Review

Multimodal Machine Learning in Healthcare: A Tutorial and Review

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

Posted Date: 3 June 2026

doi: 10.20944/preprints202512.1445.v2

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.v2

Review

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: Ruizhe_Ma@uml.edu

Abstract

Deep learning has transformed healthcare by enabling the analysis of complex, high-dimensional, and heterogeneous data. However, traditional unimodal approaches often fail to capture the multifaceted nature of human health, as patient information is inherently distributed across multiple data modalities. Multimodal machine learning (MML) has therefore emerged as a framework for integrating complementary sources such as medical images, clinical text, electronic health records (EHRs), and physiological signals to support more comprehensive modeling of health and disease. This narrative review provides a structured overview of MML in healthcare, focusing on representative data modalities, fusion strategies, advanced architectures, and clinically relevant design trade-offs. In particular, we distinguish between stage-based fusion strategies, which determine when modalities are combined, and feature integration mechanisms, which define how modality representations are merged. We synthesize applications across major domains, including brain disorders, cancer prediction, chest-related conditions, and skin diseases, highlighting both the potential benefits and the limitations of multimodal approaches. We further discuss key challenges related to data heterogeneity, cross-modal alignment, missing modalities, and the complexity of effective fusion design, along with broader issues in clinical translation. Finally, we outline future directions centered on foundation models, causal reasoning, privacy-preserving learning, and integrated healthcare data infrastructures.

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 reshaped modern healthcare by enabling the analysis of complex, high-dimensional data across tasks such as medical image interpretation, clinical prediction, and language understanding [1–3]. Yet many clinically important decisions do not depend on a single source of information. Diagnosing disease, estimating prognosis, and selecting treatment often require reasoning across imaging findings, clinical notes, laboratory values, physiological signals, demographics, and molecular profiles. Models built around only one modality can therefore miss interactions that are central to patient status and care [1,2]. This limitation has helped drive growing interest in multimodal machine learning (MML), which aims to integrate complementary data sources into a more complete representation of health and disease.

In healthcare, the motivation for MML is especially strong because the available data are inherently heterogeneous. Medical images capture anatomical and functional structure, clinical text records symptoms and physician reasoning, tabular records summarize demographics and laboratory results, time-series signals reflect dynamic physiology, and omics data provide molecular context [4–6]. When these modalities are meaningfully aligned, multimodal models can improve diagnosis, prognosis, risk stratification, and treatment planning by uncovering patterns that are difficult to recover from any modality in isolation [1,3,4,6]. Representative examples include combining pathology or radiology

 doi:10.20944/preprints202512.1445.v2

images with genomic data in precision oncology [7], integrating diverse clinical signals for decision support [8], and using multimodal multitask learning to improve performance across medical tasks [9]. At the same time, the value of MML depends not only on model capacity, but also on data quality, modality complementarity, alignment, and practical availability in real clinical workflows [10,11].

These considerations make healthcare MML both promising and challenging. Real-world clinical data are often incomplete, weakly synchronized, institution-specific, and collected under different protocols. As a result, the central questions are not merely how to fuse modalities, but when multimodality is genuinely useful, how to handle missing or mismatched inputs, and what is required for reliable clinical translation.

Motivated by this need, the present review provides a structured overview of MML in healthcare from both methodological and translational perspectives. We first outline representative healthcare data modalities and the main challenges they introduce, including heterogeneity, alignment, and missing-modality issues. We then examine stage-based fusion strategies, feature integration mechanisms, and more advanced approaches such as attention-based models, cross-modal embeddings, generative methods, and graph neural networks. Finally, we synthesize representative applications across major clinical domains and discuss the conditions under which multimodal systems are most likely to be clinically useful. By emphasizing both technical foundations and deployment realities, this review aims to clarify where MML adds value, where its limitations remain substantial, and what is needed for responsible clinical adoption.

2. Review Scope and Methodological Positioning

This study is conducted as a narrative review that aims to provide a structured and comprehensive overview of multimodal machine learning in healthcare. The literature was identified through broad searches of PubMed, Scopus, Web of Science, IEEE Xplore, and arXiv, with primary focus on publications from 2015 to 2025. The search included terms such as “multimodal machine learning”, “healthcare AI”, “data fusion”, “medical imaging”, “clinical text”, and related keywords.

Studies were selected based on their relevance to multimodal learning in healthcare, methodological contribution, and influence on the field. Since this work was designed as a narrative review rather than a systematic review, formal study counts and dual-reviewer screening procedures were not applied. Accordingly, this review should be interpreted as a structured, concept-driven synthesis intended to summarize methodological foundations, representative applications, and important challenges for clinical translation.

3. 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.

3.1. Representative Healthcare Data Modalities

This subsection highlights the modalities that most frequently appear in multimodal systems and clarifies why they are complementary rather than interchangeable. Table 1 provides a concise comparison of the representative modalities discussed below.

3.1.1. Medical Imaging

Medical imaging provides anatomical, functional, and morphological evidence that is often difficult to recover from any other modality. X-ray and computed tomography (CT) are frequently used for cardiopulmonary assessment, emergency triage, and oncologic workups, whereas magnetic resonance imaging (MRI) remains especially important for neurology, musculoskeletal analysis, and cardiovascular characterization [12–15]. Ultrasound offers real-time, radiation-free imaging at relatively low cost, and dermoscopy provides fine-grained skin-surface information that is highly relevant for lesion analysis [16–19]. In practice, imaging rarely acts alone in healthcare MML. Its value is often amplified when

 doi:10.20944/preprints202512.1445.v2

paired with clinical context, pathology, reports, or patient metadata, especially when preprocessing accounts for artifact removal, region-of-interest extraction, and protocol variability [20–22].

