Beyond Medical Imaging - A Review of Multimodal Deep Learning in Radiology

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

01-02-2022 / 03-02-2022

CITATION

Heiliger, Lars; Sekuboyina, Anjany; Menze, Bjoern; Egger, Jan; Kleesiek, Jens (2022): Beyond Medical Imaging - A Review of Multimodal Deep Learning in Radiology. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.19103432.v1>

DOI

[10.36227/techrxiv.19103432.v1](https://doi.org/10.36227/techrxiv.19103432.v1)

Beyond Medical Imaging: A Review of Multimodal Deep Learning in Radiology

Lars Heiliger*, Anjany Sekuboyina*, Bjoern Menze, Jan Egger, and Jens Kleesiek

Abstract—Healthcare data are inherently multimodal. Almost all data generated and acquired during a patient’s life can be hypothesized to contain information relevant to providing optimal personalized healthcare. Data sources such as ECGs, doctor’s notes, histopathological and radiological images all contribute to inform a physician’s treatment decision. However, most machine learning methods in healthcare focus on single-modality data. This becomes particularly apparent within the field of radiology, which, due to its information density, accessibility, and computational interpretability, constitutes a central pillar in the healthcare data landscape and traditionally has been one of the key target areas of medically-focused machine learning. Computer-assisted diagnostic systems of the future should be capable of simultaneously processing multimodal data, thereby mimicking physicians, who also consider a multitude of resources when treating patients. Before this background, this review offers a comprehensive assessment of multimodal machine learning methods that combine data from radiology and other medical disciplines. It establishes a modality-based taxonomy, discusses common architectures and design principles, evaluation approaches, challenges, and future directions. This work will enable researchers and clinicians to understand the topography of the domain, describe the state-of-the-art, and detect research gaps for future research in multimodal medical machine learning.

Index Terms—deep learning, multimodal, radiology

I. INTRODUCTION

In a clinical setting, various sources of data streams exist, that contain information concerning a patient. This includes radiological imaging data such as radiographs, magnetic resonance imaging, computed tomography imaging, nuclear medicine or molecular imaging etc. Abundant non-imaging data is also collected from or associated with every patient in the form of radiology reports, laboratory tests, electroencephalographs (EEG), etc. Complementing this, a patient can also be associated with on-clinical data in the form of demographic information, genetic information, patient history and so on. Diagnosing a patient often involves the medical expert collating and analysing data from all these data sources (cf. Fig. 1). We refer this data, beyond just the imaging modality, as multimodal data in medicine.

Radiological imaging constitutes a significant portion of patient data and plays a major role in patient diagnosis.

*Equal contribution.

Lars Heiliger, Jan Egger, and Jens Kleesiek are with the Institute for AI in Medicine, University Hospital Essen, Essen, Germany e-mail: {lars.heiliger, jens.kleesiek}@uk-essen.de

Anjany Sekuboyina is with the Technical University of Munich, Munich, Germany, and the University of Zurich, Zurich, Switzerland e-mail: anjany.sekuboyina@tum.de

Bjoern Menze is with the University of Zurich, Zurich, Switzerland

Fig. 1. Medical data concerning a patient. Clinical information including imaging and non-imaging data is acquired. Alongside this, non-clinical data could also be obtained based on genome sequencing, user surveys etc.

It is usually the role of a radiologist to summarise the findings in these images, thereby assisting the physician in making a clinical decision. Understandably, the number of radiological examinations have been consistently increasing over the decades. This increase, coupled with an identified shortage of radiologists [1] therefore establishes the necessity of computer-assisted support systems that can automatically process radiological images. Consequently, medical image analysis has been an active sub-field of computer sciences and the recent advances in machine learning (ML) have only accelerated the research towards building said support systems based on medical images [2]. We also see these systems slowly percolating into medical use [3].

The domain of machine learning dealing with multimodal data processing is termed *multimodal machine learning* or *multimodal deep learning* (MMDL), and has been a rapidly growing sub-domain in the computer vision and machine learning fields. Multimodal machine learning draws parallels with multi-sensory inputs to humans, such as auditory and visual inputs. It has applications in numerous fields such as audio-visual speech recognition (stack of images, audio), visual question answering (image, text), media summarising (stack of images, text), conditioned image generation (image, text), autonomous driving (stack of image, Radar, Lidar, other sensors) etc. In each of these cases, the ultimate goal is

to construct a feature representation that conglomerates the information present across the input modalities.

In a clinical setting, it is desirable that a computer-based support system emulate the medical expert, and hence should be capable of consuming not only a radiology image but also the patient’s supporting data, reason with it, and process the (possibly) multimodal information, eventually reaching a clinically feasible decision. In this work, we focus on MMDL applied to radiology. Specifically, we focus on research that reasons over medical images in tandem with supporting, non-imaging data modalities. Note that in this work, multiple modalities refer to a combination of imaging and non-imaging data. There exists cases where the data could contain multiple imaging modalities (e.g. MR and CT of a same patient, or multiple MR contrasts [4]). For the purpose of this work, such data is considered to be of a single modality, i.e the *imaging* modality.

