Journal of Biomedical Informatics 146 (2023) 104482

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A scoping review on multimodal deep learning in biomedical images and texts

Zhaoyi Sun [a], Mingquan Lin [a], Qingqing Zhu [b], Qianqian Xie [a], Fei Wang [a], Zhiyong Lu [b], Yifan Peng [a,*]

[a] Population Health Sciences, Weill Cornell Medicine, New York, NY 10016, USA
[b] National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA

ARTICLE INFO

Keywords:
Multimodal learning
Medical images
Clinical notes
Scoping review

ABSTRACT

Objective: Computer-assisted diagnostic and prognostic systems of the future should be capable of simultaneously processing multimodal data. Multimodal deep learning (MDL), which involves the integration of multiple sources of data, such as images and text, has the potential to revolutionize the analysis and interpretation of biomedical data. However, it only caught researchers' attention recently. To this end, there is a critical need to conduct a systematic review on this topic, identify the limitations of current work, and explore future directions.

Methods: In this scoping review, we aim to provide a comprehensive overview of the current state of the field and identify key concepts, types of studies, and research gaps with a focus on biomedical images and texts joint learning, mainly because these two were the most commonly available data types in MDL research.

Result: This study reviewed the current uses of multimodal deep learning on five tasks: (1) Report generation, (2) Visual question answering, (3) Cross-modal retrieval, (4) Computer-aided diagnosis, and (5) Semantic segmentation.

Conclusion: Our results highlight the diverse applications and potential of MDL and suggest directions for future research in the field. We hope our review will facilitate the collaboration of natural language processing (NLP) and medical imaging communities and support the next generation of decision-making and computer-assisted diagnostic system development.

1. Introduction

Multimodal deep learning (MDL), which involves the integration of multiple modalities, such as medical images, unstructured text, and structured Electronic Health Records (EHRs) has gained significant attention in biomedical research [1]. This approach has been proven to improve the accuracy and efficiency of various tasks in clinical decision-making with imaging and structured EHR (i.e., -omics data, lab test data, demographic data) [2–4]. The heterogeneous data available to clinicians allows for multiple viewpoints to be considered when making decisions and constructing computer-aided diagnosis and prognosis systems. However, the application of MDL with medical imaging data and unstructured free-text data (i.e., clinical reports) is still in its infancy. The emergence of related research has only recently surfaced. For example, in the field of natural language processing (NLP), pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) [5] and Generative Pre-trained Transformer 3 (GPT- 3) [6], have garnered world-renowned accomplishments in various downstream tasks.

Furthermore, multimodal language models, including Contrastive Language Image Pretraining (CLIP) [7] and the more recent KOSMOS-1 [8], have demonstrated remarkable performances in addressing general domain tasks. This notable progress has simultaneously facilitated the models' applicability within the medical domain. As a result, we believe it is imperative to comprehensively synthesize the past five years' research on MDL in biomedical images and texts, including an overview of research objectives and methodologies, elucidating development trends, and exploring potential broader clinical applications in the future.

Our review is inspired by several related review articles. Heiliger et al. [9] provided a comprehensive overview of existing multimodal learning methods and related databases in radiology, proposing a

* Corresponding author at: 425 East 61st Street, Suite 301, New York, NY 10065, USA.
E-mail addresses: zhs4003@med.cornell.edu (Z. Sun), mil4012@med.cornell.edu (M. Lin), qingqing.zhu@nih.gov (Q. Zhu), qix4002@med.cornell.edu (Q. Xie), few2001@med.cornell.edu (F. Wang), luzh@ncbi.nlm.nih.gov.Yifan (Z. Lu), yip4002@med.cornell.edu (Y. Peng).

https://doi.org/10.1016/j.jbi.2023.104482
Received 31 March 2023; Received in revised form 18 July 2023; Accepted 28 August 2023
Available online 29 August 2023

modality-based taxonomy based on the structural and design principles of the model. However, it was method-oriented, which might not facilitate clinicians' comprehension of the development of MDL in the medical field from the standpoint of specific applications. Cui et al. [10] explored the various fusion strategies employed in disease diagnosis and prognosis. However, the multimodal fusion discussed in these articles primarily included structured data from EHRs, with limited attention to unstructured text. Similarly, numerous systematic reviews have synthesized the employment of multimodal artificial intelligence (AI), machine learning, and the Internet of Medical Things (IoMT) within the realm of biomedicine [11–13]. Nonetheless, these investigations exhibited a notable absence of detailed discussions on implementing multimodal language models in the medical domain.

Additionally, the outstanding achievements of deep learning are accompanied by increasing model complexity and a lack of interpretability of AI models that prevents their applicability to clinical scenarios [14]. Therefore, it becomes necessary to come up with solutions to address this challenge and move toward more transparent AI. Compared to single-model AI, MDL presents unique challenges as explanations of multimodal data are often separated. For example, there are SHAP values for the EHR and a heatmap for the brain images - a visualization of the brain areas affected. But few visualization/explanation methods integrate the data and results, especially with longitudinal data. While many review studies organize and report challenges and opportunities of explainable AI, however, they do not focus on MDL [15–17].

To our knowledge, our paper represents the first review of multimodal deep learning focusing on medical image and text data, explainability, and human evaluation. Our motivation is to foster the application of multimodal language models in the medical field in a more comprehensible manner. Our target readers include clinicians and computer scientists. Specifically, we aim to provide clinicians with insights into the current performance of various pre-training models on different clinical tasks, as well as opportunities to evaluate model interpretability and contribute to developing new public datasets. Meanwhile, we hope that computer scientists will advance the clinical translation of models by focusing on clinical tasks, recognizing the significance of external validation, and increasing model transparency in the clinical translation process.

The review questions and objectives for this scoping review are as follows: The primary research question is: What is the current state of the literature on MDL in biomedical images and texts?

This question will be addressed by exploring the following subquestions: What databases were utilized in these studies? What were multimodal fusion techniques employed in these studies? Which image and text modalities were incorporated in these studies? What metrics were utilized to evaluate the model's performance in these studies? Did these studies employ external validation? Did these studies explicate the model's interpretability?

The organization of the review is as follows: Section 2 describes the protocol used in planning and executing this systematic review. Section 3 discusses the research directions of five tasks: report generation, visual question answering, cross-modal retrieval, diagnostic classification, and semantic segmentation. Section 4 summarizes the limitations and challenges of the current approaches and highlights future research directions. Lastly, Section 5 concludes the final remarks.

2. Methods

Our scoping review follows the Preferred Reporting Items for Systematic Reviews and Meta- Analyses (PRISMA) guidelines [18].

2.1. Eligibility criteria

Our scoping review focused on research on multimodal deep learning techniques applied to medical images and unstructured text. The inclusion criteria for our review consisted of English- language articles published between 2018 and 2022, including both conference papers and journal articles. We chose this time frame to capture the most up-to-date research in this rapidly evolving field. Additionally, we refer to relevant preprint articles to ensure we can consider cutting-edge research that has yet to be published in peer-reviewed venues.

2.2. Information sources

A search of multiple databases was carried out, including PubMed (https://pubmed.ncbi.nlm.nih.gov/), the Association for Computing Machinery (ACM) Digital Library (https://dl.acm.org/), the Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library (https://ieeexplore.ieee.org/Xplore/home.jsp), Google Scholar (https://scholar.google.com/), and Semantic Scholar (https://www.semanticscholar.org/). The most recent search was executed on January 8, 2023.

2.3. Search strategy

All the studies collected in this research were confined to the medical field. Initially, our search comprised three keyword groups: image modality (e.g., medical images and radiology images), text modality (e.g., text and report), and multimodal fusion learning (e.g., multimodal learning, joint fusion, and contrastive learning). We combined these keywords to carry out the first round of collection across five databases. To ensure the comprehensiveness of the articles collected, we conducted a second round of collection on Google Scholar, by adding a fourth application- oriented keyword group (i.e., report generation, visual question answering, and cross-modal retrieval).

2.4. Study selection

Title and abstract screening were conducted independently by two reviewers (ZS and ML). In cases of disagreement, studies were subjected to full-text review, and a consensus was reached through discussions. Subsequently, each article was reviewed and labeled according to the tasks. These tasks encompassed report generation, visual question answering, cross-modal retrieval, diagnostic classification, semantic segmentation, and other related tasks, with the possibility for a single article to correspond to multiple tasks. During the screening and the full-text review stages, we excluded review articles, non-medical articles, poor-quality articles, and unimodal studies (i.e., studies focusing solely on images or text). Articles containing modalities without images or text (e.g., omics data, lab test data, and demographic data) were also excluded.

2.5. Data extraction and synthesis

In our study, we undertook a systematic analysis of each downstream task. Firstly, we explored commonly used datasets for the task at hand, as well as their primary contents. Secondly, we expounded on the commonly employed multimodal frameworks and development trends of the methodology (e.g., fusion embedding, transformer-based attention models, and contrastive language-image pre-training). Subsequently, we summarized the specific image and text modalities covered in the articles, such as chest X-rays (CXR) and radiology reports. Lastly, we sorted out commonly used evaluation metrics for each downstream task, such as the area under the receiver operating characteristic curve (AUC), F1-score, and bilingual evaluation understudy (BLEU) [19]. Of particular note, we considered whether clinical experts were invited for external validation and explanation of the model's interpretability. We believe this has significant implications for enhancing the accuracy of computer-aided diagnosis and prognosis in the future.

3. Results

3.1. Included studies and datasets

A total of 361 articles were retrieved from five databases, from which 77 articles were ultimately included in our review. Fig. 1 shows the flowchart of our article screening process. During the screening process, we excluded 137 articles based on their titles and abstracts, according to our predetermined exclusion criteria (Section 2.4). Subsequently, a full-text review was conducted on the remaining articles, which resulted in an additional 13 articles being excluded. Specifically, these articles were discarded based on evaluations of their full texts, including 3 non-medical articles, 6 articles that lacked a text modality, 1 article that lacked an imaging modality, 2 articles on unimodal learning, and 1 poor-quality article.

Table 1 encapsulates the medical multimodal datasets employed in the articles collected in this scoping review, encompassing the dataset name, image type, text type, and the corresponding website for each dataset.

3.2. Report generation

Report generation aims at generating descriptives from EHR and medical images automatically. It could ease the work burden upon clinicians and improve the quality of the reports themselves. Since the training process of report generation typically requires both medical images and text reports written by clinicians, it can be naturally considered a multimodal learning process.