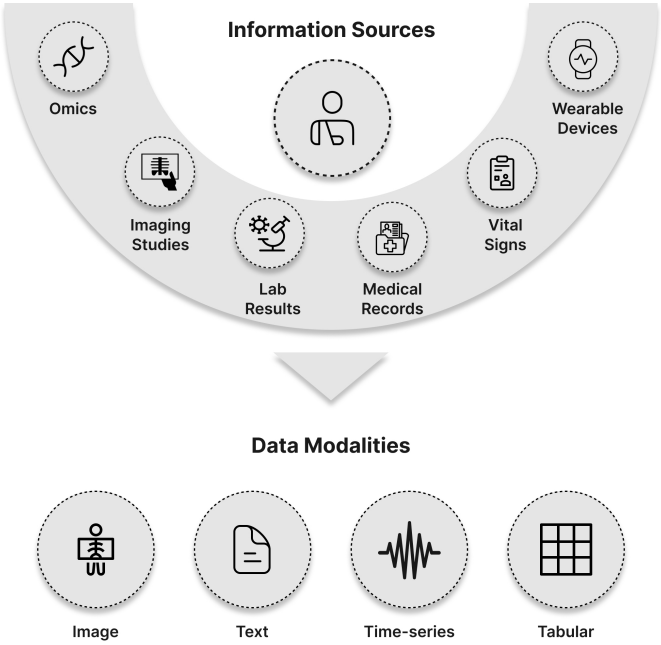

Figure 1. Overview of representative healthcare data modalities used in multimodal machine learning.

3.1.2. Text Data

Textual data are a major source of clinical context because they capture symptoms, clinician reasoning, diagnoses, procedures, and follow-up decisions that imaging or physiological signals often miss [23,24]. Common examples include discharge summaries, progress notes, referral letters, radiology reports, and medication records. Because these sources are largely unstructured, natural language processing (NLP) is needed to extract clinically meaningful information and align it with other modalities [25,26]. Methods such as transfer learning, active learning, and synthetic note generation are increasingly used to reduce annotation burden and improve data efficiency in healthcare NLP pipelines [27,28].

3.1.3. Time Series Data

Time series data arise from vital signs, bedside monitors, laboratory trajectories, medication histories, and wearable sensors. Their value lies in preserving temporal evolution, which is essential for deterioration prediction, event detection, and longitudinal disease modeling [29–31]. At the same time, these data are often irregular, noisy, and incomplete, making alignment with imaging or clinical documentation difficult [32,33]. Classical forecasting tools such as autoregressive integrated moving average (ARIMA) remain useful in some settings, but deep sequence models such as Long Short-Term Memory (LSTM) networks are often better suited to the nonlinear dynamics of healthcare monitoring data [34–37].

3.1.4. Tabular Data

Tabular data remain indispensable because they provide concise patient-level summaries of demographics, diagnoses, and standardized clinical scores [38–40]. In real-world healthcare models, these structured variables are often the easiest modality to integrate and therefore commonly serve as a practical partner to imaging or text [20,41]. In more complex settings, genomic and other omics measurements add molecular specificity that can refine prognosis, subtype discovery, and treatment-response prediction, especially in oncology [7,42]. However, these data are often sparse, inconsistently collected, and difficult to standardize across institutions.

 doi:10.20944/preprints202512.1445.v2

Table 1. Comparison of representative healthcare data modalities used in multimodal machine learning.

Modality	Typical Data	Main Contribution	Common Limitations	Fusion Suitability
Medical imaging	X-ray, CT, MRI, ultrasound, dermoscopy	Captures anatomical patterns	Protocol variability, cost, artifacts	High
Text data	Clinical notes, reports	Provides contextual information	Unstructured, noisy	High
Time-series data	Vital signs, sensors	Captures temporal patterns	Missing data, irregular sampling	High
Tabular data	Demographics, labs	Structured features	Limited expressiveness	Moderate

3.2. Challenges in Multimodal Machine Learning

MML in healthcare brings together data from different sources to support better patient care. At the same time, this integration introduces several challenges that must be addressed if models are to be both effective and dependable.

3.2.1. Heterogeneity of Modalities

Healthcare data take many forms, including structured laboratory results, unstructured clinical text, images such as X-rays and MRIs and continuous streams from wearable sensors. Each modality has its own characteristics and therefore requires preprocessing and analysis methods tailored to that type of data. For instance, NLP is needed to extract relevant information from clinical notes [1,43], whereas computer vision (CV) methods are needed to interpret medical images [44]. Likewise, advances in next-generation sequencing have made it possible to analyze complex genomic data for clinical use [45], while wearable sensors generate continuous streams that come with their own preprocessing challenges [46]. Differences in scale, units, and statistical properties across modalities make them difficult to combine into a coherent model. These differences also create challenges for preprocessing, feature extraction, and normalization, all of which are foundational to robust machine-learning models [47]. As a result, multimodal healthcare systems depend on strong fusion strategies that can handle the diversity and complexity of these datasets [48].

3.2.2. Alignment

Another important challenge is alignment, particularly the temporal and spatial synchronization of data from different modalities. Data collected from different sources are often not aligned in time or space [49]. Wearable devices can produce continuous physiological streams, whereas imaging studies and laboratory tests are usually collected intermittently [50–52]. Clinical notes may be recorded hours or even days after patient encounters, creating additional temporal discrepancies [53]. Spatial misalignment can also arise when imaging data capture different anatomical views or resolutions, or when anatomical variation between patients introduces added complexity [54]. These forms of misalignment make it harder to correlate events across modalities and often complicate model design, especially when architectures must also handle missing or asynchronous data points [55]. Addressing alignment is therefore essential for effective multimodal integration and for the development of reliable machine-learning models.