Prior work. We draw inspiration from two previous reviews around multimodal machine learning, [5] introduced a taxonomy to the general field of multimodal machine learning and identified the key approaches employed therein, visual representation learning, feature translation, feature space alignment, and fusion. [6] specifically addressed the approaches employing fusion for combining knowledge across modalities, focusing on medical imaging and electronic health records (EHR). Both these approaches review the field from a methodological perspective, not taking into account the data pipeline. Contrarily, we structure our work around data modalities, with the belief that such an overview reveals an interesting trend in the methods employed towards learning from multimodal data. We intend this work to augment the information provided in [5] and [6] in multimodal machine learning sketching the entire field from data and methods to the medical tasks that have been addressed.

II. PRELIMINARIES

We first identify the various data modalities, imaging and non-imaging. We then identify the combinations of these multiple modalities as they have been commonly used in the literature we review. Finally, we also introduce a taxonomy, different from [5], [6], that enables us to investigate the research works from the perspective of the data.

A. Data modalities and their combinations

We identify three major data modalities (cf. Fig. 1) that are routinely used in radiological diagnosis:

- **Imaging data.** This includes N -dimensional imaging information acquired in a clinical practice. This includes, x-ray, computed tomography (CT), magnetic resonance imaging (MRI), functional MRI (fMRI), nuclear or molecular imaging etc.
- **Non-imaging data.** This includes all the non-imaging data that constitutes as *supporting data* that enables a radiologist to make a more informed decision about the patient or the content in the image. This can further be sub-divided into:

- **Unstructured data**, especially free-text radiology reports describing the images, or image captions.
- **Structured data**, which includes spreadsheet-like data (discrete or continuous-valued) concerning the patient’s state. In a clinical setting, this would include information about blood gas analyses, blood pressure, heart rate, EEG reports etc. They could also include non-clinical information such as patient’s age, sex, genomic sequences, and so on.

Focusing on radiology, we treat images to be the primary modality supported by the non-imaging modalities described above. This results in the two modality combinations: **{imaging + unstructured non-imaging data}** and **{imaging + structured non-imaging data}**. We propose this first level of categorization from an architectural standpoint. Specifically, the processing of unstructured data is considerably different (e.g. free text using recurrent models) compared to structured data (e.g. vector information using dense layers).

B. Taxonomy

Branching out from the data-based categorization introduced earlier, we further sort the research works into a methodology-based taxonomy. This is based on the ‘sections’ of a neural network where each modality can be linked to, as illustrated in Fig. 2. We also complete each modality combination by listing out the publicly-accessible datasets in the category, with the intent of facilitating an easy entry-point for researchers. In further detail:

- **Modality fusion.** We start by reviewing the most direct multimodal approaches, wherein fusion of information from each modality happens in a supervised learning setting. This is usually achieved by a naive combination of information in their feature representations, for example concatenation, addition, mean etc. As described in [6], fusion approaches can further be categorised into early, joint, and late fusion, depending on the stage in the neural network where the features are fused.
- **Representation learning.** These methods deal with learning *enriched* feature representations from data leveraging information from multiple modalities. Our interpretation of this task circumscribes the approaches dealing with self-supervised learning, weakly-supervised learning, latent-space alignment, co-learning [5] etc. This category encompasses works that combine modalities in ways other than naive feature fusion, and becomes highly relevant in settings with abundant, unlabeled multimodal data and scarce, possibly unimodal, labeled data.
- **Modality translation.** Methods in this category deal with translating data from one modality (e.g. x-ray images) to another modality (e.g. radiology reports). This is especially challenging because the neural networks are supposed to learn highly non-linear mappings, typically between an image and its supporting data.
- **Datasets.** Lastly, we provide a list the publicly available medical, multimodal datasets that contain radiological images supported by data from other modalities. We hope

that such an overview of methodology and data provides the reader a holistic picture of the modality combination.

C. Search Strategy

To the best of our knowledge, this is the first review in the field of medical imaging that provides a thorough overview of the diverse field of multimodal machine learning from a modality-standpoint. For the survey, we followed a tree-based search on PubMed with keywords based on the proposed taxonomy, e.g. for image + unstructured non-image works, a prospective search string would be "deep learning AND medical imaging AND radiology reports AND representation learning". This search is also augmented by screening Google Scholar. Search phrases were also enriched using surrogates, as shown in Table I.

Note that an initial search of PubMed and Google Scholar, extending the search terms from [6], retrieved results which did not capture the diversity of the research in multimodal machine learning in medical imaging. We believe this to be due to a lack of *keywords* describing the sub-domains of multimodal machine learning other than feature fusion in the medical context.

D. Target Readers

This review is primarily aimed at AI-informed radiologists and computer scientists. For a radiologist, our focus on a data-modality-based categorization would be of interest, providing them an overview of the datasets as well as benchmarks. Furthermore, we believe that our review will help to identify gaps in data availability from a modality perspective, thus fostering the curation of more publicly-accessible datasets. For a computer scientist, our second level of categorization, based on the technical aspects of learning across modalities, will provide an overview of the algorithmic advances. We also hope that this review's focus on highlighting the medical tasks that MMDL has been employed to address will facilitate algorithmic development based on the task's state-of-the-art easier.

