp. 267–279.
26. Brunelli, A.; Charloux, A.; Bolliger, C.T.; Rocco, G.; Sculier, J.P.; Varela, G.; Licker, M.J.; Ferguson, M.K.; Faivre-Finn, C.; Huber, R.M.; et al. ERS/ESTS clinical guidelines on fitness for radical therapy in lung cancer patients (surgery and chemo-radiotherapy). *European Respiratory Journal* **2009**, *34*, 17–41.
27. Wiener, R.S.; Schwartz, L.M.; Woloshin, S. When a test is too good: how CT pulmonary angiograms find pulmonary emboli that do not need to be found. *BMJ* **2013**, *347*.
28. Battista, J.J.; Rider, W.D.; Van Dyk, J. Computed tomography for radiotherapy planning. *International Journal of Radiation Oncology, Biology, Physics* **1980**, *6*, 99–107.
29. Grover, V.P.B.; Tognarelli, J.M.; Crossey, M.M.E.; Cox, I.J.; Taylor-Robinson, S.D.; McPhail, M.J.W. Magnetic resonance imaging: principles and techniques: lessons for clinicians. *Journal of Clinical and Experimental Hepatology* **2015**, *5*, 246–255.
30. Frisoni, G.B.; Fox, N.C.; Jack, C.R.; Scheltens, P.; Thompson, P.M. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology* **2010**, *6*, 67–77.
31. Guermazi, A.; Roemer, F.W.; Haugen, I.K.; Crema, M.D.; Hayashi, D. MRI-based semiquantitative scoring of joint pathology in osteoarthritis. *Nature Reviews Rheumatology* **2013**, *9*, 236–251.
32. Parisot, S.; Ktena, S.I.; Ferrante, E.; Lee, M.; Guerrero, R.; Glocker, B.; Rueckert, D. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Medical Image Analysis* **2018**, *48*, 117–130.
33. Pölsterl, S.; Wolf, T.N.; Wachinger, C. Combining 3D image and tabular data via the dynamic affine feature map transform. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 688–698.
34. Ryman, S.G.; Poston, K.L. MRI biomarkers of motor and non-motor symptoms in Parkinson's disease. *Parkinsonism & Related Disorders* **2020**, *73*, 85–93.
35. Yoo, Y.; Tang, L.Y.W.; Li, D.K.B.; Metz, L.; Kolind, S.; Traboulsee, A.L.; Tam, R.C. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **2019**, *7*, 250–259.
36. Stadler, A.; Schima, W.; Ba-Ssalamah, A.; Kettenbach, J.; Eisenhuber, E. Artifacts in body MR imaging: their appearance and how to eliminate them. *European Radiology* **2007**, *17*, 1242–1255.
37. Yu, D.; Sheikholeslami, G.; Zhang, A. FindOut: Finding outliers in very large datasets. *Knowledge and Information Systems* **2002**, *4*, 387–412.
38. Carovac, A.; Smajlovic, F.; Junuzovic, D. Application of ultrasound in medicine. *Acta Informatica Medica* **2011**, *19*, 168–171.
39. Merz, E.; Abramowicz, J.S. 3D/4D ultrasound in prenatal diagnosis: is it time for routine use? *Clinical Obstetrics and Gynecology* **2012**, *55*, 336–351.
40. Brattain, L.J.; Telfer, B.A.; Dhyani, M.; Grajo, J.R.; Samir, A.E. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal Radiology* **2018**, *43*, 786–799.
41. Karaoğlu, O.; Bilge, H.Ş.; Uluer, I. Removal of speckle noises from ultrasound images using five different deep learning networks. *Engineering Science and Technology, an International Journal* **2022**, *29*, 101030.
42. Vestergaard, M.E.; Macaskill, P.H.P.M.; Holt, P.E.; Menzies, S.W. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology* **2008**, *159*, 669–676.
43. Braun, R.P.; Rabinovitz, H.S.; Oliviero, M.; Kopf, A.W.; Saurat, J.H. Dermoscopy of pigmented skin lesions. *Journal of the American Academy of Dermatology* **2005**, *52*, 109–121.