3.2.3. Fusion Strategies

Effective fusion strategies are indispensable for combining information from multiple modalities, but they are also one of the main sources of modeling difficulty. Early fusion may create high-dimensional feature spaces that increase computational cost and overfitting risk, especially in small datasets [56]. Intermediate or joint fusion can preserve modality-specific encoders, yet it introduces difficult design choices about where fusion should occur and how to prevent one modality from dominating the others [57]. Late fusion is operationally simpler and often more robust to heterogeneous

 doi:10.20944/preprints202512.1445.v2

pipelines, but it may miss clinically meaningful cross-modal interactions. Hybrid fusion can partially reconcile these strengths, although it typically requires more computation, more careful tuning, and more training data [58,59]. In practice, strong regularization, calibration, and cross-validation are usually needed, and the cost of training large multimodal systems remains non-trivial [60].

A closely related challenge is that real-world healthcare datasets are rarely complete or standardized across institutions. A patient may have MRI but not PET, imaging and notes may not be collected at the same visit, and laboratory units, coding systems, scanner protocols, or documentation templates may vary substantially from one hospital to another. Without stronger interoperability and standardization across institutions, a multimodal model may learn site-specific shortcuts rather than clinically meaningful relationships [8,10,48]. These issues directly affect missing-modality handling, external validation, fairness, and the ability to deploy models across institutions.

3.3. Techniques for Multimodal Machine Learning

Fusion strategies are central to MML because they determine how complementary clinical signals are combined. In this review, we distinguish between two complementary views of fusion: *stage-based fusion strategies*, which describe *when* modalities are combined in the pipeline, and *feature integration mechanisms*, which describe *how* modality representations are mathematically merged. This distinction improves terminological consistency because the same feature integration mechanism, such as concatenation or attention, may appear inside different stage-based designs.

3.3.1. Stage-Based Fusion Strategies

Stage-based fusion refers to the point in the learning pipeline at which information from different modalities is combined. The main approaches are early fusion, intermediate (joint) fusion, late fusion, and hybrid (mixed) fusion. Table 2 summarizes their principal trade-offs.

a) Early Fusion:

Early fusion [61–63] combines raw inputs or modality-specific features near the beginning of the learning pipeline. In practice, it is often implemented through concatenation or related feature-integration operations, allowing the model to learn joint patterns and cross-modal correlations from the outset.

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

 doi:10.20944/preprints202512.1445.v2

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

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.

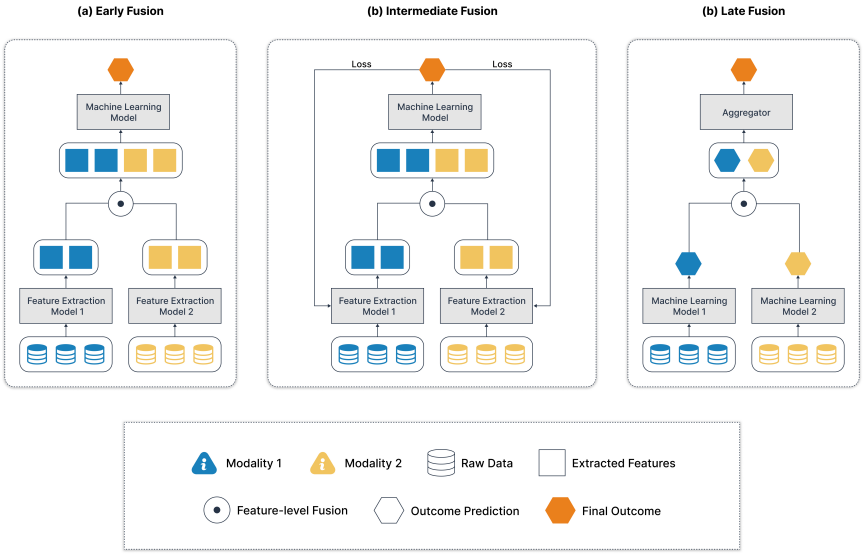

Figure 2. Comparison of stage-based fusion strategies in multimodal learning: (**a**) early fusion integrates raw or extracted features before the final predictor; (**b**) intermediate fusion combines modality-specific representations within a unified model; and (**c**) late fusion aggregates outputs from separately processed modalities.

The main benefit of early fusion is that it allows the model to learn relationships and dependencies across modalities from the very beginning. By processing the combined features simultaneously, the model can capture complex patterns that may not be evident when the modalities are analyzed separately [57]. This unified representation can also simplify the architecture and reduce the need for complicated synchronization mechanisms. Performance may improve further when the modalities provide genuinely complementary information. However, early fusion also has several drawbacks. Concatenating feature vectors from many modalities can create a very high-dimensional feature space, increasing computational cost and the risk of overfitting when training data are limited [64]. It also assumes that the modalities are aligned, an assumption that can be difficult to satisfy in real-world applications. In addition, if one modality has many more features or much stronger predictive power, it may dominate learning and obscure useful information from the others [65]. As the number of modalities and feature dimensions grows, the model can also become harder to scale efficiently.

b) Intermediate Fusion or Joint Fusion:

Intermediate fusion [63,66], also known as joint fusion [61,62], fuses multiple modalities at one or more intermediate layers of a model rather than only at the input or output stage (see Figure 2(b)). This contrasts with early fusion, where modalities are combined at the input level and may lose modality-specific nuances before higher-level feature extraction. In intermediate fusion, the loss function is back-propagated through each modality-specific feature extractor so that those representations can be fine-tuned during training [61].

 doi:10.20944/preprints202512.1445.v2

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. Parameterized by $\theta^{(m)}$, this model 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.