Taxonomy Usage: If our reader is a radiologist with imaging data and radiology reports (unstructured non-imaging), they will be pointed towards Sec. III-A, dealing with *imaging and unstructured non-imaging* modality combination to get an idea of the various tasks this data can be used to address. Similarly, if our reader is a computer scientist working on the task of representation learning from MRI images they could look at Sec. III-B3, to gather ideas about how to learn better representations if metadata could be collected.

III. A REVIEW OF MULTIMODAL MACHINE LEARNING IN RADIOLOGY

A. Imaging + Unstructured Non-imaging

In this section, we review works that process images and free-text. The latter modality is typically in natural language albeit specialised to the medical domain, e.g. radiology reports. Typical tasks in this combination involve image classification (augmented by textual information), domain translation (report generation), image retrieval, caption generation etc.

1) *Fusion:* As most public datasets utilize rule-based NLP methods for extracting image labels from radiology reports, the number of works on fusion leveraging medical imaging and text data is small.

Following the state-of-the-art before the prevalence of the transformer architecture, [7] proposed the text-image-embedding network (TieNet), an end-to-end trainable CNN+RNN architecture, for extracting discriminatory image-text representations of pairs of chest radiographs (CXR) and reports. The image encoding pipeline draws inspiration of the ResNet [8] architecture and an LSTM [9] encodes the text. Following [6], the fusion categorises as joint fusion. The conducted experiments show an improved AUC in multi-label disease classification when utilizing multimodal versus unimodal inputs. The differences to the best performing unimodal models (trained on reports only) are 0.013, 0.003, and 0.005 for the three different test sets under consideration. With the advent of entirely attention-based architectures in NLP [10] and their adaption to computer vision [11], an unified architecture for jointly processing images and texts is feasible. [12] introduced some of the proposed transformer-based vision-and-language models to the medical domain. More precisely, they implement LXMERT [13], VisualBERT [14], UNITER [15], and PixelBERT [16] trained with paired CXR and report data. In their experiments, the transformer-based joint fusion approach performs better than TieNet and unimodal text-only models (TieNet, BERT [17], ClinicalBERT [18]). In both conducted experiments, the best performing multimodal model is VisualBERT achieving an AUC of 0.987 and 0.987 whereas the best unimodal model is ClinicalBERT with an AUC of 0.972 and 0.974, respectively.

2) *Representation Learning:* As one of the first works, [19] learned joint embeddings of chest radiographs and radiology reports and applied them to the task of image-text retrieval. The representations are learned under an unsupervised adversarial training regime and compared with fully supervised learned representations. The image is featurized by a DenseNet121 [20], and the report is encoded by a concatenation of term frequency-inverse document frequency (TF-IDF) over bi-grams and several distributed embeddings. The experiments show that with representations learned with unsupervised adversarial training, only a limited amount of supervision is needed to be on par with fully supervised methods in the performed retrieval tasks.

[21] train a neural network jointly on chest radiographs and associated radiology reports in a semi-supervised manner to assess the severity of pulmonary edema. The image is processed by a series of residual blocks and BERT encodes the radiology reports. The model is trained by minimizing a loss function that adds a joint embedding loss which encourages representations of matched pairs to be closer than those of mismatched pairs to two cross-entropy losses. The paired data stem from the MIMIC-CXR dataset and the labels are extracted from the reports. The conducted experiments show the superiority of the joint embeddings by comparing the macro-F1 metric of their pre-trained image encoder (0.51) with its architectural equivalent that was solely trained on the images in a supervised way (0.43).

Fig. 2. **Taxonomy for the methods with a modality combination.** Each method is reviewed based on how it uses each modality in a modality combination, (a) fusion of the modality features, (b) learning enriched representations from multiple modalities, or (c) translating one modality to the other.

TABLE I
SURROGATE MAPPING: SURROGATES USED TO ENRICH THE SEARCH TERMS FOR THIS REVIEW.

Entity	Surrogate
Deep Learning	Deep Learning, Neural Network, Convolutional Neural Network, Vision Transformer, Recurrent Neural Network
Medical Imaging	X-ray, CT, MRI, Ultrasound, Angiography
Unstructured Non-Imaging Data	Text, Report, Radiology Report, Discharge Letter
Structured Non-Imaging Data	Tabular, Time Series, Longitudinal, ECG, EEG, Metadate, Patient data
Taxonomy	Multimodal Datasets, Fusion, Feature Fusion, Representation Learning, Self-Supervised Learning, Contrastive Learning, Unsupervised Learning, Semi-Supervised Learning, Translation, Report Generation