Table 2 provides an overview of the application of multimodal deep learning on report generation. Common image data used in the medical field include X-rays, computerized tomography (CT), magnetic resonance imaging (MRI), and pathological images. A common dataset for this task is the IU X-Ray [20] dataset, which comprises 7,470 frontal and lateral chest radiographs and 3,955 corresponding reports. Another widely-used dataset is the MIMIC-CXR [21,22] dataset, including 377,110 images and 227,827 reports. Furthermore, there exist datasets specifically designed for image classification and assistance in report generation, such as the CheXpert dataset [23], which comprises 224,316 images and 14 labels marked as present, absent, or uncertain.

Most studies employ convolutional neural networks (CNNs) to process medical images. Regarding text processing, Long Short-Term Memory (LSTM) was previously a popular method. For example, Yuan et al. [24] developed a CNN encoder and hierarchical LSTM decoder that utilized a visual attention mechanism based on multi-view in radiology. In the recent two years, the Transformer architecture has seen increasing use in report generation. Chen et al. [25] proposed the VMEKNet model, which combines the Transformer architecture with visual memory and external knowledge, resulting in improved performance in both qualitative and quantitative experiments and clinical diagnosis. Another notable contribution is the AlignTransformer proposed by You et al. [26], which effectively addresses data bias and is particularly well-suited for long-sequence report generation. The use of self-supervised learning techniques, such as CLIP, has also garnered attention for its ability to retrieve reports for report generation purposes. The CXR-RePaiR model proposed by Endo et al. [27] employed the CLIP approach with retrieval-based mechanisms and achieved outstanding metrics in language generation tasks. Similarly, the RepsNet model proposed by Tanwani et al. [28] incorporates the principle of self-supervised contrastive alignment. Recent research has focused on improving the factual correctness and completeness of generated reports through reward mechanisms. Miura et al. [29] developed a model that applies a reward mechanism to reinforcement learning, resulting in significant improvements in clinical performance. This approach was further refined by Delbrouck et al. [30] and improved by 14.2% in factual correctness and 25.3% in completeness.

Evaluation metrics for report generation can be classified into three

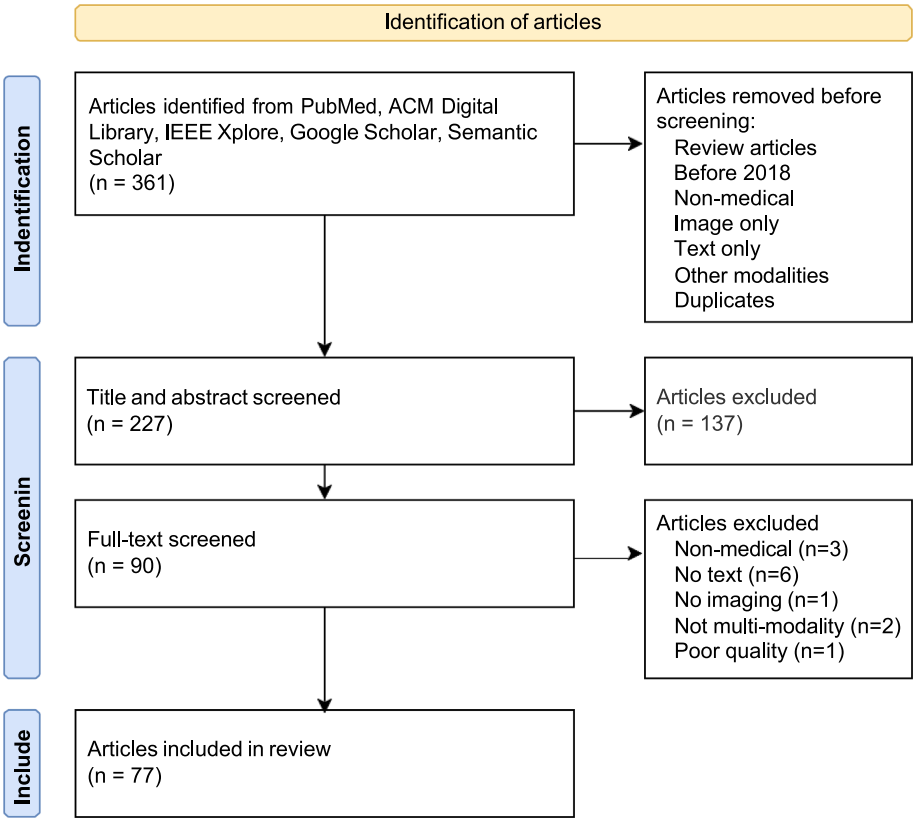

Fig. 1. Flowchart of article selection.

Table 1
Multimodal medical image-text datasets.

Dataset	Image type	Text type	URL
MURA	Bone X-rays	Annotations	https://stanfordmlgroup.github.io/competitions/mura/
DeepLesion	CT	Annotations	https://nihcc.app.box.com/v/DeepLesion
COV-CTR	CT	Radiology reports	https://github.com/mlii0117/COV-CTR
COVID-19 CT	CT	Radiology reports	https://covid19ct.github.io
COVID Rural	CT, CXR	Annotations	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226443
COVID-19 Image Data Collection	CT, CXR	Annotations	https://github.com/ieee8023/covid-chestxray-dataset
COVIDx	CXR	Annotations	https://github.com/lindawangg/COVID-Net
MS-CXR	CXR	Annotations	https://aka.ms/ms-cxr
QaTa-COV19	CXR	Annotations	https://www.kaggle.com/datasets/aysendegerli/qatacov19-dataset
Shenzhen Tuberculosis	CXR	Annotations	https://www.kaggle.com/datasets/raddar/tuberculosis-chest-xrays-shenzhen
SIIM-ACR	CXR	Annotations	https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/data
VinBigData Chest X-ray	CXR	Annotations	https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection/data
RSNA	CXR	Image captions	https://rsna.org/challenge-datasets/2018
CheXpert	CXR	Radiology reports	https://stanfordmlgroup.github.io/competitions/chexpert
IU X-Ray	CXR	Radiology reports	https://openi.nlm.nih.gov
MIMIC-CXR	CXR	Radiology reports	https://physionet.org/content/mimic-cxr/2.0.0
MIMIC-CXR-JPG	CXR	Radiology reports	https://physionet.org/content/mimic-cxr-jpg/2.0.0
NIH-CXR	CXR	Radiology reports	https://nihcc.app.box.com/v/ChestXray-NIHCC
PadChest	CXR	Radiology reports	https://bimcv.cipf.es/bimcv-projects/padchest
RadGraph	CXR	Radiology reports	https://physionet.org/content/radgraph/1.0.0
MoNuSeg	Pathology images	Annotations	https://monuseg.grand-challenge.org/Data
ARCH	Pathology images	Image captions	https://warwick.ac.uk/fac/cross_fac/tia/data/arch
PathVQA	Pathology images	Medical questions	https://github.com/UCSD-AI4H/PathVQA
TCGA	Pathology images	Pathology reports	https://portal.gdc.cancer.gov/repository
PEIR	Pathology images, radiology images	Image captions	https://peir.path.uab.edu/library
MedICaT	Radiology images	Image captions	https://github.com/allenai/medicat
ROCO	Radiology images	Image captions	https://github.com/razorx89/roco-dataset
ImageCLEF VQA-Med	Radiology images	Medical questions	https://www.imageclef.org
SLAKE	Radiology images	Medical questions	https://www.med-vqa.com/slake
VQA-RAD	Radiology images	Medical questions	https://osf.io/89kps

categories: text quality, medical correctness, and explainability [49]. These metrics are typically intended to be generated automatically, rather than manually, to facilitate automation of the report generation process. The text quality is commonly evaluated using metrics such as BLEU [19], METEOR [50], and ROUGE-L [51]. Medical correctness is evaluated using metrics such as AUC, precision, recall, and F1 [41,46]. Yu et al. introduced a composite metric, RadCliQ, aimed at quantifying the similarity between model-generated reports and those produced by radiologists, and the percentage of decreased errors [52]. Additionally, the explainability-related metrics factENT and factENTNLI, proposed by Miura et al. [29], have been shown to effectively evaluate the factual correctness and completeness of the model. In the reviewed literature, 10 articles sought external validation through the involvement of radiologists or other clinical experts. Furthermore, 14 articles provided validation of the interpretability of the models through various methods.

3.3. Visual question answering

In the clinical domain, Visual Question Answering (VQA) represents a computer-assisted diagnostic technique that offers clinical decision-making support for image analysis [53].

Table 3 is an overview of the application of MDL on VQA. Commencing in 2018, ImageCLEF has been conducting an annual challenge for medical VQA, evaluating and ranking the performance of participating models. The mainstream VQA datasets in the medical domain include VQA-MED-2018 [54], VQA-MED-2019 [55], and VQA-MED-2020 [56], which were proposed by the challenge tasks. These datasets encompass radiographic images along with corresponding question–answer pairs. For instance, VQA-MED-2020 comprises 4,500 radiographic images and 4,500 question–answer pairs [56]. Additionally, VQA-RAD consists of 315 radiological images and 3,500 question–answer pairs [57]. The PathVQA dataset contains 1,670 pathological images and 32,799 question–answer pairs [58]. Liu et al. [59] introduced the SLAKE, a bilingual dataset that encompasses semantic labels and structural medical knowledge, incorporating more modalities and body parts. The SLAKE includes 642 images, 14,028 question–answer pairs, and 5,232 medical knowledge triplets.

A typical VQA model consists of four essential components: an image feature extractor, a question feature extractor, a multimodal fusion component, and a classifier or generator. For the image feature extractor, CNN-based pre-trained models such as ResNet [60] or VGGNet [61] are often employed to extract high-dimensional features from medical images. Liu et al. [62] introduced a bi-branch model that leverages both ResNet152 and VGG16 to extract sequence/spatial features and retrieve the similarity of image features, thereby enhancing the semantic understanding of images. For question feature extraction, recurrent neural networks (RNNs) such as Long-Short-Term Memory (LSTM) [63] and Gated Recurrent Unit (GRU) [64] are commonly utilized. Additionally, BERT-based models [5] have seen increasing use for extracting textual features. With regards to multimodal fusion, models from general domain VQA such as Stacked Attention Networks (SAN) [65], Bilinear Attention Networks (BAN) [66], Multimodal factorized bilinear (MFB) [67], and Multimodal Factorized High-order (MFH) [68] are often adopted. Sharma et al. [69] utilized MFB as a feature fusion technique to design an attention-based model that maximizes learning while minimizing complexity. Liu et al. [70] proposed a pre-training model called the Contrastive Pre-training and Representation process (CPRD), which effectively resolves the issue of limited MED-VQA data and demonstrates excellent performance.