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 may be a simple operation such as concatenation or addition, or a more complex mechanism such as 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 offers both strengths and trade-offs. A key advantage is balanced processing: it allows deep, modality-specific feature extraction before fusion, helping preserve the unique characteristics of each modality. It can also model complex cross-modal relationships because fusion at intermediate layers enables the model to learn interactions that may be missed by purely early or late fusion approaches. However, intermediate fusion also increases model complexity, since separate networks for each modality and additional fusion layers raise computational requirements and the risk of overfitting. It also introduces architectural design challenges, particularly in choosing the most effective fusion method and determining where in the network the modalities should be combined.

Applications of intermediate fusion span a variety of fields. In multimodal sentiment analysis, Zadeh et al. [67] introduced the Memory Fusion Network (MFN), which processes language, visual, and acoustic modalities through separate LSTM networks and fuses their outputs at intermediate layers using a multi-view gated memory mechanism. In medical image analysis, Suk et al. [59] introduced a deep learning framework for Alzheimer's disease diagnosis that processes MRI and PET images through 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. [68] introduced a neural machine translation model that processes textual and visual information separately

 doi:10.20944/preprints202512.1445.v2

and fuses them at intermediate layers to improve translation quality. Together, these examples illustrate the flexibility and effectiveness of intermediate fusion for difficult multimodal problems.

c) Late Fusion:

In MML, late fusion [51,61,63], also known as decision-level fusion [69–71], processes each modality independently using separate models or pipelines and combines their outputs at the end of the learning process to make a final prediction [56,72]. This approach is simple and modular because each modality can be handled with the most appropriate method without requiring compatibility at the feature level. It is especially useful when the modalities are highly heterogeneous or when combining raw data or features directly is infeasible. In late fusion, each modality m in the set of modalities $\mathcal{M}$ is processed by its own machine learning model $f^{(m)}$, parameterized by $\theta^{(m)}$. The input data $\mathbf{x}^{(m)}$ for each modality 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 $\mathbf{y}^{(m)}$ may represent class probabilities, regression estimates, or other task-relevant predictions. They 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 in various ways, including averaging, majority voting, weighted summation, or training a meta-classifier on the outputs of the individual models. For instance, in weighted summation, weights $w^{(m)}$ are assigned to each modality's output based on its 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 the outputs are combined, the final prediction is derived from the aggregated result (see Figure 2(c)). In 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. [64] 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 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 [10,56,66], 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

 doi:10.20944/preprints202512.1445.v2

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 [56,66,73].

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 [74–77].

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 [78].

An example of hybrid fusion in practice is the Tensor Fusion Network (TFN) introduced by Zadeh et al. [73] 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 [73].

In healthcare practice, no single stage-based strategy is universally optimal. Early fusion can work well when measurements are synchronized and paired datasets are sufficiently large; intermediate fusion is often preferable when each modality needs its own encoder; late fusion is attractive when one expects missing modalities or separate institutional pipelines; and hybrid fusion is most useful when clinically relevant interactions exist at multiple levels. Importantly, multimodal systems may fail to outperform strong unimodal baselines when the added modality is noisy, weakly aligned, available only for a small subset of patients, or operationally unavailable at inference time.

Table 2. Comparison of stage-based fusion strategies in multimodal learning.

Strategy	Fusion Point	Main Strengths	Main Limitations
Early fusion	Input or feature level before the main predictor	Learns cross-modal interactions from the beginning and maintains a simple pipeline	High-dimensional inputs, strong alignment requirements, and risk of modality dominance
Intermediate fusion	One or more hidden layers after modality-specific encoders	Preserves modality-specific representations while modeling rich cross-modal dependencies	Requires careful design, more tuning, and higher implementation complexity
Late fusion	Decision level after separate unimodal models	Modular, flexible, and robust when modalities are heterogeneous or partially missing	May miss important cross-modal interactions due to late integration
Hybrid fusion	Multiple stages across the pipeline	Balances unimodal specialization with cross-modal interaction and supports complex multimodal settings	More computationally expensive and harder to design and optimize

 doi:10.20944/preprints202512.1445.v2

3.3.2. Feature Integration Mechanisms

Feature integration mechanisms describe how modality representations are mathematically merged once a model reaches a fusion point. The most common mechanisms are concatenation, operation-based fusion, and learning-based fusion. These mechanisms can appear inside early, intermediate, or hybrid multimodal models. Figure 3 provides a schematic overview, while broader taxonomies are discussed elsewhere [56,79–81]. Table 3 highlights the main differences among these mechanisms.

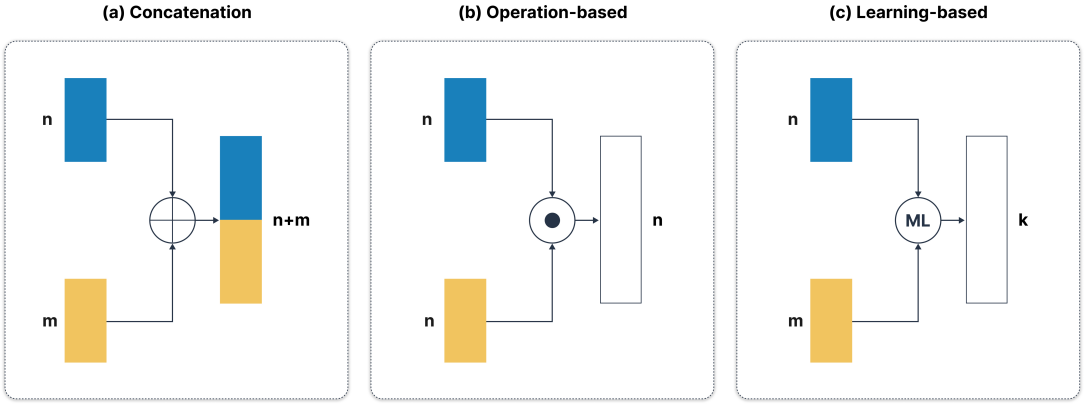

Figure 3. Feature integration mechanisms: (**a**) end-to-end concatenation of feature vectors; (**b**) element-wise operations or attention mechanisms for aligned feature representations; and (**c**) learning-based methods for constructing shared multimodal representations.