[22] propose the ConVIRT (contrastive visual representation learning from text) framework to learn visual representations from paired medical images and textual data in an unsupervised way. The introduced method contrasts the image representations from an image encoding pipeline with text representations from a text encoding pipeline. The former is a ResNet-50 architecture, whereas the BERT encoder compresses the text to a representation. The minimization of a bidirectional contrastive loss function maximizes the agreement between true image-text pairs as opposed to random image-text pairs. This pre-training leverages the MIMIC-CXR and a non-public musculoskeletal dataset collected from the Rhode Island Hospital system. In their empirical analysis, the weights of the image encoder serve as initialization for the considered downstream tasks. In detail, they conduct four medical image classification tasks and both an image-image and text-image retrieval task. ConVIRT outperforms ImageNet pre-training as well as other strong in-domain initialization methods that also exploited paired image-text data. Furthermore, ConVIRT achieves parity with them with an order of magnitude fewer data. Besides exploring that label-efficiency of multimodal pre-training, the authors provide empirical evidence of the benefit of ConVIRT over unimodal unsupervised image representation learning approaches (SimCLR, MoCo v2). Similarly, the work of [23] also utilizes contrastive learning by performing contrastive language-image pre-training (CLIP) as proposed in [24]. As the work focused on report generation, its results are discussed in more detail in the subsequent translation subsection.

[25] introduce a contrastive learning framework for learning global-local representations for images using attention mechanism (GLoRIA). Motivated by the observation that pathologies often occupy only small portions of the medical image, they

learn global and localized features jointly. Hence, the objective function consists of global contrastive loss as in [22] that aligns image and text representations of a positive pair and a local contrastive loss that aligns attention-weighted image representations and their respective text representations. The medical image and the reports are encoded by a ResNet-50 and a BERT architecture, respectively. The conducted experiments are based on the CheXpert [26], RSNA Pneumonia [27], [28], and the SIIM Pneumothorax¹ dataset. When compared with [22] and other multimodal representation learning methods, GLORIA achieves state-of-the-art performance in the tasks of image-text retrieval and image classification by being more label-efficient.

Similar to [25], [29] propose the localized representation learning from vision and text (LoVT) pre-training method, that extends ConVIRT by also learning localized features. Within LoVT, a ResNet-50 and BERT encode the image and the text, respectively. In contrast to [25], the text encoding works on sentence-level rather than word-level. Further, an attention-based alignment model computes cross-modal representations that are aligned with the unimodal representations by using a local contrastive loss. The latter is minimized along with a global contrastive loss to learn the framework. Their comprehensive experiments include 18 localized downstream tasks ranging from semantic segmentation to object detection on five public CXR datasets. The image-text self-supervised methods (i.e. LoVT or ConVIRT) perform better than its unimodal (self-supervised) competitors in 15 out of 18 and LoVT achieves the best results in 11 out of 18 downstream tasks. Their experiments provide additional evidence for label-efficiency and shows that multimodal pre-training is superior

¹<https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation>

to unimodal pre-training in most of the conducted experiment runs.

[30] propose a transformer-based image-text pre-training framework that jointly learns the representations of both image and text data from mixed data (i.e. paired and unpaired medical images and reports). The encoding of each modality draws inspiration from the transformer architecture [10] whereby both encoders share weights. In order to model the correlation between the image and text data, the authors introduce two attention-based modules that can be placed between the encoder and the decoder. While the UNIT (UNified Transformer) module computes cross-modality cross-attention and therefore requires both modalities during training and testing, the UWOX (UNIT WithOut Cross fusion) module uses self-attention with shared weights, and hence, does not need image and text data during training. The framework is learned by minimizing the masked word prediction loss, the masked patch prediction loss, and an optional pair matching loss. The network is trained with the MIMIC-CXR and NIH14-CXR [27] datasets, and the OpenI-CXR dataset serves as an external validation set. They achieve better results compared to unimodal transformer models in the experiments on the tasks of classification, retrieval, and image generation. Considering classification, for instance, UWOX trained with images and the image-only transformer achieve an AUC of 0.763 and 0.739, respectively.

3) *Translation*: We review modality translation for image+text in two parts: 1) Image-to-text, consisting of tasks such as radiology report generation, image captioning, and visual question answering (VQA). 2) Text-to-image, with tasks such as image retrieval and image synthesis. Arguably, image-to-text translation, especially radiology report generation is a fertile research domain with numerous works.

Report generation. In [31], a set of baselines for text image-to-text generation are provided, starting from image conditioned n -gram models and nearest neighbor-based report retrieval based on query image (CNN) to jointly learning from image and text using CNN-RNN models. This sets the stage for most of the works in this domain. In the landmark work of TieNet [7], the image embeddings are utilized to conditionally generate attention encoded text embeddings using an LSTM, resulting in natural language-based radiology reports. Commonly, RNN-based generation is further split into two levels, a word-decoder and a sentence decoder [32]. The sentence decoder generates topics for the sentences based on the image embedding, and the word generator takes the topics and generates word sequences while attending to the image embedding. Novelties in this aspect involve obtaining better representations for the query image, for example using a knowledge graph [33] or triplet and matching loss functions [30]. As a natural next step, in recent works, transformers have replaced RNN architectures for language modelling [34]. Note that in most of the works presented above, paired image and reports are necessary. In [35], this requirement is eliminated thanks to a modality-independent representation learning augmented by a pre-constructed knowledge graph (co-occurrence of labels in the text corpus).