 doi:10.20944/preprints202512.1445.v1

44. Raju, M.A.; Mia, M.S.; Sayed, M.A.; Uddin, M.R. Predicting the outcome of English Premier League matches using machine learning. In Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2020, pp. 1–6.
45. Kawahara, J.; Daneshvar, S.; Argenziano, G.; Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics* **2018**, *23*, 538–546.
46. Iqbal, I.; Younus, M.; Walayat, K.; Kakar, M.U.; Ma, J. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics* **2021**, *88*, 101843.
47. Yap, J.; Yolland, W.; Tschandl, P. Multimodal skin lesion classification using deep learning. *Experimental Dermatology* **2018**, *27*, 1261–1267.
48. Gessert, N.; Nielsen, M.; Shaikh, M.; Werner, R.; Schlaefer, A. Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX* **2020**, *7*, 100864.
49. Kittler, H.; Pehamberger, H.; Wolff, K.; Binder, M.J.T.I.O. Diagnostic accuracy of dermoscopy. *The Lancet Oncology* **2002**, *3*, 159–165.
50. Spasic, I.; Nenadic, G.; et al. Clinical text data in machine learning: systematic review. *JMIR Medical Informatics* **2020**, *8*, e17984.
51. Mustafa, A.; Azghadi, M.R. Automated machine learning for healthcare and clinical notes analysis. *Computers* **2021**, *10*, 24.
52. Li, Q.; Spooner, S.A.; Kaiser, M.; Lingren, N.; Robbins, J.; Lingren, T.; Solti, I.; Ni, Y. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC Medical Informatics and Decision Making* **2015**, *15*, 1–12.
53. Johnson, A.; Pollard, T.; Horng, S.; Celi, L.A.; Mark, R. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet* **2023**.
54. Huang, S.C.; Shen, L.; Lungren, M.P.; Yeung, S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3942–3951.
55. Casey, A.; Davidson, E.; Poon, M.; Dong, H.; Duma, D.; Grivas, A.; Grover, C.; Suárez-Paniagua, V.; Tobin, R.; Whiteley, W.; et al. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making* **2021**, *21*, 179.
56. Afzal, M.; Hussain, J.; Abbas, A.; Maqbool, H. Multi-class clinical text annotation and classification using BERT-based active learning. *Available at SSRN* **2022**, *4081033*.
57. Xie, X.; Niu, J.; Liu, X.; Chen, Z.; Tang, S.; Yu, S. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis* **2021**, p. 101985.
58. Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V.; et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Medical Informatics* **2019**, *7*, e12239.
59. Locke, S.; Bashall, A.; Al-Adely, S.; Moore, J.; Wilson, A.; Kitchen, G.B. Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care* **2021**, *38*, 4–9.
60. Chen, Y.; Lasko, T.A.; Mei, Q.; Denny, J.C.; Xu, H. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics* **2015**, *58*, 11–18.
61. Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; McLachlan, S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association* **2018**, *25*, 230–238.
62. Zeger, S.L.; Irizarry, R.A.; Peng, R.D. On time series analysis of public health and biomedical data. *Annual Review of Public Health* **2006**, *27*, 57–79.
63. Jarrett, D.; Yoon, J.; Bica, I.; Qian, Z.; Ercole, A.; van der Schaar, M. Clairvoyance: A pipeline toolkit for medical time series. *arXiv preprint arXiv:2310.18688* **2023**.
64. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent neural networks for multi-variate time series with missing values. *Scientific Reports* **2018**, *8*, 6085.
65. Liu, Z.; Wu, L.; Hauskrecht, M. Modeling clinical time series using Gaussian process sequences. In Proceedings of the Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2013, pp. 623–631.
66. Walker, B.; Krones, F.; Kiskin, I.; Parsons, G.; Lyons, T.; Mahdi, A. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. *Computing in Cardiology* **2022**.