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)) [79]. 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 [82] 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 [83]. 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

 doi:10.20944/preprints202512.1445.v2

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 [20,34,84]. In order to reduce the problems of high dimensionality, dimensionality reduction techniques like Principal Component Analysis (PCA) [85] or Linear Discriminant Analysis (LDA) [86] can be applied before or after concatenation to reduce the feature space. Feature selection [82] 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 [10] (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 [10] and applies element-wise or channel-wise operations 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:

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

 doi:10.20944/preprints202512.1445.v2

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 [87, 88]. 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 [7,79,89]. Examples include [7] 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 [79,90–92]. Common applications of these attention mechanisms include tasks like image-caption generation [93] and audio-visual emotion recognition [94].

c) Learning-Based Fusion:

Learning-based fusion is a more advanced approach to the integration of multi-modal data, where machine learning models [82] 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.

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 [95]. 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) [96,97], 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 [97], 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

 doi:10.20944/preprints202512.1445.v2

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 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.

Table 3. Comparison of common feature integration mechanisms in multimodal learning.

Mechanism	Core Idea	Main Strengths	Main Limitations
Concatenation	Stacks modality feature vectors into a single representation	Simple, transparent, and easy to use as a baseline while retaining all modality features	Can create high-dimensional inputs, include redundant features, and require normalization or feature selection
Operation-based fusion	Combines aligned features using addition, multiplication, averaging, or similar element-wise operations	Produces compact fused representations and can efficiently highlight shared or complementary signals	Requires matched feature dimensions and may fail to capture complex nonlinear interactions
Learning-based fusion	Learns the fusion strategy using models such as autoencoders, CCA variants, or attention-based networks	More adaptive and expressive, with strong ability to capture nonlinear and task-specific relationships	Higher data and computational demands, reduced interpretability, and risk of overfitting without regularization

4. 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

 doi:10.20944/preprints202512.1445.v2

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.

4.1. 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 [98], 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 [98] 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" [99] 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 [100] 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 [77] and ViLBERT [76] 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) [101] 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) [102] 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 [103] 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 [104], 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 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. [105] 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 [106].

 doi:10.20944/preprints202512.1445.v2

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.

4.2. 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 [107] and Kernel CCA [108] 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 [97]. 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) [109] and ALIGN [110] 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 [111]. 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. An additional paradigm shift has been the emergence of multimodal large language models (MLLMs), which use language as an interface for cross-modal reasoning over images, reports, structured records, and other clinical evidence [11]. In healthcare, these systems are attractive because they can support retrieval, summarization, report drafting, and clinician-facing question answering in a single framework. At the same time, they raise concerns about hallucination, provenance tracking, calibration, and the risk of generating plausible but clinically unsafe explanations. As a result, MLLMs should be viewed as promising reasoning layers rather than as drop-in replacements for carefully validated task-specific models. As multimodal embeddings are increasingly applied to a lot of applications, ensuring their robustness, interpretability, and fairness has been a major goal of research [56,112]. 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

 doi:10.20944/preprints202512.1445.v2

features across modalities [113,114]. There is a growing need for further research to make multimodal embeddings understandable to both domain experts and end-users.

4.3. 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) [115] and Generative Adversarial Networks (GANs) [116], 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 [117] and Stable Diffusion [118] 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. [57] 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. [119]. 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, mitigating, and explaining biases to ensure that generated outputs do not inadvertently perpetuate societal stereotypes or mislead end-users [120].

 doi:10.20944/preprints202512.1445.v2

4.4. 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) [121] 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) [122], Graph Attention Networks (GATs) [123], Graph Isomorphism Networks (GINs) [124], and GraphSAGE [125] 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 [126]. 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 [103,127]. 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 [128]. 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 multimodal 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.

5. Applications of Multimodal Machine Learning in Healthcare

The integration of multimodal 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 multimodal deep learning in healthcare. Table 4 summarizes the major application domains reviewed in this section.

5.1. 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. [129] integrated T1-weighted structural MRI and phenotypic/demographic features to construct population graphs for classifying both autism spectrum disorder and Alzheimer's disease (AD). Similarly, Huang and Chung [130] introduced an edge-

 doi:10.20944/preprints202512.1445.v2

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 [131] 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. [81] exploited MRI, PET, and clinical/neuropsychological measures in a deep learning network for AD diagnosis, while Zhou et al. [132] 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. [133] 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. [134] 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. [59] 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. [135], 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. [92] 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. [136] 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. [137] 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 [31] used multimodal temporal signals—including behavioral and physiological data—to recognize bipolar disorder and depression states. Qiu et al. [138] 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. [139] 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. [140], 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. [35] 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.