As an example of retrieval-based generation, KERP [36]

utilises the image embedding to *retrieve* text templates that are then paraphrased to generate the final report. Every stage of KERP, namely encoding, retrieval, and paraphrasing are graph-based. For image-based report retrieval, [37] used a CNN to generate core labels (or original image labels) and fine finding labels (structured label sets obtained by parsing the reports). Based on these two, a 'nearest' report from the database is retrieved and post-processed to generate the final report. A recent manifestation of retrieval-based approach [23] uses representations obtained using contrastive learning to measure similarity.

Deriving from the report generation task, is the task of *image captioning*. Arguably, this is a simpler task where the requirement on the generated text to be free-flowing, natural, and clinically precise is relatively weaker. However, there exists a few works [38] that address this task, which are also built on the popular CNN-RNN framework.

Visual question answering. Compared to report generation, the clinical relevance of VQA in radiology is limited. Most of the approaches in this realm deal with the ImageCLEF dataset. Approaches attempting VQA typically have an architecture involving two encoders, one for the query image and one for question-text, followed by a text decoder. Note that, as in image captioning, the requirement of the text language to be natural and clinical accurate is weaker for VQA, relative to report generation.

Image synthesis In computer vision, image generation is a well-explored research field, supported by the rise of generative adversarial networks (GAN). The generated images are typically used for augmenting the training data for another downstream task. For example: class-conditioned generation of ImageNet examples can help augment data for training better classifiers. However, in radiology, conditioned image generation is relatively less explored. We opine that this to be because of higher standards for the generated image to be used for any downstream task. To the best of our knowledge, [40], is one work addressing synthesis of images (chest radiographs) from radiology reports using progressive-generative adversarial networks. [41] is another peculiar work in this domain that synthesizes images from text embeddings in order to evaluate quality of the textual embeddings. A text-to-image generator is used to synthesis images, and the Wasserstein distance between the distribution of the validation images and that of the generated images is used as a proxy to evaluate the quality of the text embeddings.

4) *Datasets*: One of the landmark datasets combining images and text is the MIMIC-CXR [42]–[44], a dataset containing 227,835 imaging studies (CXR) for 65,379 patients along with semi-structured free-text radiology reports. In the JPEG version of the same dataset (MIMIC-CXR-JPG [45]), the studies are labelled for 14 structured labels extracted using two rule-based NLP tools [26], [46]. In fact, there exist many large chest radiograph datasets such as the Indiana Chest X-ray collection [47] consisting of 3,996 radiology reports and 8,121 associated images (frontal and lateral radiographs). The very large PADCHEST dataset [48] from the Hospital San Juan Hospital, consists of 160,000 images obtained from 67,000 patients annotated for 174 radiological findings, 19 differential

TABLE II

PERFORMANCE OVERVIEW WORKS UTILIZING IMAGING + UNSTRUCTURED NON-IMAGING DATA INCLUDED IN THE REVIEW (MM: MULTIMODAL, UM: UNIMODAL). WE MARK PAPERS ON REPRESENTATION LEARNING THAT DID NOT INCLUDE UNIMODAL PRE-TRAINING BASELINES BY AN ASTERISK (*).

Reference	MM Framework	Fusion / Pre-Training Task	Task	Metric	MM Performance	UM Performance
[7]	TieNet	Joint Fusion	Classification	AUC	0.989	0.976 (TieNet)
[12]	VisualBERT	Joint Fusion	Classification	AUC	0.987	0.972 (ClinicalBERT)
				AUC	0.987	0.974 (ClinicalBERT)
[21]*	unnamed	Ranking-based	Classification	Macro-F1	0.51	0.43 (unnamed)
[22]	ConVIRT	Contrastive	Classification	AUC	0.927	0.908 (ResNet-50)
				AUC	0.881	0.876 (ResNet-50)
				Accuracy	0.924	0.903 (ResNet-50)
				AUC	0.890	0.870 (ResNet-50)
[25]*	GLoRIA	Contrastive	Classification	AUC	0.881	0.814 (ResNet-50)
				AUC	0.886	0.763 (ResNet-50)
			Segmentation	Dice	0.634	0.635 (ResNet-50)
[29]	LoVT	Contrastive	Object Detection	mAP	0.181	0.222 (ResNet-50)
				fROC	0.621	0.625 (ResNet-50)
			Segmentation	Dice	0.512	0.462 (ResNet-50)
				Dice	0.441	0.385 (ResNet-50)
[30]	UWOX	Masked {Word, Patch} Prediction	Classification	AUC	0.763	0.739 (Transformer, image-only)

TABLE III

TASK-WISE OVERVIEW OF WORKS DEALING WITH MODALITY TRANSLATION BETWEEN IMAGING AND UNSTRUCTURED NON-IMAGING DATA

Category	Reference
Report Generation	[31], [7], [32], [33], [30], [34], [35], [36], [37], [23], [38]
Visual Question Answering	[39]
Image Synthesis	[40], [41]

diagnoses, and 104 anatomic locations all arranged hierarchically by mapping to the Unified Medical Language System (UMLS). PADCHEST also provides free text radiology reports in the Spanish language. These datasets are frequently used in research dealing with multi-label image classification, report generation, and image retrieval, as discussed in the subsequent sections.