 doi:10.20944/preprints202512.1445.v1

67. Han, J.H. Comparing models for time series analysis. *Wharton Research Scholars* **2018**.
68. Lee, Y.S.; Tong, L.I. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems* **2011**, *24*, 66–79.
69. Arumugam, V.; Natarajan, V. Time Series Modeling and Forecasting Using Autoregressive Integrated Moving Average and Seasonal Autoregressive Integrated Moving Average Models. *Instrumentation, Mesures, Métrologies* **2023**, *22*.
70. Kaushik, S.; Choudhury, A.; Sheron, P.K.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in Big Data* **2020**, *3*, 4.
71. Zheng, W.L.; Amorim, E.; Jing, J.; Ge, W.; Hong, S.; Wu, O.; Ghassemi, M.; Lee, J.W.; Sivaraju, A.; Pang, T.; et al. Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. *Resuscitation* **2021**, *169*, 86–94.
72. Morid, M.A.; Sheng, O.R.L.; Dunbar, J. Time series prediction using deep learning methods in healthcare. *ACM Transactions on Management Information Systems* **2023**, *14*, 1–29.
73. Ceccarelli, F.; Mahmoud, M. Multimodal temporal machine learning for bipolar disorder and depression recognition. *Pattern Analysis and Applications* **2022**, *25*, 493–504.
74. Salekin, M.S.; Zamzmi, G.; Goldgof, D.; Kasturi, R.; Ho, T.; Sun, Y. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in Biology and Medicine* **2021**, *129*, 104150.
75. M. Masud, M.; Hayawi, K.; Samuel Mathew, S.; Dirir, A.; Cheratta, M. Effective patient similarity computation for clinical decision support using time series and static data. In Proceedings of the Proceedings of the Australasian Computer Science Week Multiconference, 2020, pp. 1–8.
76. Di Martino, F.; Delmastro, F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review* **2023**, *56*, 5261–5315.
77. Knaus, W.A.; Draper, E.A.; Wagner, D.P.; Zimmerman, J.E. APACHE II: a severity of disease classification system. *Critical care medicine* **1985**, *13*, 818–829.
78. Pierson, E.; Cutler, D.M.; Leskovec, J.; Mullainathan, S.; Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine* **2021**, *27*, 136–140.
79. Herdman, M.; Gudex, C.; Lloyd, A.; Janssen, M.; Kind, P.; Parkin, D.; Bonsel, G.; Badia, X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research* **2011**, *20*, 1727–1736.
80. Pölsterl, S.; Wolf, T.N.; Wachinger, C. Combining 3D image and tabular data via the dynamic affine feature map transform. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer, 2021, pp. 688–698.
81. Krones, F.H.; Walker, B.; Parsons, G.; Lyons, T.; Mahdi, A. Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest. *arXiv preprint arXiv:2403.06027* **2024**.
82. Ayesha, S.; Hanif, M.K.; Talib, R. Performance enhancement of predictive analytics for health informatics using dimensionality reduction techniques and fusion frameworks. *IEEE Access* **2021**, *10*, 753–769.
83. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 6679–6687.
84. Reda, I.; Khalil, A.; Elmogy, M.; Abou El-Fetouh, A.; Shalaby, A.; Abou El-Ghar, M.; Elmaghraby, A.; Ghazal, M.; El-Baz, A. Deep learning role in early diagnosis of prostate cancer. *Technology in cancer research & treatment* **2018**, *17*, 1533034618775530.
85. Vanguri, R.S.; Luo, J.; Aukerman, A.T.; Egger, J.V.; Fong, C.J.; Horvat, N.; Pagano, A.; Araujo-Filho, J.d.A.B.; Geneslaw, L.; Rizvi, H.; et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. *Nature cancer* **2022**, *3*, 1151–1164.
86. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Scientific data* **2016**, *3*, 1–9.
87. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nature medicine* **2019**, *25*, 24–29.
88. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics* **2016**, *17*, 333–351.
89. Piwek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The rise of consumer health wearables: promises and barriers. *PLoS medicine* **2016**, *13*, e1001953.
90. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *New England Journal of Medicine* **2019**, *380*, 1347–1358.