The issue of data scarcity and lack of multilevel reasoning ability in Med-VQA has prompted the development of the Mixture of Enhanced Visual Features (MEVF) [87]. MEVF is a meta- learning-based approach that utilizes Model-Agnostic Meta-Learning (MAML) [88] and Convolutional Denoising Auto-Encoder (CDAE) [89] to effectively address the problem of insufficient data during image feature extraction. The proposed method has gained widespread use in subsequent studies and

Table 2
Overview of MDL models for report generation.

Ref.	Method	Dataset	Image type	Text type	Metrics	External validation	Explainability
Yuan et al. [24]	CNN, LSTM	CheXpert, IU X-Ray	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	–	√
Ni et al. [31]	CNN, LSTM	MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	√	√
Nishino et al. [32]	CNN, GRU, BERT	JCT, MIMIC-CXR	CXR	Radiology reports	BLEU, ROUGE, CRS	–	–
Miura et al. [29]	CNN, Transformer	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, CIDEr, BERTScore, factENT, factENTNLI	√	√
Chen et al. [33]	CNN, Transformer	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	–	√
You et al. [26]	CNN, Transformer, multi-head attention	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	√	–
Alfarghaly et al. [34]	CNN, word2vec, GPT-2	IU X-Ray	CXR	Radiology reports	BLEU, METEOR, ROUGE-L, CIDEr	√	√
Delbrouck et al. [35]	GRU, Fusion	MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE	–	–
Liu et al. [36]	CNN, BERT, multi- head attention	COVID-19 CT, CX-CHR	CT, CXR	Radiology reports	BLEU, ROUGE-L, CIDEr	√	√
Pahwa et al. [37]	CNN, Transformer	IU X-Ray, PEIR Gross	CXR, pathology images	Radiology reports, image captions	BLEU, METEOR, ROUGE-L	–	–
Zhou et al. [38]	CNN, BioSentVec, LSTM	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L, CIDEr, nKTD	–	√
Endo et al. [27]	CLIP	CheXpert, MIMIC-CXR	CXR	Radiology reports	Semb, BLEU, F1	–	–
Chen et al. [25]	CNN, TF-IDF, Transformer	IU X-Ray	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	–	–
Wang et al. [39]	BLIP	ImageCLEF 2020	Radiological images	Image captions	BLEU, METEOR, ROUGE-L, CIDEr, SPICE, BERTScore	–	–
Yan et al. [40]	CNN, BERT	COV-CTR, IU X-Ray, MIMIC-CXR	CT, CXR	Radiology reports	BLEU, METEOR, ROUGE-L	–	√
Tanwani et al. [28]	CNN, BERT, BAN	IU X-Ray	CXR	Radiology reports	BLEU	–	√
Keicher et al. [41]	CLIP	MIMIC-CXR	CXR	Radiology reports	AUC	–	–
Chen et al. [42]	CNN, Transformer, cross-modal memory	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	–	√
Qin et al. [43]	CNN, Transformer, cross-modal memory	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	√	√
Ma et al. [44]	CNN, LSTM, CMCL	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	√	–
Hassan et al. [45]	CNN, BERT, GRU	IU X-Ray	CXR	Radiology reports	BLEU, ROUGE	–	–
Moon et al. [46]	CNN, BERT, attention masking	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	BLEU, Precision, Recall, F1	√	√
You et al. [47]	CNN, Transformer, GRU	IU X-Ray	CXR	Radiology reports	BLEU, METEOR, ROUGE-L, CIDEr, SPICE, BERTScore	–	√
Delbrouck et al. [30]	CNN, BERT, semantic graph- based reward	IU X-Ray, MIMIC-CXR, RadGraph	CXR	Radiology reports	BLEU, ROUGE-L, F1cXb, factENT, factENTNLI, RGE, RGER, RGER	√	√
Serra et al. [48]	CNN, Transformer	CheXpert, MIMIC-CXR	CXR	Radiology reports	BLEU, METEOR, ROUGE-L	√	–

has been further improved by the introduction of the Question Conditioned Reasoning (QCR) and Type Conditioned Reasoning (TCR) modules by Zhan et al. [73], which enhance the model's reasoning ability. Do et al. [74] have proposed a Multiple Meta- model Quantifying (MMQ) model that achieves remarkable accuracy with the addition of metadata. The latest trends indicate that BERT and attention-based models are currently the most effective and are expected to be the future of VQA models. The RespsNet-10 proposed by Tanwani et al. [28] achieved an accuracy of 0.804 on the ImageCLEF 2018 and ImageCLEF 2019 datasets. Meanwhile, the study by Zhan et al. investigated the contrastive representation learning model UnICLAM with adversarial masking and obtained an accuracy of 0.831 on the SLAKE dataset [86].

Accuracy is the most widely used evaluation metric for VQA, typically associated with classification models and closed-ended questions. Meanwhile, some generation models designed to tackle open-ended problems may also employ alternative metrics, such as BLEU or WBSS [90], for evaluation purposes. While 12 articles have demonstrated the interpretability of the models, there has been a lack of studies that have sought to evaluate the results of VQA models from clinicians.

3.4. Cross-modal retrieval

Cross-modal retrieval encompasses two primary types of retrieval: image-to-text retrieval, which involves retrieving associated text for a given image, and text-to-image retrieval, which involves retrieving the associated image for a given text.

Table 4 summarizes an overview of the application of MDL on cross-modal retrieval. In the medical field, cross-modal retrieval tasks frequently involve radiological images and reports, such as those found in MIMIC-CXR [22] and CheXpert [23] datasets. The ROCO dataset, comprising over 81,000 radiology image-text pairs, is also widely employed in cross-modal retrieval tasks [91]. In addition, a small number of pathological captioning datasets exist. One is the ARCH dataset proposed by Gamper et al [92]. It comprises 7,579 image and description pairs extracted from medical articles on PubMed and pathology textbooks.

Most cross-modal retrieval tasks rely on matching image and text features through contrastive learning. This process involves both global and local feature matching, together with attention mechanisms. For

Table 3
Overview of MDL models for VQA.

Ref.	Method	Dataset	Image type	Text type	Metrics	External validation	Explainability
Liu et al. [71]	CNN, ETM, MFH	ImageCLEF 2018	Radiology images	Medical questions	WBSS, BLEU, CBSS	–	–
Ren et al. [72]	CNN, Transformer	ImageCLEF 2019	Radiology images	Medical questions	Accuracy, BLEU, WBSS	–	–
Zhan et al. [73]	QCR, TCR, MEVF, LSTM, BAN	VQA-RAD	Radiology images	Medical questions	Accuracy	–	–
Liu et al. [70]	CPRD, LSTM, BAN	SLAKE, VQA- RAD	Radiology images	Medical questions	Accuracy	–	√
Do et al. [74]	MMQ, LSTM, SAN/BAN	PathVQA, VQA- RAD	Radiology images, pathology images	Medical questions	Accuracy	–	–
Khare et al. [75]	CNN, BERT, self-attention	ImageCLEF 2019, VQA-RAD	Radiology images	Medical questions	Accuracy	–	√
Pan et al. [76]	MAML and CDAE, GRU, multi-view attention	VQA-RAD, VQA- RADPh	Radiology images	Medical questions	Accuracy	–	√
Gong et al. [77]	CNN, LSTM, cross- modal self-attention	VQA-RAD	Radiology images	Medical questions	Accuracy	–	–
Sharma et al. [69]	CNN, BERT, MFB	ImageCLEF 2019	Radiology images	Medical questions	Accuracy, AUC-ROC, AUC-PRC	–	√
Eslami et al. [78]	CLIP, MEVF, QCR	ROCO, SLAKE, VQA-RAD	Radiology images	Medical questions	Accuracy	–	–
Tanwani et al. [28]	CNN, BERT, BAN	VQA-RAD	Radiology images	Medical questions	Accuracy	–	√
Chen et al. [79]	Vision Transformer, BERT, co-attention	ImageCLEF 2019, MedICaT, ROCO, SLAKE, VQA- RAD	Radiology images	Medical questions	Accuracy	–	–
Wang et al. [80]	CDAE, LSTM, attention-based multi- granularity fusion	VQA-RAD	Radiology images	Medical questions	Accuracy	–	√
Naseem et al. [81]	CNN, LSTM, Transformer	PathVQA	Radiology images	Medical questions	Accuracy	–	√
Liu et al. [62]	CNN, Transformer	ImageCLEF 2018, ImageCLEF 2019, VQA-RAD	Radiology images	Medical questions	Accuracy, BLEU	–	–
Haridas et al. [82]	CNN, BERT, ViLBERT	SLAKE	Radiology images	Medical questions	Accuracy	–	–
Moon et al. [46]	CNN, BERT, attention masking	VQA-RAD	Radiology images	Medical questions	Accuracy	–	√
Chen et al. [83]	Vision Transformer, BERT, co-attention	ImageCLEF 2019, SLACK, VQA- RAD	Radiology images	Medical questions	Accuracy	–	√
Pan et al. [84]	MAML, CDAE, GRU, attention-based multimodal alignment	PathVQA, VQA- RAD	Radiology images, pathology images	Medical questions	Accuracy	–	√
Li et al. [85]	M2I2, Transformer, self-supervised pretraining	ImageCLEF 2022, PathVQA, SLAKE, VQA-RAD	Radiology images, pathology images	Medical questions	Accuracy	–	√
Zhan et al. [86]	Vision Transformer, BERT, adversarial masking	ROCO, SLAKE, VQA-RAD	Radiology images	Medical questions	Accuracy	–	√

Table 4
Overview of MDL models for cross-modal retrieval.