The ImageCLEF [39] series of challenges from 2003 to 2020 are also a popular source of data for joint image and text modalities, mainly aimed at language independent image annotation, multimodal information retrieval, and image retrieval [49], [50]. This data mainly consists of radiology (or biomedical) images extracted by crawling over PubMed articles along with the image captions, thus including multiple imaging modalities (x-rays, MRI, CT, angiographs etc.). The ROCO dataset [51] and the MedICaT [52] datasets are also similarly extracted (figures and captions) and are used for image captioning and concept tagging tasks.

Very recently, the CANDID-PTX dataset [53] consisting of 19,237 chest radiographs and their corresponding radiology reports was released. Interestingly, CANDID-PTX contains region-annotations, visual segmentation annotations for pneumothorax, bounding boxes for acute rib fractures, and line markings for chest tubes. We envision a rich mix of computer vision and natural language processing methods on this new dataset.

5) *Remarks:* Interestingly, we did not find much literature dealing with fusion of image and text modalities. This is expected because most of the datasets (e.g. MIMIC-CXR-JPG, PADCHEST etc.) rely on rule-based mining of labels from the associated radiology reports. We opine that this field will be more lucrative if the labels were human-generated or generated using a means independent of any modality. From a representation learning perspective, we see that BERT

[17] seems to be a popular inspiration for many NLP-related approaches ([22], [25], [29]), and we believe this to hold, given the prevalence of transformer-architectures in solving various tasks. In general, representation learning utilizing this modality combination achieves superior performance when compared to supervised unimodal in many downstream tasks (cf. Table II). Modality translation from images to text is the most researched sub-domain we have reviewed (cf. Table III). Tagged as report generation or image captioning, current state-of-the-art involves exploiting transformer architectures for both image and language processing [34]. Fostering the research in this modality combination further, we believe there is abundant publicly-accessible data. However, as reported in Table IV, most of them focus on chest radiographs. We believe a diversification here would lead to more interesting research in an already fertile field.

B. Imaging + Structured Non-imaging

In this section, we review the combination of medical imaging with structured non-imaging data. In our understanding, structured non-imaging refers to tabular data, typically sourced from the electronic health records, that are usually represented in a spreadsheet. We envision the columns to indicate measurements (discrete as well as continuous) whereas the rows are determined by an existence of an inherent temporal order in the data. In its absence the rows of the spreadsheet denote the different patients (i.e. cross-sectional data). A temporal order implies the rows to represent points in time. Each row could represent a single point in time (i.e. time series data) or a point in time coupled with a patient identifier (i.e. longitudinal data). The variables could include demographic data (patient's age, sex, height, and weight), vital signs (heart rate, blood pressure, and temperature), or other laboratory tests or assessments.

TABLE IV
DATASET OVERVIEW: A LIST OF PUBLICLY ACCESSIBLE MULTIMODAL DATASETS COMBINING IMAGING + UNSTRUCTURED NON-IMAGING

Acronym	Publication	Image Modality	Non-image Modality	# Imaging Studies
MIMIC-CXR	[42], [43], [44]	CXR	Radiology Reports	377,110
MIMIC-CXR-JPG	[45]	CXR	Radiology Reports	377,110
OpenI	[47]	CXR	Radiology Reports	8,121
PADCHEST	[48]	CXR	Radiology Reports	>160,000
ImageCLEF	[49], [50]	Diverse	Image Captions	Evolving
ROCO	[51]	Diverse	Image Captions	>81,000
MEDICAT	[52]	Diverse	Image Captions	217,060
CANDID-PTX	[53]	CXR	Radiology Reports	19,237

TABLE V
DATASET OVERVIEW: A LIST OF PUBLICLY ACCESSIBLE MULTIMODAL DATASETS COMBINING IMAGING + STRUCTURED NON-IMAGING

Acronym	Publication	Image Modality	Type of Structured Non-imaging data	# Imaging Studies
ABIDE I	[54]	MRI	Cross-sectional	1,112
ABIDE II	[55]	MRI	Cross-sectional	2,156
ADNI	-	MRI	Longitudinal	> 7,000
UK Biobank	[56]	MRI, X-ray, Ultrasound	Cross-sectional	100,000
RadFusion	[57]	CT	Cross-Sectional	1,794
TCIA	[58]	MRI, CT, Nuclear	Both	-
MIMIC-III (+imaging counterpart)	[59]	CXR	Longitudinal	-

1) *Fusion*: Considering fusion techniques dealing with medical imaging and clinical tabular data, we point the reader to the systematic review [6] that, amongst others, propose implementation guidelines to assess the performance differences of early, joint, and late fusion. Below, we outline works that were published after [6].