5.2. Multimodality Approaches in Cancer Prediction

Nie et al. [141] used multimodal 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. [89] fused radiology, pathology, genomic, and clinical information to discover multimodal prognostic biomarkers for improving cancer outcome predictions. Focusing on breast cancer, Duanmu et al. [142] leveraged integrative imaging (MRI), molecular profiles, and demographic data in a deep learning framework to estimate pathological complete response to neoadjuvant chemotherapy. Yala et al. [143] constructed a mammography-based deep model for enhanced breast cancer risk assessment, while Yan et al. [84] proposed a richer fusion network that

 doi:10.20944/preprints202512.1445.v2

combines multiple imaging and non-imaging features to improve breast cancer classification. Liu and Hu [144] employed denoising autoencoders on genomic data to extract deep genomic features associated with breast cancer subtypes, and Li et al. [145] fused pathological images with genomic data to predict breast cancer survival outcomes. Similarly, Holste et al. [83] demonstrated an end-to-end approach for fusing breast MRI with tabular clinical descriptors to boost classification accuracy. Kharazmi et al. [146] targeted basal cell carcinoma detection by combining dermoscopic images and patient profiles in a feature-fusion system, whereas Hyun et al. [147] applied PET-based radiomics to distinguish histological subtypes in lung cancer. Vanguri et al. [42] 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 [90] developed a pan-cancer prognosis model using multimodal deep learning, an approach echoed by Cheerla and Gevaert [148], who combined multi-omic data for pan-cancer survival prediction. Beyond these examples, Rubinstein et al. [149] introduced an unsupervised technique for tumor detection in dynamic PET/CT scans of the prostate, while Reda et al. [150] highlighted how deep learning can aid in the early diagnosis of prostate cancer. Schulz et al. [91] similarly deployed a multimodal deep learning pipeline to forecast prognosis in renal cancer patients. Guo et al. [151] 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. [7] 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.

5.3. Multimodality Approaches in Chest Related Conditions

Palepu and Beam [152] 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. [20] 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. [153] 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. [154] 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. [155] 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. [156] 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 al. [157] combined clinical, radiological, and procedural data to forecast functional outcomes following thrombectomy, showing how integrated models can inform more personalized stroke management

 doi:10.20944/preprints202512.1445.v2

strategies. Walker et al. [34] 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. [158] 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. [159] 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. [41] 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. [160] 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. [161] 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.

5.4. Multimodality Approaches in Skin Related Conditions and Other Diseases

Taleb et al. [162] 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. [163] 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. [164] 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. [165] 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. [37] 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. [166] 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. [167] 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. [18] utilized a multimodal deep learning framework that draws on dermoscopic images and metadata to enhance skin lesion classification, while Gessert et al. [19] 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. [21] 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.

Across these application domains, one recurring lesson is that multimodal learning is most valuable when each modality contributes non-redundant clinical information and when the paired data can be aligned reliably. Imaging plus structured clinical variables is often easier to deploy than imaging plus omics because tabular EHR data are collected more routinely, whereas multi-omics cohorts remain relatively small. Likewise, multimodal gains can shrink or disappear when an added modality is noisy, weakly synchronized, unavailable at inference time, or already implicitly encoded

 doi:10.20944/preprints202512.1445.v2

in another source. These recurring patterns motivate the more critical synthesis in the following discussion section.

Table 4. Overview of major healthcare application domains for multimodal machine learning, including representative fusion strategies in the reviewed studies.

Domain	Common Modalities	Reported Fusion Strategy	Representative Tasks	Main Clinical Value
Brain disorders	MRI, PET, EEG, cognitive scores, demographics, genetics	Mostly joint/intermediate fusion; also hybrid, graph-based, and stage-wise fusion	Classification; staging; progression forecasting; outcome prediction	Combines structural, functional, cognitive, and molecular evidence for improved neurological characterization
Cancer prediction	Radiology, pathology, genomics, clinical variables, demographics	Mostly early/feature-level and joint/intermediate fusion; some hybrid approaches	Diagnosis; subtype classification; prognosis; treatment-response prediction; survival modeling	Improves patient stratification and biomarker discovery by linking imaging with molecular context
Chest-related conditions	Chest X-ray or CT, laboratory tests, vital signs, EHR data, ECG, clinical text	Late fusion in several COVID-19 studies; otherwise early and joint fusion	Cardiomegaly detection; risk prediction; pneumonia/COVID-19 severity assessment; outcome prediction	Enhances cardiopulmonary decision support by combining imaging with physiological context
Skin-related conditions and other diseases	Dermoscopic and clinical images, metadata, EHR data, blood parameters, audio–video–physiological signals	Mostly early and joint fusion; some hybrid multitask fusion	Lesion classification; disorder detection; mortality prediction; pain assessment; prognosis	Extends multimodal learning to diverse clinical settings with improvements in diagnosis and monitoring

6. 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. 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.

6.1. When Multimodality Helps—and When It Does Not

Multimodal learning is often beneficial in healthcare because different data sources describe different aspects of disease: imaging reflects anatomy, text captures clinical interpretation, physiological signals preserve temporal dynamics, and omics data reveal molecular variation. However, the gains are not universal. Multimodal systems are most helpful when the modalities are complementary, reliably aligned, and available at inference time. By contrast, multimodal models may offer only marginal improvement—or even underperform strong unimodal baselines—when additional modalities are noisy, weakly paired, sparsely available, or operationally difficult to collect in routine care. The key question is therefore not whether multimodality is always better, but under what clinical and infrastructural conditions it is worth the added complexity. Table 5 summarizes practical guidance for selecting among early, intermediate, late, and hybrid fusion strategies under different healthcare deployment conditions.

 doi:10.20944/preprints202512.1445.v2

Table 5. Practical guidance for choosing multimodal fusion strategies in healthcare.