Ref.	Method	Dataset	Image type	Text type	Metrics	External validation	Explainability
Hsu et al. [93]	CNN, TF-IDF, DAN	MIMIC-CXR	CXR	Radiology reports	MRR, nDCG@K	–	–
Lara et al. [94]	CNN, TF-IDF	TCGA-PRAD	Pathology images	Pathology reports	Precision, MAP, GM- MAP, P@10, P@30	–	–
Ni et al. [31]	CNN, LSTM	MIMIC-CXR	CXR	Radiology reports	Accuracy, Precision, Recall, BLEU , ROUGE-L, METEOR	√	√
Zhang et al. [95]	CNN, CLIP	CheXpert, MIMIC-CXR	CXR	Radiology reports	Precision@K	√	√
Wang et al. [96]	Unified transformer	IU X-Ray, MIMIC-CXR, NIH-CXR	CXR	Radiology reports	Precision@K	–	–
Ji et al. [97]	CNN, Transformer	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	Recall@K	–	–
Huang et al. [98]	CNN, BERT, self-attention	CheXpert	CXR	Radiology reports	Precision@K	–	√
Chen et al. [79]	Vision Transformer, BERT, co-attention	ROCO	Radiology images	Image captions	Recall@K	–	–
Maleki et al. [99]	Vision Transformer, Text Transformer, self-attention	ARCH	Pathology images	Image captions	Recall@K	–	–
Moon et al. [46]	CNN, BERT, attention masking	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	Hit@K, Recall@K, Precision@K, MRR	–	√
Chen et al. [83]	Vision Transformer, BERT, co-attention	ROCO	CXR	Radiology reports	Recall@K	–	√
Wang et al. [100]	CLIP	CheXpert, COVID, MIMIC-CXR, RSNA	CXR	Radiology reports	Precision@K	–	√

example, Huang et al. [98] introduced GLoRIA which enables cross-modal retrieval through the averaging of global and local similarity metrics. In a separate study, Chen et al. [79] developed self-supervised multimodal masked autoencoders, achieving excellent performances for image-to-text retrieval and text-to-image retrieval on the ROCO dataset.

Maleki et al. [99] proposed LILE, a dual attention network that uses Transformers and an additional self-attention loss term to enhance internal features for text retrieval and image retrieval on the ARCH dataset.

Widely used measurements for assessing the performance of cross-modal retrieval are precision@K [46,94,98] and Recall@K [46,79,97,99], which quantify the accuracy of the first K retrieval results. Another commonly used metric is the mean reciprocal rank (MRR) [46,93]. Out of the 12 studies in our collection, only 2 works incorporated external validation, while 6 studies assessed the interpretability of their model.

3.5. Computer-aided diagnosis

MDL-based computer-aided diagnosis (CAD) is the use of generated output from multimodal data as an assisting tool for a clinician to make a diagnosis. Incorporating text modality in this context has been shown to provide supplementary features that can enhance performance in image classification. Currently, research in CAD mainly focuses on utilizing chest X-ray images in conjunction with corresponding radiological reports. It is expected that future pathological datasets will expand this field of research.

Table 5 summarizes the application of multimodal deep learning on CAD. There exist several commonly employed multimodal fusion strategies, including image-text embedding and contrastive learning. Image-text embedding refers to merging image and text features, which are then trained using supervised learning. For example, Wang et al. [101] introduced a Text-Image Embedding network (TieNet), which utilized a multi-task CNN-RNN framework and achieved an AUC of over 0.9 in thorax disease classification. In contrast, contrastive learning often involves image-text alignment and self-supervised learning. Tiu et al.

Table 5
Overview of MDL models for computer-aided diagnosis.

Ref.	Method	Dataset	Image type	Text type	Metrics	External validation	Explainability
Wang et al. [101]	CNN, LSTM	IU X-Ray, NIH-CXR	CXR	Radiology reports	AUC	–	–
Daniels et al. [106]	DNN	IU X-Ray, NIH-CXR	CXR	Radiology reports	AUC, Precision	–	–
Yan et al. [107]	CNN	DeepLesion	CT	Annotations	AUC, F1	–	√
Weng et al. [108]	CNN, BERT, Early fusion	TCGA, TTH	Pathology images	Pathology reports	AUC	–	–
Lara et al. [94]	CNN, TF-IDF	TCGA-PRAD	Pathology images	Pathology reports	Accuracy	–	–
Chauhan et al. [109]	CNN, BERT	MIMIC-CXR	CXR	Radiology reports	AUC, F1	√	√
Zhang et al. [95]	CNN, CLIP	CheXpert, COVIDx, MURA, RSNA	X-rays	Annotations, radiology report	AUC, Accuracy	√	√
Sonsbeek et al. [110]	CNN, BERT	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	AUC	–	√
Wang et al. [96]	Unified transformer	IU X-Ray, MIMIC-CXR, NIH-CXR	CXR	Radiology reports	AUC	–	–
Ji et al. [97]	CNN, Transformer	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	AUC	–	–
Liao et al. [111]	CNN, BERT	CheXpert, MIMIC-CXR	CXR	Radiology reports	AUC	–	–
Huang et al. [98]	CNN, BERT, self-attention	CheXpert, RSNA	CXR	Radiology reports	AUC, F1	–	√
Zheng et al. [112]	CNN, BERT, self-attention	Multimodal COVID-19 Pneumonia Dataset	CT, CXR, ultrasound	Doctor-patient dialogues	AUC, Accuracy, Precision, Sensitivity, Specificity, F1	–	–
Zhou et al. [113]	Vision Transformer, BERT	COVID-19 Image Data Collection, MIMIC-CXR, NIH-CXR, Shenzhen Tuberculosis, VinBigData Chest X-ray	CXR	Radiology reports	AUC	√	√
Yan et al. [40]	CNN, BERT	COV-CTR, IU X-Ray, MIMIC-CXR	CT, CXR	Radiology reports	AUC	–	√
Monajatipoor et al. [103]	Vision Transformer, BERT	IU X-Ray	CXR	Radiology reports	AUC	–	–
Jacenków et al. [114]	CNN, BERT	MIMIC-CXR	CXR	Radiology reports	AUC	√	–
Hassan et al. [45]	CNN, BERT, GRU	IU X-Ray	CXR	Radiology reports	AUC	–	–
Moon et al. [46]	CNN, BERT, attention masking	IU X-Ray, MIMIC-CXR	CXR	Radiology reports	AUC, F1	–	√
You et al. [47]	CNN, Transformer, GRU	IU X-Ray	CXR	Radiology reports	Accuracy	–	√
Chen et al. [83]	Vision Transformer, BERT, co-attention	MedICaT, MELINDA, MIMIC-CXR, ROCO	CXR	Radiology reports	Accuracy	–	√
Wang et al. [115]	Vision Transformer, BERT	CheXpert, COVIDx, MIMIC-CXR, RSNA	CXR	Radiology reports	AUC	–	–
Wang et al. [100]	CLIP	CheXpert, COVID, MIMIC- CXR, RSNA	CXR	Radiology reports	Accuracy	–	√
Tiu et al. [102]	Vision Transformer, CLIP	CheXpert, MIMIC-CXR	CXR	Radiology reports	AUC, MCC, F1	√	√

[102] proposed a self-supervised learning framework, CheXzero, which achieved expert-level performance in zero- shot thoracic disease classification without requiring manual labeling. Monajatipoor et al. [103] developed BERTHop, which leverages PixelHop++ [104] and VisualBERT [105] to enable the learning of associations between clinical images and notes. This model achieved an AUC of 0.98 on the IU X-Ray dataset [20].

Studies on COVID-19 diagnosis have recently been another popular trend. Zheng et al. [112] designed a multimodal knowledge graph attention embedding framework for diagnosing COVID-19, based on clinical images and doctor-patient dialogues. The proposed model performed better than single modality approaches, with an AUC of 0.99. In addition, the MedCLIP proposed by Wang et al. [100] achieved better performance than supervised models for the zero-shot classification task of COVID-related datasets.

The metrics employed to assess the performance of diagnostic classification primarily comprise the AUC and the F1-score. Additionally, the Matthews correlation coefficient (MCC) is utilized to assess the dissimilarity between model and expert classifications [102]. Out of the 24 studies gathered, 4 incorporated external validation, while 11 studies focused on elucidating the interpretability of the model.

3.6. Semantic segmentation

This group of studies investigates the effectiveness of image-text contrastive learning, which involves utilizing semantic segmentation to extract visual features that can be juxtaposed with textual features to facilitate the comprehension of the relationship between images and their corresponding textual descriptions (Table 6). Additionally, local alignment assessment in contrastive learning is evaluated using semantic segmentation techniques.

Typical datasets employed for semantic segmentation include SIIM [116] and RNSA [117]. The SIIM dataset consists of 12,047 chest radiographs, along with corresponding manual annotations. Similarly, the RNSA dataset includes 29,700 frontal view radiographs for evaluating evidence of pneumonia. Boecking et al. have recently proposed the MS-CXR [118] dataset, which comprises 1153 image-sentence pairs with annotated bounding boxes and corresponding phrases validated by radiologists. This dataset covers eight distinct cardiopulmonary radiology findings.

Image-text alignment and local representation learning are commonly used in MDL for semantic segmentation. These techniques can help improve the model's accuracy by enabling it to better understand the spatial relationships between different regions in the image and the relationship between visual and textual information [119]. Li et al. [120] proposed LViT, which used medical text annotations to improve the quality of image data and guide the generation of pseudo labels, leading to better segmentation performance. Müller et al. [121] devised a novel pre-training approach, LoVT, which aimed to specifically address localized medical imaging tasks. Their method exhibited superior performance on 10 out of 18 localized tasks in comparison to commonly employed pre-training techniques.

In all the research studies that we have gathered, Dice [122] has been utilized as a metric for measuring the similarity between predicted segmentation and ground truth. Additionally, mean intersection over union (mIoU) and contrast-to-noise ratio (CNR) have also been employed. Out of the 5 studies in our collection, no work incorporated external validation, while 2 studies assessed the interpretability of their model.