 doi:10.20944/preprints202512.1445.v1

91. Shaik, T.; Tao, X.; Li, L.; Xie, H.; Velásquez, J.D. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion* **2024**, *102*, 102040.
92. Xiao, R.; Ding, C.; Hu, X. Time Synchronization of Multimodal Physiological Signals through Alignment of Common Signal Types and Its Technical Considerations in Digital Health. *Journal of Imaging* **2022**, *8*, 120.
93. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* **2018**, *8*, 6085.
94. Zhu, X.; Suk, H.I.; Shen, D. Multi-modality canonical feature selection for Alzheimer's disease diagnosis. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17. Springer, 2014, pp. 162–169.
95. Bannach, D.; Amft, O.; Lukowicz, P. Automatic event-based synchronization of multimodal data streams from wearable and ambient sensors. In Proceedings of the Smart Sensing and Context: 4th European Conference, EuroSSC 2009, Guildford, UK, September 16-18, 2009. Proceedings 4. Springer, 2009, pp. 135–148.
96. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* **2017**.
97. Zitova, B.; Flusser, J. Image registration methods: a survey. *Image and vision computing* **2003**, *21*, 977–1000.
98. Lipton, Z.C.; Kale, D.C.; Elkan, C.; Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* **2015**.
99. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **2018**, *41*, 423–443.
100. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y.; et al. Multimodal deep learning. In Proceedings of the ICML, 2011, Vol. 11, pp. 689–696.
101. Wang, Y.; Xu, X.; Yu, W.; Xu, R.; Cao, Z.; Shen, H.T. Combine early and late fusion together: A hybrid fusion framework for image-text matching. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021, pp. 1–6.
102. Suk, H.I.; Lee, S.W.; Shen, D.; Initiative, A.D.N.; et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* **2014**, *101*, 569–582.
103. Dean, J.; Corrado, G.; Monga, R.; Chen, K.; Devin, M.; Mao, M.; Ranzato, M.; Senior, A.; Tucker, P.; Yang, K.; et al. Large scale distributed deep networks. *Advances in neural information processing systems* **2012**, *25*.
104. Huang, S.C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* **2020**, *3*, 136.
105. Huang, S.C.; Pareek, A.; Zamanian, R.; Banerjee, I.; Lungren, M.P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports* **2020**, *10*, 22147.
106. Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; Luo, Y. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine* **2022**, *5*, 171.
107. Ayesha, S.; Hanif, M.K.; Talib, R. Performance enhancement of predictive analytics for health informatics using dimensionality reduction techniques and fusion frameworks. *IEEE Access* **2021**.
108. Dolly, J.M.; Nisa, A.K. A survey on different multimodal medical image fusion techniques and methods. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT). IEEE, 2019, pp. 1–5.
109. Hermessi, H.; Mourali, O.; Zagrouba, E. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing* **2021**, *183*, 108036.
110. Snoek, C.G.; Worring, M.; Smeulders, A.W. Early versus late fusion in semantic video analysis. In Proceedings of the Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 399–402.
111. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* **2010**, *16*, 345–379.
112. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* **2017**, *34*, 96–108.
113. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.