Fusion strategy	Best used when	Use caution when	Healthcare deployment note
Early fusion	Modalities are well aligned at the patient or feature level and consistently available at inference time	Missing data are frequent, feature scales differ substantially, or dimensionality is high relative to sample size	Simplest end-to-end option, but typically requires careful preprocessing, normalization, and imputation design
Intermediate (joint) fusion	Each modality requires its own encoder and clinically meaningful cross-modal interactions are expected	Paired multimodal training data are limited or the architecture cannot be carefully tuned and validated	Often provides a strong balance between representation power and flexibility, but with higher validation effort
Late fusion	Modalities originate from separate pipelines or institutions, or one modality is often unavailable at runtime	Fine-grained cross-modal dependencies are critical and cannot be recovered from separate unimodal predictions	Easiest to maintain and integrate into existing workflows; supports graceful fallback when a modality is missing
Hybrid fusion	Important interactions occur at multiple levels and the task justifies a more complex modeling pipeline	Interpretability, computational budget, dataset size, or deployment simplicity are primary constraints	Best suited for high-value use cases where the added complexity is justified by clear clinical benefit

6.2. Clinical Translation and Deployment Considerations

Beyond benchmark performance, clinically useful MML systems must satisfy operational constraints that are often under-discussed in the modeling literature. First, multimodal data availability is uneven. A model trained on imaging, text, genomics, and dense laboratory panels may look attractive in a tertiary-care dataset but become unusable in community or low-resource settings where only a subset of modalities is routinely available [8,10]. Second, external validation remains essential because multimodal models can silently encode institution-specific workflows, scanner settings, note-writing habits, or demographic imbalances. Third, regulatory and workflow considerations matter: clinicians need traceable inputs, calibrated outputs, fallback behavior when a modality is missing, and evidence that the model improves decisions rather than merely improving retrospective AUC [11,48].

For these reasons, deployable multimodal systems should be designed with modality dropout, missing-data pathways, and institution-aware validation protocols in mind. They should also be accompanied by interoperable data standards, robust preprocessing pipelines, and post-deployment monitoring for calibration drift, bias amplification, and failure modes. In many practical settings, a slightly less accurate model that uses routinely available modalities and can be externally validated may be more valuable than a highly complex multimodal architecture that depends on data rarely available in clinical workflows.

6.3. 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:

 doi:10.20944/preprints202512.1445.v2

- **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.

6.4. 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.

6.5. Generative Models in Healthcare

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

- **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.

6.6. 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.

 doi:10.20944/preprints202512.1445.v2

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.

6.7. Interpretability, Fairness, and Ethical Considerations

As multi-modal models become more complex, concerns about interpretability and fairness become harder to ignore. 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.
- **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.

6.8. Path Forward

Looking ahead, several key themes stand out:

1. **Unified Foundation Models in Healthcare:** Inspired by CLIP, ALIGN, and multimodal 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. However, their usefulness will depend on whether they can operate under the practical constraints emphasized throughout this review: missing modalities, uneven modality availability, institution-specific variation, and the need for calibrated and traceable outputs. Accordingly, foundation-model research in healthcare should prioritize modality-dropout robustness, provenance tracking, uncertainty estimation, and evaluation under realistic deployment conditions rather than only benchmark performance.
2. **Causality and Counterfactual Reasoning:** Current MML approaches excel at correlational reasoning but often fail to capture causal relationships. This limitation is especially important in healthcare because multimodal datasets frequently contain confounding from site-specific workflows, documentation habits, treatment-selection effects, and demographic imbalance. Developing causal representation learning methods that disentangle these effects may improve

 doi:10.20944/preprints202512.1445.v2

generalization, make predictions more clinically interpretable, and support more reliable reasoning about interventions rather than only associations.

3. **Multimodal Reinforcement Learning (RL):** Interactive clinical tasks—such as robotic procedures, closed-loop monitoring, or adaptive therapy optimization—may benefit from combining RL with multimodal understanding. Yet this direction also inherits the challenges discussed earlier, including noisy streams, asynchronous inputs, safety constraints, and limited tolerance for errors. Progress will therefore require simulation-to-clinic transfer strategies, uncertainty-aware policies, human oversight, and explicit fallback mechanisms when one or more modalities are missing or unreliable.
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 centralizing sensitive data. This direction is particularly important for addressing one of the major bottlenecks identified in this review: the lack of external validation across heterogeneous institutions. Future work should therefore focus not only on privacy guarantees, but also on harmonization across sites, communication-efficient training, bias monitoring, and robust aggregation when participating institutions differ in modalities, sample sizes, and patient populations.
5. **Standardization and Interoperable Data Infrastructures:** Progress in multimodal healthcare will depend not only on better models but also on better data plumbing. Harmonized acquisition protocols, consistent coding and laboratory units, interoperable record standards, and institution-spanning quality control are basic but essential requirements for reliable cross-site fusion and external validation. In practical terms, this is the clearest path toward reducing alignment errors, preventing site-specific shortcut learning, improving missing-modality handling, and making multimodal systems portable across real clinical environments.

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 may expand what multimodal systems can do, but the field will move forward only if those advances are tied directly to the unresolved problems identified throughout this review: heterogeneity, alignment, missing modalities, institutional variation, interpretability, fairness, robustness, and workflow integration. By addressing these methodological and translational challenges together, researchers can ensure that multimodal systems are not only technically impressive but also clinically trustworthy, safe, and useful.

7. Conclusion

MML has become an important part of next-generation healthcare by bringing together diverse data types. From foundational fusion strategies—early, intermediate, late, and hybrid—to advanced techniques involving attention-based transformers, cross-modal embeddings, generative modeling, graph neural networks, and emerging MLLM-style reasoning layers, MML has demonstrated strong potential for capturing the complexity of patient health. Across neurology, oncology, cardiopulmonary medicine, dermatology, and monitoring applications, multimodal systems can improve diagnostic accuracy, disease staging, prognosis prediction, and clinical decision support when the contributing modalities are truly complementary and operationally available.