3.7. Other related tasks

During our article collection, we identified several works that, while not fitting into the aforementioned categories, are of considerable importance. These works include studies centered on medical image generation, object detection, multimodal predictive modeling, MDL-related databases, and libraries of pre-training models. Chambon et al. [123] fine-tuned the Stable Diffusion model to generate CXR images with realistic-looking abnormalities by employing domain-specific text prompts. In a separate publication, they introduced RoentGen, a model adept at synthesizing CXR images predicated upon text prompts present in radiological reports, resulting in a 25% enhancement in the representation capabilities of pneumothorax [124]. Qin et al. [125] scrutinized the implementation of pre-trained vision language models (VLM) for medical object detection and devised an approach to incorporate expert medical knowledge and image-specific information within the prompt, thereby augmenting the performance of zero-shot learning. Lin et al. [126] developed a survival prediction model using radiation reports and images to forecast ICU mortality. This model outperformed traditional single-modal machine learning methods with a higher C-index. Bai et al. [127] designed an interactive VQA system that empowers patients to upload their own multimodal data, choose the appropriate model in the library, and communicate with an AI robot for model evaluation. Delbrouck et al. [128] presented ViLMedic, a Vision-and-Language medical library, consisting of over 20 pre-trained models for various downstream tasks. This resource facilitates the real-world clinical translation of these models. Kovaleva et al. [129] released the first publicly available visual dialog datasets for radiology, highlighting the belief that integrating patients' medical history information would enhance the performance of traditional VQA models. Li et al. [130] summarized the performance of four pre-trained models for multimodal vision-and-language feature learning and visualized their attention mechanism. Evidenced by these studies, we believe multimodal vision-and- language learning will continue to expand its range of applications in the future, with more related databases and model libraries being established to promote its clinical use.

4. Discussion

Our scoping review identifies research related to MDL in biomedical images and texts on different downstream tasks, with specific attention to the datasets employed, model methodology, evaluation metrics,

Table 6
Overview of MDL models for semantic segmentation.

Ref.	Method	Dataset	Image type	Text type	Metrics	External validation	Explainability
Huang et al. [98]	CNN, BERT, self-attention	CheXpert, SIIM-ACR	CXR	Annotations	Dice	–	√
Müller et al. [121]	CNN, BERT, CLIP	COVID Rural, NIH-CXR, Object CXR, RSNA, SIIM-ACR	CXR	Annotations	Dice	–	–
Boecking et al. [118]	CNN, BERT	MIMIC-CXR, MS-CXR, RSNA	CXR	Annotations	Dice, mIoU, CNR	–	–
Li et al. [120]	CNN, Vision Transformer, BERT	MoNuSeg, QaTa-COV19	CXR, Pathology images	Annotations	Dice, mIoU	–	√
Wang et al. [115]	Vision Transformer, BERT	RNSA, SIIM-ACR	CXR	Annotations	Dice	–	–

external validation, and interpretability. Overall, the evidence suggests that deep learning models on multimodal medical image and text data can potentially improve diagnostic accuracy and clinical decision-making, showing promising results in several medical fields, including oncology, radiology, and pathology. However, our review also reveals challenges related to data imbalance, clinical knowledge, model fairness, and human evaluation. These findings are highly relevant to clinicians, researchers, and computer scientists interested in leveraging recent advances in artificial intelligence and deep learning to improve patient care and health outcomes.

In the realm of MDL, acquiring high-quality annotated data is crucial for the development and evaluation of models, yet several challenges persist in obtaining such datasets like MIMIC-CXR. First, the annotation of medical data is a laborious and time-consuming task that requires domain expertise and specialized tools to ensure accuracy and consistency, particularly when annotating both image and text data. This can result in insufficient annotated samples for certain modalities, leading to imbalanced datasets that adversely affect model performance. To address data scarcity and reduce the burden of expert annotation, multimodal *meta*-learning, and few-shot learning are poised to remain popular research topics in the medical field [41,80]. Second, the current trend in medical datasets predominantly features radiology images and their accompanying reports, with a limited representation of other imaging modalities such as pathological images, ultrasound, endoscopy, and text modalities such as clinical notes. This limits the broader clinical application of multimodal models. Future work should construct more multimodal datasets for different medical scenarios, and integrate these heterogeneous data into a system to realize multimodal cross-scenario learning. Thirdly, data privacy concerns are pronounced in the medical domain, necessitating the protection of sensitive patient information. However, this often leads to a lack of publicly available datasets, exacerbating the issue of insufficient and unbalanced data.

Advocating for open-access initiatives can help address this challenge by enabling researchers to access larger and more diverse datasets for model training and evaluation. In addition, implementing advanced privacy-preserving techniques, such as differential privacy and federated learning, can further alleviate privacy concerns while allowing researchers to utilize medical data (paper: Federated learning and differential privacy for medical image analysis).

Incorporating clinical knowledge into medical NLP has been identified as a major research direction that can enhance the model's performance and broaden its application in clinical practice [131–133]. However, the current research is limited in terms of the integration of clinical knowledge into MDL models. Incorporating clinical knowledge into the encoding stage can help learn useful visual features, leading to more accurate predictions. Specifically, clinical knowledge can provide insights into specific image features that are more clinically relevant, such as lesions or abnormalities, and guide the model to focus on these features during the encoding process. Chen et al. [25] integrated external knowledge into the features of TF-IDF and achieved improved performance in both report generation and diagnostic tasks. Furthermore, clinical knowledge can be particularly beneficial in scenarios with limited or new data, such as COVID-19-related datasets, where overfitting is more likely to occur. Liu et al. [36] incorporated external knowledge into the COVID-19 CT report generation task, generating fewer irrelevant words and higher BLEU scores. In addition, Chen et al. [83] demonstrated that aligning, reasoning, and learning using clinical knowledge could achieve higher accuracy than each approach individually in VQA. Future research could explore more sophisticated ways to integrate clinical knowledge into models, such as knowledge graphs and ontologies. Moreover, researchers could examine how clinical knowledge from diverse sources, such as electronic health records, medical literature, or expert opinions, can be integrated to enhance the models' performance and adaptability. It is also important to assess the clinical relevance and impact of models in real-world clinical settings by conducting clinical trials and involving clinicians and patients in the development and validation process.

Human evaluation is essential for assessing the practicality of the model in real-world clinical scenarios and providing insights into the model's decision-making process. However, human evaluation was not widely employed in the studies we collected. Out of the five downstream tasks covered in this review, report generation incorporated more external validation, as observed in 10 of 25 articles. Notably, no studies were found to introduce external validation for VQA or semantic segmentation tasks. The observed phenomenon could be attributed to the fact that human evaluation is time-consuming and costly [134]. Additionally, the absence of standardized protocols for human evaluation of MDL models in medical settings poses a significant challenge to the comparison and generalization of findings across studies [135]. Furthermore, the interdisciplinary collaboration between clinicians and computer scientists can be a formidable obstacle, owing to differences in their respective backgrounds and training that can hinder effective communication and seamless teamwork. Besides, clinicians often have limited availability to engage in such collaborative efforts, while computer scientists may face stringent deadlines for developing and testing models. In the future, there is a need to develop and adopt standardized protocols for the human evaluation of MDL models in medical applications.

Moreover, interdisciplinary workshops can help bridge the gap between clinicians and computer scientists and facilitate effective collaboration. Finally, effective automated metrics could provide a more objective and efficient approach to evaluating MDL models.

The fairness and explainability of MDL models also exhibit deficiencies. The absence of interpretability of the models engenders challenges in fostering trust in their predictions, thereby limiting their adoption in clinical practice. The lack of transparency in these"black boxes" further compounds the issue as it hinders the detection of errors and biases, thereby resulting in potential harm to patients [136]. Out of the 77 articles we collected, only 35 provided an exposition of the interpretability of the model, leveraging techniques such as heat maps and factual metrics.

Among them, the visual interpretation of CNN models, which are based on attention mechanisms, has gained increasing traction in the medical field [137]. However, it is worth noting that a significant number of articles do not explicitly consider the inclusion of interpretability as an improvement, and only a few employ a formal counterfactual evaluation [49]. Future MDL research endeavors must prioritize the development of interpretable models. Standardized methods are needed for evaluating and quantifying the interpretability of these models. Additionally, it is essential to engage in a continuous dialogue between clinicians, researchers, and computer scientists to ensure that the development of MDL models aligns with the values and needs of the medical community.

While our scoping review provides a comprehensive overview of the current state of MDL in biomedical images and texts, several limitations must be considered. First, our search strategy may have missed some relevant studies, as we focused on a limited set of databases and search terms. Second, we tried to understand the current state of the literature from their downstream tasks and applications. Still, there was a lack of a systematic summary of the methodology, particularly regarding the multimodal fusion strategy. Third, the heterogeneity of the included studies makes it difficult to compare and synthesize the evidence across different domains and contexts. Finally, our scoping review did not include a formal quality assessment of the studies, which may have affected the reliability and validity of the evidence. However, we believe the breadth and depth of the evidence we gathered will provide a robust foundation for future research and improvement.

5. Conclusion

In this scoping review, we systematically examined the current state of research on MDL in biomedical images and texts based on various

downstream tasks, including report generation, visual question answering, cross-modal retrieval, computer-aided diagnosis, and semantic segmentation. Our findings suggest that MDL can potentially improve diagnostic accuracy and clinical decision-making, but it also poses challenges related to data imbalance, clinical knowledge, human evaluation, and model fairness. We also discussed several areas for further investigation and improvement, such as developing more robust evaluation standards, collaborating with interdisciplinary institutions or individuals, and exploring new data sources and modalities. Our review has important implications for clinicians, researchers, and computer scientists interested in leveraging the latest advances in MDL to improve patient care and health outcomes.

6. Statement of Significance

Problems	Multimodal deep learning (MDL) in biomedical images and structured EHR data improves clinical decision-making, but the application of MDL methods with medical images and texts is still in its infancy, and explaining MDL methods remains a challenge.
What is Already Known	Previous review articles have focused on MDL methods, and MDL with medical images and structured EHR data.
What this Paper Adds	This scoping review provides a comprehensive synthesis of MDL research in biomedical images and texts over the past five years, focusing on 5 clinical tasks, explainability, and human evaluation to foster the application of multimodal language models in the medical field.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Library of Medicine under Award No. 4R00LM013001, NSF CAREER Award No. 2145640, and Amazon Research Award. This work is also supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

[1] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit. Med. 3 (2020) 136, https://doi.org/10.1038/s41746-020-00341-z.

[2] G. Holste, S.C. Partridge, H. Rahbar, D. Biswas, C.I. Lee, A.M. Alessio, End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, 2021: pp. 3294–3303. https://doi.org/10.1109/iccvw54120.2021.00368.

[3] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, M.P. Lungren, Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection, Sci. Rep. 10 (2020) 22147, https://doi.org/10.1038/s41598-020-78888-w.