 doi:10.20944/preprints202512.1445.v1

114. Calixto, I.; Liu, Q.; Campbell, N. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287* **2017**.
115. Behrad, F.; Abadeh, M.S. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications* **2022**, *200*, 117006.
116. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Computation* **2020**, *32*, 829–864.
117. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* **2017**.
118. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv* **2022**, *abs/2205.12005*.
119. Xu, H.; Ye, Q.; Yan, M.; Shi, Y.; Ye, J.; Xu, Y.; Li, C.; Bi, B.; Qian, Q.; Wang, W.; et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv* **2023**, *abs/2302.00402*.
120. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Advances in Neural Information Processing Systems, 2019, Vol. 32.
121. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* **2019**.
122. Xu, X.; Wu, C.; Rosenman, S.; Lal, V.; Che, W.; Duan, N. Bridgetower: Building bridges between encoders in vision-language representation learning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, pp. 10637–10647.
123. Cui, C.; Yang, H.; Wang, Y.; Zhao, S.; Asad, Z.; Coburn, L.A.; Wilson, K.T.; Landman, B.A.; Huo, Y. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review. *arXiv* **2022**, *abs/2203.15588*.
124. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 248–255.
125. Dwivedi, S.; Goel, T.; Tanveer, M.; Murugan, R.; Sharma, R. Multi-modal fusion based deep learning network for effective diagnosis of Alzheimer's disease. *IEEE MultiMedia* **2022**.
126. Holste, G.; Partridge, S.C.; Rahbar, H.; Biswas, D.; Lee, C.I.; Alessio, A.M. End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3294–3303.
127. Yan, R.; Zhang, F.; Rao, X.; Lv, Z.; Li, J.; Zhang, L.; Liang, S.; Li, Y.; Ren, F.; Zheng, Chunhou, e.a. Richer fusion network for breast cancer classification based on multimodal data. *BMC Medical Informatics and Decision Making* **2021**, *21*, 1–15.
128. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2010**, *2*, 433–459.
129. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B.; Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear discriminant analysis. *Robust data mining* **2013**, pp. 27–33.
130. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *35*, 221–231.
131. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Medical image analysis* **2017**, *42*, 60–88.
132. Braman, N.; Gordon, J.W.; Goossens, E.T.; Willis, C.; Stumpe, M.C.; Venkataraman, J. Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 667–677.
133. Vale Silva, L.A.; Rohr, K. Pan-cancer prognosis prediction using multimodal deep learning. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020, pp. 568–571.
134. Schulz, S.; Woerl, A.; Jungmann, F.; Glasner, C.; Stenzel, P.; Strobl, S.; Fernandez, A.; Wagner, D.; Haferkamp, A.; Mildenberger, Peter, e.a. Multimodal deep learning for prognosis prediction in renal cancer. *Frontiers in Oncology* **2021**, *11*.
135. Agrawal, V.; Dhekane, S.; Tuniya, N.; Vyas, V. Image caption generator using attention mechanism. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2021, pp. 1–6.

 doi:10.20944/preprints202512.1445.v1

136. Ghaleb, E.; Niehues, J.; Asteriadis, S. Joint modelling of audio-visual cues using attention mechanisms for emotion recognition. *Multimedia Tools and Applications* **2023**, *82*, 11239–11264.
137. Jaques, N.; Taylor, S.; Sano, A.; Picard, R. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2017, pp. 202–208.
138. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 8992–8999.
139. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International conference on machine learning. PMLR, 2013, pp. 1247–1255.
140. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
141. Xu, K. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* **2015**.
142. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
143. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning. PMLR, 2021, pp. 5583–5594.
144. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. In Proceedings of the Advances in Neural Information Processing Systems, 2021, Vol. 34, pp. 9694–9705.
145. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the European Conference on Computer Vision, Cham, 2020; pp. 104–120.
146. Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.F.; Wang, W.Y.; Zhang, L. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6629–6638.
147. Zou, X.; Tang, C.; Zhang, W.; Sun, K.; Jiang, L. Hierarchical Attention Learning for Multimodal Classification. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023, pp. 936–941.
148. Islam, M.M.; Iqbal, T. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10285–10292.
149. Hardoon, D.R.; Szedmak, S.; Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* **2004**, *16*, 2639–2664.
150. Lai, P.L.; Fyfe, C. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* **2000**, *10*, 365–377.
151. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, 2021, pp. 8748–8763.
152. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning. PMLR, 2021, pp. 4904–4916.
153. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240.
154. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Proceedings of the Conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, Vol. 2019, p. 6558.
155. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Proceedings of the International Conference on Machine Learning. PMLR, 2018, pp. 2668–2677.
156. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215.
157. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* **2013**.