Following these guidelines, [60] implemented early, joint, and late multimodal fusion models for pulmonary embolism detection. The authors utilize a non-open dataset comprising CT scans and EHR data to train two unimodal and seven multimodal fusion models. The computer vision backbone is the PENet [61] architecture and a feedforward network encodes the tabular data. The late fusion approach performs best among the fusion models with an AUC of 0.947, and outperforms both, the image-only and the tabular-only (ElasticNet [62]), models with an AUC of 0.791 and 0.911, respectively. Besides proposing the RadFusion dataset, [57] also explore unimodal and multimodal models to benchmark the task of pulmonary embolism detection. The image-only and tabular-only model are a PENet model and an ElasticNet, respectively. Referring to [6], the implemented multimodal architecture is a late fusion model that is constructed from both unimodal models by mean pooling the predictions. With an AUC of 0.946, the multimodal late fusion model outperforms a PENet and an ElasticNet with AUCs of 0.796 and 0.922, respectively. In their retrospective study, [63] implement multimodal fusion models, two joint fusion and one late fusion architecture, for breast cancer classification. The non-open image and tabular data are composed of dynamic contrast-enhanced (DCE) MRI images and 18 associated clinical EHR measurement like demographic information, clinical indication, and mammographic breast density. Reduced to 2D maximum intensity projection, the DCE-MRI images are encoded by an ResNet-50 whereas feedforward network processes the tabular data. In their experiments, all fusion models exhibit a better performance (AUC: 0.898) when compared with their unimodal, image-only (0.849) and text-only (0.807), counterparts. Among the

fusion models, the joint fusion approach is superior at what the joint fusion model with learned image and tabular encoders works best.

2) *Representation Learning*: The SimCLR [64] based contrastive learning framework by [65] introduces the multi-instance contrastive learning method (MICLe) that incorporates pathology information into the positive pair creation. In contrast to traditional contrastive learning in computer vision where a positive pairs consist of two transformations of the same image, MICLe builds positive pairs from two transformations of possibly distinct images that share the same pathology (i.e. from the same patient). The downstream tasks comprise dermatology condition classification and multi-label CXR classification. The empirical analysis underlines the benefit of extending the pool of positive pairs by using patient metadata. Further, the authors created additional evidence for label-efficiency of (multimodal) pre-training and also showed that their MICLe pre-training outperformed unimodal SimCLR pre-training.

Similar to [65], [66] developed MedAug, a method that uses patient metadata for the selection of positive pairs. Integrated into a MoCo [67], [68] based contrastive learning framework, MedAug also requires the images of positive pairs to originate from the same patient. The authors explore study number and laterality based selection criteria for the creation of positive pairs. Their results in the downstream task of pleural effusion classification complement the work of [65], as the conducted experiments show that the performance depends on the metadata. More precisely, they showed that choosing all images of the same patient hurt the downstream performance, whereas a study number based selection criterion obtained the best performance.

3) *Translation*: One could imagine that *translation* from image to structured data (or tabular) domain, or vice versa would comprise of inferring EHR information or patient metadata given an image, or synthesising an image given patient metadata. The clinical relevance of metadata-to-image transla-

tion is unclear, speaking for the general lack of works in this realm. [69], for example, synthesis images based on patient metadata (scanner type or scan center). Since very few data fields are employed for conditional image generation, tagging the work as tabular-to-image translation is questionable. The converse (image-to-tabular) also has similar characteristics, i.e. a significant number of works that could potentially fall into the image-to-tabular translation typically just infer very few metadata (AD assessment [70], genomic signatures [71], [72], patient age estimation [73] etc.), making them typical classification or regression networks.

4) *Datasets*: The Autism Brain Imaging Data Exchange (ABIDE) and the Alzheimer’s Disease Neuroimaging Initiative (ADNI) are the most popular datasets consisting of imaging studies augmented with tabular clinical data. ABIDE [54], [55], for instance, consists of more than 1000 resting state functional MRI imaging data supported by phenotypic data such as gender, age, and assessment scores of a series of tests such as handedness score, performance IQ score, visual IQ score, score for social responses, medication name etc. This data is collected to check if a subject is on the autism spectrum or not. The preprocessed version of ABIDE’s neuroimaging data [74] is more accessible and hence easier to use. At a larger scale, ADNI is a series of studies such as ADNI 1, 2, 3, and GO, aimed at studying mild cognitive impairment and its progression into Alzheimer’s disease. ADNI deals with more than 7000 MRI and PET images from about 1700 subjects, supported by numerous non-imaging features² both clinical and genetic. ADNI also has a simplified counterpart more commonly used in medical image analysis literature, TADPOLE [75], which consists of a subset of ADNI-3 samples and features. TADPOLE does not include raw images but processed structural information about the images such as brain sub-region volumes, cortical thicknesses, ROI averages, etc. Further details can be found on the dataset’s webpage³.

The UK Biobank dataset [56] is another work in this direction consisting of multimodal imaging data (brain, cardiac, and abdominal MRI, ultrasound, and dual x-ray absorptiometry scans) as well as phenotypic data and genetic data from 100,000 participants. The data is mostly used for population studies, especially to investigate cross-sectional associations. Similarly, associations between genetic variations and phenotypic imaging are also studied. However, the authors claim that the data acquisition is still in progress and the future directions are yet to be ascertained in prospective studies.