[4] Y. Zhou, S.-C. Huang, J.A. Fries, A. Youssef, T.J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, M.P. Lungren, RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR, ArXiv [Eess.IV]. (2021). http://arxiv.org/abs/2111.11665.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv [Cs.CL], 2018. https://aclanthology.org/N19-1423.pdf.

[6] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, ArXiv [Cs.CL]. (2020) 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a- Abstract.html (accessed February 27, 2023).

[7] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR, 18–24 Jul 2021: pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

[8] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O.K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, F. Wei, Language Is Not All You Need: Aligning Perception with Language Models, ArXiv [Cs.CL]. (2023). http://arxiv.org/abs/2302.14045.

[9] Lars Heiliger, Anjany Sekuboyina, Bjoern Menze, Jan Egger, and Jens Kleesiek, Beyond Medical Imaging: A Review of Multimodal Deep Learning in Radiology, (2022). https://www.researchgate.net/profile/Jan-Egger-2/publication/358581125_Beyond_Medical_Imaging_A_Review_of_Multimodal_Deep_Learning_in_Radiology/links/620a1e5a7b05f82592ea5bda/Beyond-Medical-Imaging-A-Review-of-Multimodal-Deep-Learning-in-Radiology.pdf (accessed January 2, 2023).

[10] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L.A. Coburn, K.T. Wilson,BA. Landman, Y. Huo, Deep Multimodal Fusion of Image and Non-image Data in Disease Diagnosis and Prognosis: A Review, ArXiv [Cs.LG]. (2022). http://arxiv.org/abs/2203.15588.

[11] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nat. Med. 28 (2022) 1773–1784, https://doi.org/10.1038/s41591-022-01981-2.

[12] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: A scoping review, NPJ Digit Med. 5 (2022) 171, https://doi.org/10.1038/s41746-022-00712-8.

[13] G. Muhammad, F. Alshehri, F. Karray, A.E. Saddik, M. Alsulaiman, T.H. Falk, A comprehensive survey on multimodal medical signals fusion for smart healthcare systems, Inf. Fusion. 76 (2021) 355–375, https://doi.org/10.1016/j.inffus.2021.06.007.

[14] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, L. Cilar, Interpretability of machine learning-based prediction models in healthcare, Wiley Interdiscip, Rev. Data Min. Knowl. Discov. 10 (2020) e1379.

[15] E. Tjoa, C. Guan, A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, IEEE Trans Neural Netw Learn Syst. 32 (2021) 4793–4813, https://doi.org/10.1109/TNNLS.2020.3027314.

[16] Y. Zhang, Y. Weng, J. Lund, Applications of Explainable Artificial Intelligence in Diagnosis and Surgery, Diagnostics (Basel). 12 (2022), https://doi.org/10.3390/diagnostics12020237.

[17] B.H.M. van der Velden, H.J. Kuijf, K.G.A. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022), 102470, https://doi.org/10.1016/j.media.2022.102470.

[18] A.C. Tricco, E. Lillie, W. Zarin, K.K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M.D.J. Peters, T. Horsley, L. Weeks, S. Hempel, E.A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M.G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M.T. Macdonald, E.V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, S.E. Straus, PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation, Ann. Intern. Med. 169 (2018) 467–473, https://doi.org/10.7326/M18-0850.

[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, 2002. https://aclanthology.org/P02-1040.pdf (accessed January 24, 2023).

[20] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2016) 304–310, https://doi.org/10.1093/jamia/ocv080.

[21] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, ArXiv [Cs.CV], 2019. http://arxiv.org/abs/1901.07042.

[22] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data (2019). https://www.nature.com/articles/s41597-019-0322-0.

[23] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J. K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A. Y. Ng, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, AAAI. 33 (2019) 590–597, https://doi.org/10.1609/aaai.v33i01.3301590.

[24] J. Yuan, H. Liao, R. Luo, J. Luo, Automatic Radiology Report Generation Based on Multi- view Image Fusion and Medical Concept Enrichment, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, 2019, pp. 721–729, https://doi.org/10.1007/978-3-030-32226-7_80.

[25] W. Chen, H. Pan, K. Zhang, X. Du, Q. Cui, VMEKNet: Visual Memory and External Knowledge Based Network for Medical Report Generation, in: PRICAI 2022: Trends in Artificial Intelligence, Springer Nature, Switzerland, 2022: pp. 188–201. https://doi.org/10.1007/978-3-031-20862-1_14.

[26] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, 2021, pp. 72–82, https://doi.org/10.1007/978-3-030-87199-4_7.

[27] M. Endo, R. Krishnan, V. Krishna, A.Y. Ng, P. Rajpurkar, Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model, in: S. Roy, S. Pfohl, E. Rocheteau, G.A. Tadesse, L. Oala, F. Falck, Y. Zhou, L. Shen,

G. Zamzmi, P. Mugambi, A. Zirikly, M.B.A. McDermott, E. Alsentzer (Eds.), Proceedings of Machine Learning for Health, PMLR, 2021: pp. 209–219. https://proceedings.mlr.press/v158/endo21a.html.
[28] A.K. Tanwani, J. Barral, D. Freedman, RepsNet: Combining Vision with Language for Automated Medical Reports, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, 2022: pp. 714–724. https://doi.org/10.1007/978-3-031-16443-9_68.
[29] Y. Miura, Y. Zhang, E.B. Tsai, C.P. Langlotz, D. Jurafsky, Improving factual completeness and consistency of image-to-text radiology report generation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020. https://doi.org/10.18653/v1/2021.naacl-main.416.
[30] J.-B. Delbrouck, P. Chambon, C. Bluethgen, E. Tsai, O. Almusa, C.P. Langlotz, Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards, ArXiv [Cs.CL]. (2022). http://arxiv.org/abs/2210.12186.
[31] J. Ni, C.-N. Hsu, A. Gentili, J. McAuley, Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.176.
[32] T. Nishino, R. Ozaki, Y. Momoki, T. Taniguchi, R. Kano, N. Nakano, Y. Tagawa, M. Taniguchi, T. Ohkuma, K. Nakamura, Reinforcement learning with imbalanced dataset for data-to-text medical report generation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.202.
[33] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating Radiology Reports via Memory-driven Transformer, ArXiv [Cs.CL]. (2020). http://arxiv.org/abs/2010.16056.
[34] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, A. Fahmy, Automated radiology report generation using conditioned transformers, Inf. Med. Unlocked 24 (2021), 100557, https://doi.org/10.1016/j.imu.2021.100557.
[35] J.-B. Delbrouck, C. Zhang, D. Rubin, QIAI at MEDIQA 2021: Multimodal Radiology Report Summarization, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021: pp. 285–290. https://doi.org/10.18653/v1/2021.bionlp-1.33.
[36] G. Liu, Y. Liao, F. Wang, B. Zhang, L. Zhang, X. Liang, X. Wan, S. Li, Z. Li, S. Zhang, S. Cui, Medical-VLBERT: Medical Visual Language BERT for COVID-19 CT Report Generation With Alternate Learning, IEEE Trans Neural Netw Learn Syst. 32 (2021) 3786–3797, https://doi.org/10.1109/TNNLS.2021.3099165.
[37] E. Pahwa, D. Mehta, S. Kapadia, D. Jain, A. Luthra, MedSkip: Medical report generation using skip connections and integrated attention, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, 2021: pp. 3409–3415. https://doi.org/10.1109/iccvw54120.2021.00380.
[38] Y. Zhou, L. Huang, T. Zhou, H. Fu, L. Shao, Visual-textual attentive semantic consistency for medical report generation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2021: pp. 3985–3994. https://doi.org/10.1109/iccv48922.2021.00395.
[39] X. Wang, J. Li, ImageSem Group at ImageCLEFmedical Caption 2022 task: Generating Medical Image Descriptions based on Vision-Language Pre-training, 2022. http://ceur-ws.org/Vol-3180/paper-124.pdf (accessed January 5, 2023).
[40] B. Yan, M. Pei, Clinical-BERT: Vision-language pre-training for Radiograph Diagnosis and Reports Generation, (2022). https://www.aaai.org/AAAI22Papers/AAAI-4013.YanB.pdf.
[41] M. Keicher, K. Mullakaeva, T. Czempiel, K. Mach, A. Khakzar, N. Navab, Few-shot Structured Radiology Report Generation Using Natural Language Prompts, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2203.15723.
[42] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal Memory Networks for Radiology Report Generation, ArXiv [Cs.CL]. (2022). http://arxiv.org/abs/2204.13258.
[43] H. Qin, Y. Song, Reinforced Cross-modal Alignment for Radiology Report Generation, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022: pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
[44] X. Ma, F. Liu, S. Ge, X. Wu, Competence-based Multimodal Curriculum Learning for Medical Report Generation, ArXiv [Cs.CL]. (2022). http://arxiv.org/abs/2206.14579.
[45] A. Hassan, M. Sirshar, M.U. Akram, M. Umar Farooq, Analysis of multimodal representation learning across medical images and reports using multiple vision and language pre-trained models, in: 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST), IEEE, 2022. https://doi.org/10.1109/ibcast54850.2022.9990154.
[46] J.H. Moon, H. Lee, W. Shin, Y.-H. Kim, E. Choi, Multimodal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training, IEEE J Biomed Health Inform. PP (2022). https://doi.org/10.1109/JBHI.2022.3207502.
[47] J. You, D. Li, M. Okumura, K. Suzuki, JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022: pp. 5989–6001. https://aclanthology.org/2022.coling-1.523.
[48] F. Dalla Serra, W. Clackett, H. MacKinnon, C. Wang, F. Deligianni, J. Dalton, A.Q. O'Neil, Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022: pp. 615–624. https://aclanthology.org/2022.aacl-main.47.
[49] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, D. Capurro, A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images, ACM Comput. Surv. 54 (2022) 1–40, https://doi.org/10.1145/3522747.
[50] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005: pp. 65–72. https://aclanthology.org/W05-0909.
[51] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: T.S. B. Out (Ed.), Association for Computational Linguistics, Spain, Barcelona, 2004, pp. 74–81.
[52] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E.P. Reis,EKUN Fonseca, H.M. Ho Lee, Z.S.H. Abad, A.Y. Ng, C.P. Langlotz, V.K. Venugopal, P. Rajpurkar, Evaluating progress in automatic chest X-ray radiology report generation, BioRxiv. (2022). https://doi.org/10.1101/2022.08.30.22279318.
[53] Q. Wu, P. Wang, X. Wang, X. He, W. Zhu, Medical VQA, in: Q. Wu, P. Wang, X. Wang, X. He, W. Zhu (Eds.), Visual Question Answering: From Theory to Application, Springer Nature Singapore, Singapore, 2022: pp. 165–176. https://doi.org/10.1007/978-981-19-0964- 1_11.
[54] S.A. Hasan, Y. Ling, O. Farri, J. Liu, H. Muller, M. Lungren, Overview of ImageCLEF 2018 medical domain visual question answering task, (2018). https://ceur-ws.org/Vol-2125/paper_212.pdf (accessed February 9, 2023).
[55] A. Ben Abacha, S.A. Hasan, V.V. Datla, J. Liu, D. Demner-Fushman, H. Muller, VQA- med: Overview of the medical visual question answering task at ImageCLEF 2019, (2019). http://ceur-ws.org/Vol-2380/paper_272.pdf (accessed February 9, 2023).
[56] A. Ben Abacha, V.V. Datla, S.A. Hasan, D. Demner-Fushman, H. Muller, Overview of the VQA-med task at ImageCLEF 2020: Visual question answering and generation in the medical domain, (2020). http://star.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-2696/paper_106.pdf (accessed February 9, 2023).
[57] J.J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Sci. Data 5 (2018), 180251, https://doi.org/10.1038/sdata.2018.251.
[58] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, PathVQA: 30000+ Questions for Medical Visual Question Answering, ArXiv [Cs.CL]. (2020). http://arxiv.org/abs/2003.10286.
[59] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), ieeexplore.ieee.org, 2021: pp. 1650–1654. https://doi.org/10.1109/ISBI48211.2021.9434010.
[60] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
[61] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ArXiv [Cs.CV]. (2014). http://arxiv.org/abs/1409.1556.
[62] S. Liu, X. Zhang, X. Zhou, J. Yang, BPI-MVQA: a bi-branch model for medical visual question answering, BMC Med. Imaging 22 (2022) 79, https://doi.org/10.1186/s12880-022- 00800-x.
[63] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[64] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, ArXiv [Cs.NE]. (2014). http://arxiv.org/abs/1412.3555.
[65] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked Attention Networks for Image Question Answering, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 21–29. https://doi.org/10.1109/CVPR.2016.10.
[66] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear Attention Networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2018: pp. 1571–1581.
[67] Z. Yu, J. Yu, J. Fan, D. Tao, Multimodal factorized bilinear pooling with co-attention learning for visual question answering, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017: pp. 1821–1830. https://doi.org/10.1109/iccv.2017.202.
[68] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering, IEEE Trans Neural Netw Learn Syst. 29 (2018) 5947–5959, https://doi.org/10.1109/TNNLS.2018.2817340.
[69] D. Sharma, S. Purushotham, C.K. Reddy, MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain, Sci. Rep. 11 (2021) 19826, https://doi.org/10.1038/s41598-021-98390-1.
[70] B. Liu, L.-M. Zhan, X.-M. Wu, Contrastive Pre-training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, 2021: pp. 210–220. https://doi.org/10.1007/978-3-030-87196-3_20.
[71] F. Liu, Y. Peng, M.P. Rosen, An Effective Deep Transfer Learning and Information Fusion Framework for Medical Visual Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, 2019: pp. 238–247. https://doi.org/10.1007/978-3-030-28577-7_20.