 doi:10.20944/preprints202512.1445.v1

158. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2014, Vol. 27.
159. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2021, pp. 8821–8831.
160. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
161. Nie, D.; Trullo, R.; Lian, J.; Wang, L.; Petitjean, C.; Ruan, S.; Wang, Q.; Shen, D. Medical Image Synthesis with Deep Convolutional Adversarial Networks. *IEEE Transactions on Biomedical Engineering* **2018**, *65*, 2720–2730.
162. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning Fair Representations. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2013, pp. 325–333.
163. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, *32*, 4–24.
164. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* **2016**.
165. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv preprint arXiv:1710.10903* **2017**.
166. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv preprint arXiv:1810.00826* **2018**.
167. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive Representation Learning on Large Graphs. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30.
168. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 684–699.
169. Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; Fu, J. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv preprint arXiv:2004.00849* **2020**.
170. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, *32*, 4–24.
171. Huang, Y.; Chung, A.C. Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 562–572.
172. Li, H.; Fan, Y. Early prediction of Alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 368–371.
173. Zhou, T.; Thung, K.H.; Zhu, X.; Shen, D. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human Brain Mapping* **2019**, *40*, 1001–1016.
174. Thung, K.H.; Yap, P.T.; Shen, D. Multi-stage diagnosis of Alzheimer's disease with incomplete multi-modal data via multi-task deep learning. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer, 2017; pp. 160–168.
175. El-Sappagh, S.; Abuhmed, T.; Islam, S.R.; Kwak, K.S. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing* **2020**, *412*, 197–215.
176. Spasov, S.E.; Passamonti, L.; Duggento, A.; Liò, P.; Toschi, N. A multi-modal convolutional neural network framework for the prediction of Alzheimer's disease. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 1271–1274.
177. Venugopalan, J.; Tong, L.; Hassanzadeh, H.R.; Wang, M.D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* **2021**, *11*, 1–13.
178. Achalia, R.; Sinha, A.; Jacob, A.; Achalia, G.; Kaginalkar, V.; Venkatasubramanian, G.; Rao, N.P. A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. *Asian Journal of Psychiatry* **2020**, *50*, 101984.
179. Qiu, S.; Chang, G.H.; Panagia, M.; Gopal, D.M.; Au, R.; Kolachalama, V.B. Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **2018**, *10*, 737–749.