At the same time, this review highlights that MML is not inherently superior to unimodal modeling. The effectiveness of MML depends not only on model architecture, but also on how well data characteristics and clinical constraints are addressed. Challenges such as cross-modal alignment, data heterogeneity, missing modalities, and variability across institutions can limit performance and generalizability. These issues are further compounded by practical considerations in clinical translation, including interpretability, robustness, fairness, and external validation. Looking forward, progress in MML for healthcare will require a balanced focus on methodological innovation and real-world deployment. Advances in foundation models, cross-modal reasoning, privacy-preserving learning,

 doi:10.20944/preprints202512.1445.v2

and integrated healthcare data infrastructures may expand the capabilities of MML, but their impact will depend on rigorous validation and alignment with clinical workflows. Overall, the successful adoption of MML will depend on integrating diverse data sources in ways that are not only technically effective, but also reliable, transparent, and clinically meaningful.

Author Contributions: Conceptualization, M.A.R, P.S, M.S.H.A and R.M.; software, M.A.R; validation, M.A.R, P.S, M.S.H.A and R.M.; formal analysis, M.A.R, P.S, M.S.H.A and R.M.; investigation, M.A.R, P.S, M.S.H.A and R.M.; resources, M.A.R, P.S, M.S.H.A 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. 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.
5. de Gomez, M.R.C. A Comprehensive Introduction to Healthcare Data Analytics. *Journal of Biomedical and Sustainable Healthcare Applications* **2024**. n. pag.
6. Seneviratne, M.G.; Kahn, M.G.; Hernandez-Boussard, T. Merging heterogeneous clinical data to enable knowledge discovery. *Pac Symp Biocomput* **2019**, *24*, 439–443.
7. 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.
8. 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.
9. Bertsimas, D.; Ma, Y. M3H: Multimodal Multitask Machine Learning for Healthcare. *arXiv preprint arXiv:2404.18975* **2024**.
10. Krones, F.; Marikkar, U.; Parsons, G.; Szmul, A.; Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Information Fusion* **2025**, *114*, 102690.
11. 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*.
12. England, N.H.S.; Improvement, N.H.S. Diagnostic imaging dataset statistical release. *Department of Health* **2016**, *421*.
13. Goldman, L.W. Principles of CT and CT technology. *Journal of Nuclear Medicine Technology* **2007**, *35*, 115–128.
14. 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.
15. 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.
16. 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.
17. 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.

 doi:10.20944/preprints202512.1445.v2

18. Yap, J.; Yolland, W.; Tschandl, P. Multimodal skin lesion classification using deep learning. *Experimental Dermatology* **2018**, *27*, 1261–1267.
19. 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.
20. 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.
21. 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.
22. 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.
23. Spasic, I.; Nenadic, G.; et al. Clinical text data in machine learning: systematic review. *JMIR Medical Informatics* **2020**, *8*, e17984.
24. Johnson, A.; Pollard, T.; Horng, S.; Celi, L.A.; Mark, R. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet* **2023**.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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**.
31. Ceccarelli, F.; Mahmoud, M. Multimodal temporal machine learning for bipolar disorder and depression recognition. *Pattern Analysis and Applications* **2022**, *25*, 493–504.
32. 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.
33. 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.
34. 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**.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.

 doi:10.20944/preprints202512.1445.v2

41. 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**.
42. 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.
43. 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.
44. 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.
45. 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.
46. Piwek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The rise of consumer health wearables: promises and barriers. *PLoS medicine* **2016**, *13*, e1001953.
47. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *New England Journal of Medicine* **2019**, *380*, 1347–1358.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* **2017**.
54. Zitova, B.; Flusser, J. Image registration methods: a survey. *Image and vision computing* **2003**, *21*, 977–1000.
55. Lipton, Z.C.; Kale, D.C.; Elkan, C.; Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* **2015**.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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*.
61. 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.
62. 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.
63. 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.
64. 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.

 doi:10.20944/preprints202512.1445.v2

65. 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.
66. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* **2017**, *34*, 96–108.
67. 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.
68. Calixto, I.; Liu, Q.; Campbell, N. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287* **2017**.
69. 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.
70. 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.
71. Behrad, F.; Abadeh, M.S. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications* **2022**, *200*, 117006.
72. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Computation* **2020**, *32*, 829–864.
73. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* **2017**.
74. 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*.
75. 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*.
76. 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.
77. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* **2019**.
78. 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.
79. 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*.
80. 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.
81. 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**.
82. 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.
83. 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.
84. 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.
85. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2010**, *2*, 433–459.
86. 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.
87. 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.

 doi:10.20944/preprints202512.1445.v2

88. 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.
89. 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.
90. 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.
91. 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*.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
99. Xu, K. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044* **2015**.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. Hardoon, D.R.; Szedmak, S.; Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* **2004**, *16*, 2639–2664.

 doi:10.20944/preprints202512.1445.v2

108. Lai, P.L.; Fyfe, C. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems* **2000**, *10*, 365–377.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* **2013**.
116. 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.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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.
122. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* **2016**.
123. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv preprint arXiv:1710.10903* **2017**.
124. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv preprint arXiv:1810.00826* **2018**.
125. 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.
126. 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.
127. 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**.
128. 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.
129. 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.
130. 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.

 doi:10.20944/preprints202512.1445.v2

131. 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.
132. 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.
133. 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.
134. 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.
135. 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.
136. 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.
137. 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.
138. 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.
139. 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.
140. 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.
141. 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.
142. 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.
143. 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.
144. Liu, Q.; Hu, P. Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer. *Cancers* **2019**, *11*, 494.
145. 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.
146. 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.
147. 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.
148. Cheerla, A.; Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **2019**, *35*, i446–i454.
149. 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.
150. 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.

 doi:10.20944/preprints202512.1445.v2

151. 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.
152. Palepu, A.; Beam, A.L. Tier: Text-image entropy regularization for clip-style models. *arXiv preprint arXiv:2212.06710* **2022**.
153. 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.
154. 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.
155. 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.
156. 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.
157. 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.
158. 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.
159. 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.
160. 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.
161. 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.
162. 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.
163. 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.
164. 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.
165. 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**.
166. 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.
167. 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.