[72] F. Ren, Y. Zhou, CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering, IEEE Access 8 (2020) 50626–50636, https://doi.org/10.1109/ACCESS.2020.2980024.
[73] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, in: Medical Visual Question Answering via Conditional Reasoning, in, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2345–2354, https://doi.org/10.1145/3394171.3413761.
[74] T. Do, B.X. Nguyen, E. Tjiputra, M. Tran, Q.D. Tran, A. Nguyen, Multiple Meta-model Quantifying for Medical Visual Question Answering, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, 2021: pp. 64–74. https://doi.org/10.1007/978-3-030-87240-3_7.
[75] Y. Khare, V. Bagal, M. Mathew, A. Devi, U.D. Priyakumar, C.V. Jawahar, MMBERT: Multimodal BERT Pretraining for Improved Medical VQA, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), ieeexplore.ieee.org, 2021: pp. 1033–1036. https://doi.org/10.1109/ISBI48211.2021.9434063.
[76] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, K. Shi, MuVAM: A Multi-View Attention-based Model for Medical Visual Question Answering, ArXiv [Cs.CV]. (2021). http://arxiv.org/abs/2107.03216.
[77] H. Gong, G. Chen, S. Liu, Y. Yu, G. Li, Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering, in: Proceedings of the 2021 International Conference on Multimedia Retrieval, Association for Computing Machinery, New York, NY, USA, 2021: pp. 456–460. https://doi.org/10.1145/3460426.3463584.
[78] S. Eslami, G. de Melo, C. Meinel, Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?, ArXiv [Cs.CV]. (2021). http://arxiv.org/abs/2112.13906.
[79] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, T.-H. Chang, Multimodal Masked Autoencoders for Medical Vision-and-Language Pre-training, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, 2022: pp. 679–689. https://doi.org/10.1007/978-3-031-16443-9_65.
[80] H. Wang, H. Pan, K. Zhang, S. He, C. Chen, M2FNet: Multi-granularity Feature Fusion Network for Medical Visual Question Answering, in: PRICAI 2022: Trends in Artificial Intelligence, Springer Nature Switzerland, 2022: pp. 141–154. https://doi.org/10.1007/978- 3-031-20865-2_11.
[81] U. Naseem, M. Khushi, J. Kim, Vision-Language Transformer for Interpretable Pathology Visual Question Answering, IEEE J Biomed Health Inform. PP (2022). https://doi.org/10.1109/JBHI.2022.3163751.
[82] HT Haridas, M.M. Fouda, Z.M. Fadlullah, M. Mahmoud, B.M. ElHalawany, M. Guizani, MED-GPVS: A deep learning-based joint biomedical image classification and visual question answering system for precision e-health, in: ICC 2022 - IEEE International Conference on Communications, IEEE, 2022. https://doi.org/10.1109/icc45855.2022.9839076.
[83] Z. Chen, G. Li, X. Wan, R. Align, Learn,, in: Enhancing Medical Vision-and-Language Pre-training with Knowledge, in, Association for Computing Machinery, New York, NY, USA, 2022, pp. 5152–5161, https://doi.org/10.1145/3503161.3547948.
[84] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, K. Shi, AMAM: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering, Knowl.-Based Syst. 255 (2022), 109763, https://doi.org/10.1016/j.knosys.2022.109763.
[85] P. Li, G. Liu, L. Tan, J. Liao, S. Zhong, Self-supervised vision-language pretraining for Medical visual question answering, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2211.13594.
[86] C. Zhan, P. Peng, H. Wang, T. Chen, H. Wang, UnICLAM:Contrastive Representation Learning with Adversarial Masking for Unified and Interpretable Medical Vision Question Answering, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2212.10729.
[87] BD. Nguyen, T.-T. Do, B.X. Nguyen, T. Do, E. Tjiputra, Q.D. Tran, Overcoming Data Limitation in Medical Visual Question Answering, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, 2019: pp. 522–530. https://doi.org/10.1007/978-3-030-32251-9_57.
[88] C. Finn, P. Abbeel, S. Levine, Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, 06–11 Aug 2017: pp. 1126–1135. https://proceedings.mlr.press/v70/finn17a.html.
[89] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction, in: Artificial Neural Networks and Machine Learning – ICANN 2011, Springer Berlin Heidelberg, 2011: pp. 52–59. https://doi.org/10.1007/978-3- 642-21735-7_7.
[90] G. Soganciogl u, H. Öztürk, A. Özgür, BIOSSES: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics 33 (2017) i49–i58, https://doi.org/10.1093/bioinformatics/btx238.
[91] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C.M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, 2018: pp. 180–189. https://doi.org/10.1007/978-3-030- 01364-6_20.
[92] J. Gamper, N. Rajpoot, Multiple instance captioning: Learning representations from histopathology textbooks and articles, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2021: pp. 16549–16559. https://doi.org/10.1109/cvpr46437.2021.01628.
[93] T.-M.H. Hsu, W.-H. Weng, W. Boag, M. McDermott, P. Szolovits, Unsupervised Multimodal Representation Learning across Medical Images and Reports, ArXiv [Cs.LG]. (2018). http://arxiv.org/abs/1811.08615.
[94] J.S. Lara, V.H. Contreras O., S. Otálora, H. Müller, F.A. González, Multimodal Latent Semantic Alignment for Automated Prostate Tissue Classification and Retrieval, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, 2020: pp. 572–581. https://doi.org/10.1007/978-3-030-59722- 1_55.
[95] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive Learning of Medical Visual Representations from Paired Images and Text, ArXiv [Cs.CV]. (2020). https://www.semanticscholar.org/paper/6dd9f99cecd38504b667d320eb2a6267a9fee35d (accessed January 4, 2023).
[96] X. Wang, Z. Xu, L.K. Tam, D. Yang, D. Xu, Self-supervised Image-text Pre-training With Mixed Data In Chest X-rays, ArXiv [Cs.CV]. (2021). https://www.semanticscholar.org/paper/c49d8a576ee4c1778eafd75f00565f75864054e4 (accessed January 4, 2023).
[97] Z. Ji, M.A. Shaikh, D. Moukheiber, S.N. Srihari, Y. Peng, M. Gao, Improving Joint Learning of Chest X-Ray and Radiology Report by Word Region Alignment, Mach Learn Med Imaging. 12966 (2021) 110–119. https://doi.org/10.1007/978-3-030-87589-3_12.
[98] S.-C. Huang, L. Shen, M.P. Lungren, S. Yeung, GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2021: pp. 3942– 3951. https://doi.org/10.1109/iccv48922.2021.00391.
[99] D. Maleki, H.R. Tizhoosh, LILE: Look In-Depth before Looking Elsewhere – A Dual Attention Network using Transformers for Cross-Modal Information Retrieval in Histopathology Archives, ArXiv [Cs.CV]. (2022). https://proceedings.mlr.press/v172/maleki22a.html.
[100] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive Learning from Unpaired Medical Images and Text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022: pp. 3876–3887. https://aclanthology.org/2022.emnlp- main.256.
[101] X. Wang, Y. Peng, L. Lu, Z. Lu, R.M. Summers, TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018: pp. 9049–9058. https://doi.org/10.1109/cvpr.2018.00943.
[102] E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng. 6 (2022) 1399–1406, https://doi.org/10.1038/s41551-022-00936-9.
[103] M. Monajatipoor, M. Rouhsedaghat, L.H. Li, C.-C. Jay Kuo, A. Chien, K.-W. Chang, BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, 2022: pp. 725–734. https://doi.org/10.1007/978-3-031-16443- 9_69.
[104] Y. Chen, M. Rouhsedaghat, S. You, R. Rao, C.-C. Jay Kuo, Pixelhop++: A Small Successive-Subspace-Learning-Based (Ssl-Based) Model For Image Classification, in: 2020 IEEE International Conference on Image Processing (ICIP), ieeexplore.ieee.org, 2020: pp. 3294–3298. https://doi.org/10.1109/ICIP40778.2020.9191012.
[105] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, VisualBERT: A Simple and Performant Baseline for Vision and Language, ArXiv [Cs.CV]. (2019). http://arxiv.org/abs/1908.03557.
[106] Z.A. Daniels, D.N. Metaxas, Exploiting Visual and Report-Based Information for Chest X-RAY Analysis by Jointly Learning Visual Classifiers and Topic Models, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), ieeexplore.ieee.org, 2019: pp. 1270–1274. https://doi.org/10.1109/ISBI.2019.8759548.
[107] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, R.M. Summers, Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019: pp. 8523–8532. https://doi.org/10.1109/cvpr.2019.00872.
[108] W.-H. Weng, Y. Cai, A. Lin, F. Tan, P.-H.C. Chen, Multimodal Multitask Representation Learning for Pathology Biobank Metadata Prediction, ArXiv [Cs.CV]. (2019). http://arxiv.org/abs/1909.07846.
[109] G. Chauhan, R. Liao, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, P. Golland, Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment, Med. Image Comput. Comput. Assist. Interv. 12262 (2020) 529–539, https://doi.org/10.1007/978-3-030-59713-9_51.
[110] T. van Sonsbeek, X. Zhen, M. Worring, L. Shao, Variational Knowledge Distillation for Disease Classification in Chest X-Rays, in: Information Processing in Medical Imaging, Springer International Publishing, 2021: pp. 334–345. https://doi.org/10.1007/978-3-030- 78191-0_26.
[111] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, W. M. Wells, Multimodal Representation Learning via Maximization of Local Mutual Information, Med. Image Comput. Comput. Assist. Interv. 12902 (2021) 273–283, https://doi.org/10.1007/978-3-030-87196-3_26.
[112] W. Zheng, L. Yan, C. Gou, Z.-C. Zhang, J. Jason Zhang, M. Hu, F.-Y. Wang, Pay attention to doctor-patient dialogues: Multimodal knowledge graph attention image-text embedding for COVID-19 diagnosis, Inf. Fusion. 75 (2021) 168–185, https://doi.org/10.1016/j.inffus.2021.05.015.
[113] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, Y. Yu, Generalized radiograph representation learning via cross-supervision between images and free-text