 doi:10.20944/preprints202512.1445.v1

180. Ghosal, S.; Chen, Q.; Pergola, G.; et al. G-MIND: an end-to-end multimodal imaging-genetics framework for biomarker identification and disease classification. In Proceedings of the Medical Imaging 2021: Image Processing. SPIE, 2021, Vol. 11596, p. 115960C.
181. Nie, D.; Lu, J.; Zhang, H.; Adeli, E.; Wang, J.; Yu, Z.; Liu, L.; Wang, Q.; Wu, J.; Shen, D. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific Reports* **2019**, *9*, 1–14.
182. Duanmu, H.; Huang, P.B.; Brahmavar, S.; Lin, S.; Ren, T.; Kong, J.; Wang, F.; Duong, T.Q. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 242–252.
183. Yala, A.; Lehman, C.; Schuster, T.; Portnoi, T.; Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **2019**, *292*, 60–66.
184. Liu, Q.; Hu, P. Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer. *Cancers* **2019**, *11*, 494.
185. Li, S.; Shi, H.; Sui, D.; Hao, A.; Qin, H. A novel pathological images and genomic data fusion framework for breast cancer survival prediction. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020, pp. 1384–1387.
186. Kharazmi, P.; Kalia, S.; Lui, H.; Wang, J.; Lee, T. A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile. *Skin Research and Technology* **2018**, *24*, 256–264.
187. Hyun, S.H.; Ahn, M.S.; Koh, Y.W.; Lee, S.J. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clinical Nuclear Medicine* **2019**, *44*, 956–960.
188. Cheerla, A.; Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **2019**, *35*, i446–i454.
189. Rubinstein, E.; Salhov, M.; Nidam-Leshem, M.; White, V.; Golan, S.; Baniel, J.; Bernstine, H.; Groshar, D.; Averbuch, A. Unsupervised tumor detection in dynamic PET/CT imaging of the prostate. *Medical Image Analysis* **2019**, *55*, 27–40.
190. Guo, Z.; Li, X.; Huang, H.; Guo, N.; Li, Q. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences* **2019**, *3*, 162–169.
191. Palepu, A.; Beam, A.L. Tier: Text-image entropy regularization for clip-style models. *arXiv preprint arXiv:2212.06710* **2022**.
192. Bagheri, A.; Groenhof, T.K.J.; Veldhuis, W.B.; de Jong, P.A.; Asselbergs, F.W.; Oberski, D.L. Multimodal learning for cardiovascular risk prediction using ehr data. In Proceedings of the Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery, 2020, pp. New York, NY, USA.
193. Grant, D.; Papież, B.W.; Parsons, G.; Tarassenko, L.; Mahdi, A. Deep learning classification of cardiomegaly using combined imaging and non-imaging ICU data. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis. Springer, 2021, pp. 547–558.
194. Baltruschat, I.M.; Nickisch, H.; Grass, M.; Knopp, T.; Saalbach, A. Comparison of deep learning approaches for multi-label chest X-Ray classification. *Scientific Reports* **2019**, *9*, 1–10.
195. Brugnara, G.; Neuberger, U.; Mahmutoglu, M.A.; Foltyn, M.; Herweh, C.; Nagel, S.; Schönenberger, S.; Heiland, S.; Ulfert, C.; Ringleb, Peter Arthur, e.a. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke* **2020**, *51*, 3541–3551.
196. Nishimori, M.; Kiuchi, K.; Nishimura, K.; Kusano, K.; Yoshida, A.; Adachi, K.; Hirayama, Y.; Miyazaki, Y.; Fujiwara, R.; Sommer, Philipp, e.a. Accessory pathway analysis using a multimodal deep learning model. *Scientific Reports* **2021**, *11*, 1–8.
197. Chauhan, G.; Liao, R.; Wells, W.; Andreas, J.; Wang, X.; Berkowitz, S.; Horng, S.; Szolovits, P.; Golland, P. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 529–539.
198. Zhou, J.; Zhang, X.; Zhu, Z.; Lan, X.; Fu, L.; Wang, H.; Wen, H. Cohesive multi-modality feature learning and fusion for COVID-19 patient severity prediction. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 2535–2549.
199. Taleb, A.; Lippert, C.; Klein, T.; Nabi, M. Multimodal self-supervised learning for medical image analysis. In Proceedings of the Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings. Springer, 2021, pp. 661–673.

 doi:10.20944/preprints202512.1445.v1

200. Purwar, S.; Tripathi, R.K.; Ranjan, R.; Saxena, R. Detection of microcytic hypochromia using CBC and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications* **2020**, *79*, 4573–4595.
201. Jin, M.; Bahadori, M.T.; Colak, A.; Bhatia, P.; Celikkaya, B.; Bhakta, R.; Senthivel, S.; Khalilia, M.; Navarro, D.; Zhang, Borui, e.a. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276* **2018**.
202. Tiulpin, A.; Klein, S.; Bierma-Zeinstra, S.M.; Thevenot, J.; Rahtu, E.; van Meurs, J.; Oei, E.H.; Saarakkala, S. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific Reports* **2019**, *9*, 1–11.
203. Rodin, I.; Fedulova, I.; Shelmanov, A.; Dylov, D.V. Multitask and multimodal neural network model for interpretable analysis of X-Ray images. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 1601–1604.