radiology reports, Nature, Machine Intelligence. (2021) 32–40, https://doi.org/10.1101/2021.11.02.21265838.
[114] G. Jacenków, A.Q. O'Neil, S.A. Tsaftaris, Indication as Prior Knowledge for Multimodal Disease Classification in Chest Radiographs with Transformers, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), ieeexplore.ieee.org, 2022: pp. 1–5. https://doi.org/10.1109/ISBI52829.2022.9761567.
[115] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, L. Yu, Multi-Granularity Cross-modal alignment for generalized medical visual representation learning, ArXiv [Cs.CV]. (2022). https://doi.org/10.48550/ARXIV.2210.06044.
[116] Society for Imaging Informatics in Medicine: SIIM-ACR pneumothorax segmentation (2019), (n.d.). https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation.
[117] G. Shih, C.C. Wu, S.S. Halabi, M.D. Kohli, L.M. Prevedello, T.S. Cook, A. Sharma, J.K. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R.R. Gill, M.C.B. Godoy, S. Hobbs, J. Jeudy, A. Laroia, P.N. Shah, D. Vummidi, K. Yaddanapudi, A. Stein, Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia, Radiol Artif Intell. 1 (2019) e180041.
[118] B. Boecking, N. Usuyama, S. Bannur, D.C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, O. Oktay, Making the most of text semantics to improve biomedical vision–language processing, in: Lecture Notes in Computer Science, Springer Nature, Switzerland, Cham, 2022, pp. 1–21, https://doi.org/10.1007/978-3-031-20059-5_1.
[119] Z. Zhao, J. Hu, Z. Zeng, X. Yang, P. Qian, B. Veeravalli, C. Guan, MMGL: Multi-Scale Multi-View Global-Local Contrastive Learning for Semi-Supervised Cardiac Image Segmentation, in: 2022 IEEE International Conference on Image Processing (ICIP), ieeexplore.ieee.org, 2022: pp. 401–405. https://doi.org/10.1109/ICIP46576.2022.9897591.
[120] Z. Li, Y. Li, Q. Li, P. Wang, Y. Zhang, D. Guo, L. Lu, D. Jin, Q. Hong, LViT: Language meets Vision Transformer in Medical Image Segmentation, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2206.14718.
[121] P. Müller, G. Kaissis, C. Zou, D. Rueckert, Joint Learning of Localized Representations from Medical Images and Reports, in: Computer Vision – ECCV 2022, Springer Nature Switzerland, 2022: pp. 685–701. https://doi.org/10.1007/978-3-031-19809-0_39.
[122] W.R. Crum, O. Camara, D.L.G. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Trans. Med. Imaging 25 (2006) 1451–1461, https://doi.org/10.1109/TMI.2006.880587.
[123] P. Chambon, C. Bluethgen, C.P. Langlotz, A. Chaudhari, Adapting Pretrained Vision- Language Foundational Models to Medical Imaging Domains, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2210.04133.
[124] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J.M.Z. Chaves, T.M. Abraham, S. Purohit, C.P. Langlotz, A. Chaudhari, RoentGen: Vision-Language Foundation Model for Chest X-ray Generation, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2211.12737.
[125] Z. Qin, H. Yi, Q. Lao, K. Li, Medical Image Understanding with Pretrained Vision Language Models: A Comprehensive Study, ArXiv [Cs.CV]. (2022). http://arxiv.org/abs/2209.15517.
[126] M. Lin, S. Wang, Y. Ding, L. Zhao, F. Wang, Y. Peng, An empirical study of usingradiology reports and images to improve ICU-mortality prediction, IEEE Int Conf Healthc Inform. 2021 (2021) 497–498. https://doi.org/10.1109/ichi52183.2021.00088.
[127] H. Bai, X. Shan, Y. Huang, X. Wang, MVQAS: A Medical Visual Question Answering System, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2021: pp. 4675–4679. https://doi.org/10.1145/3459637.3481971.
[128] J.-B. Delbrouck, K. Saab, M. Varma, S. Eyuboglu, P. Chambon, J. Dunnmon, J. Zambrano, A. Chaudhari, C. Langlotz, ViLMedic: a framework for research at the intersection of vision and language in medical AI, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022: pp. 23–34. https://doi.org/10.18653/v1/2022.acl-demo.3.
[129] O. Kovaleva, C. Shivade, S. Kashyap, K. Kanjaria, J. Wu, D. Ballah, A. Coy, A. Karargyris, Y. Guo, D.B. Beymer, A. Rumshisky, V.M. Mukherjee, Towards Visual Dialog for Radiology, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2020: pp. 60–69. https://doi.org/10.18653/v1/2020.bionlp-1.6.
[130] Y. Li, H. Wang, Y. Luo, A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), ieeexplore.ieee.org, 2020: pp. 1999–2004. https://doi.org/10.1109/bibm49941.2020.9313289.
[131] T.J. Callahan, I.J. Tripodi, H. Pielke-Lombardo, L.E. Hunter, Knowledge-Based Biomedical Data Science, Annu Rev Biomed Data Sci. 3 (2020) 23–41, https://doi.org/10.1146/annurev-biodatasci-010820-091627.
[132] A. Roy, S. Pan, Incorporating medical knowledge in BERT for clinical relation extraction, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021: pp. 5357–5366. https://doi.org/10.18653/v1/2021.emnlp- main.435.
[133] B. Hao, H. Zhu, I.C. Paschalidis, Enhancing clinical bert embedding using a biomedical knowledge base, 28th International Conference On. (2020). https://par.nsf.gov/servlets/purl/10220309.
[134] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M.M. Hoffman, W. Xie, G.L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A.E. Carpenter, A. Shrikumar, J. Xu, E.M. Cofer, C.A. Lavender, S.C. Turaga, A.M. Alexandari, Z. Lu, D.J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L.K. Wiley, M.H. S. Segler, S.M. Boca, S.J. Swamidass, A. Huang, A. Gitter, C.S. Greene, Opportunities and obstacles for deep learning in biology and medicine, J. R. Soc. Interface 15 (2018), https://doi.org/10.1098/rsif.2017.0387.
[135] H.P. Cowley, M. Natter, K. Gray-Roncal, R.E. Rhodes, E.C. Johnson, N. Drenkow, T.M. Shead, F.S. Chance, B. Wester, W. Gray-Roncal, Author Correction: A framework for rigorous evaluation of human performance in human and machine learning comparison studies, Sci. Rep. 12 (2022) 11559, https://doi.org/10.1038/s41598-022-15857-5.
[136] F. Xie, H. Yuan, Y. Ning, M.E.H. Ong, M. Feng, W. Hsu, B. Chakraborty, N. Liu, Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies, J. Biomed. Inform. 126 (2022), 103980, https://doi.org/10.1016/j.jbi.2021.103980.
[137] T. He, J. Guo, N. Chen, X. Xu, Z. Wang, K. Fu, L. Liu, Z. Yi, MediMLP: Using Grad-CAM to Extract Crucial Variables for Lung Cancer Postoperative Complication Prediction, IEEE J. Biomed. Health Inform. 24 (2020) 1762–1771, https://doi.org/10.1109/JBHI.2019.2949601.