More recently, RadFusion [57], a dataset combining high-resolution CT images with EHR data was released. RadFusion is aimed at pulmonary embolism detection using data from 1794 patients across demographic subgroups such as race, gender, and age.

Lastly, we mention two datasets The Cancer Imaging Archive (TCIA) [58] and MIMIC-III [59]. The former is a collection of medical images (CT, MR, and nuclear medicine), some collection of which are supported by image analyses, clinical, and genomic information. The latter is a dataset

of critical-care EHRs, whose data items can potentially be mapped to MIMIC-CXR-JPG and used as a multimodal dataset.

5) *Remarks*: For multimodal fusion, we found that early fusion is rarely chosen due to architectural constraints, as it involves fusing raw image with raw structured non-imaging data. We observed that the multimodal approaches consistently performed better than their unimodal counterparts (cf. Table VI). Recent efforts, such as [76], developed an architecture geared toward processing structured data using deep learning. We envision that such architectures will ultimately find their way into multimodal fusion of images and structured data modalities but the current state of research is ambiguous [77], [78]. Concerning learning representations from image and structured data, contrastive learning aided by patient metadata seems to be a go-to choice. Both [65] and [66] provides empirical evidence for the benefit of multimodal pre-training over unimodal pre-training. In [65], it is shown that multimodal pre-training is more label-efficient on the downstream task, when compared to fully supervised methods. From Table VI, we also see that existing work solely consider classification and retrieval tasks. We believe that other processing tasks (e.g. segmentation and object detection) could be improved using patient metadata as well. When dealing with images combined with structured data, we see a general lack of research in modality translation. Given an image, we envision multi-task learning of multiple metadata fields to be a prospective research direction. However, this lack of research could be attributed to the highly non-linear mapping that each of the parameter requires, thus making learning infeasible. Lastly, we believe that the datasets containing both imaging and structured auxiliary data need more pre-processing thus making them more accessible to computer scientists. As reported in Table V, all the data collections are very rich in information but difficult to navigate and construct the training and validation sets.

IV. DISCUSSION

Multimodal machine learning is a thriving field of research gaining increasing attention from the deep learning community [5]. In radiology, the motivation for using multiple modalities is very strong [79]–[81], wherein the interpretation of images improves significantly, given the data auxiliary clinical sources. With this motivation, we presented a comprehensive summary of 45 publications dealing with multimodal machine learning in radiology. In this section, we summarise our findings, state the limitations of our work, and sketch out prospective future directions for research in MMDL.

A. Trends in MMDL

Through this exercise, we observe certain trends in the field of MMDL, which can be categorised into the following categories:

- 1) *Data sources*: We find that the most accessible multimodal datasets are the MIMIC series of data, containing chest radiographs, radiology reports, electronic health reports, alongside other information such as admission

²<http://adni.loni.usc.edu/data-samples/data-types/>

³<https://tadpole.grand-challenge.org/Data/#List>

TABLE VI

PERFORMANCE OVERVIEW WORKS UTILIZING IMAGING + STRUCTURED NON-IMAGING DATA INCLUDED IN THE REVIEW (MM: MULTIMODAL, UM: UNIMODAL)

Reference	MM Framework	Fusion / Pre-Training Task	Task	Metric	MM Performance	UM Performance
[60]	PENet + ElasticNet	Late Fusion	Classification	AUC	0.947	0.911 (ElasticNet)
[57]	PENet + Elastic Net	Late Fusion	Classification	AUC	0.946	0.922 (ElasticNet)
[63]	ResNet-50 + MLP	Joint Fusion	Classification	AUC	0.898	0.849 (ResNet-50)
[65]	SimCLR + MICLe	Contrastive	Classification	Accuracy	0.688	0.646 (ResNet-50)
[66]	MedAUG	Contrastive	Classification	AUC	0.906	0.858 (ResNet-18)

status, medical costs etc. The research advancements garnered due to MIMIC call for the creation of similar large-scale, multi-modal datasets for other radiological data sources with miscellaneous anatomical focus.

- 2) Applications: Disease classification (*fusion*- and *representation learning*- based approaches) is the most common application addressed in a multimodal setting. This is inline with the motivation for the usage of multimodal data. Interestingly, domain translation from images to text seems to be a fertile research direction as well. We attribute this abundance to the cutting-edge research in the field of computer vision and natural language processing (e.g. transformers).
- 3) Approaches: We believe that the taxonomy presented in this work covers all the avenues where multimodal data can be employed (as an input and an output of a network). It also covers supervised and unsupervised approaches, the trend shifting from the former to the latter. From a modality perspective, we see CNNs being the go-to option for image processing. For text, the approaches vary from RNNs to LSTMs to transformers. Tabular is processed mostly using dense layers when being fed into a network, or is used for weak supervision, for example to choose positive and negative pairs in contrastive learning.
- 4) Prospects: With this review, we note that the growth of MMDL seems to be inline with the progress in the fields of computer vision, deep learning, and NLP. However, this growth is hindered by lack of sufficient data for benchmarking methods. This is understandable, given the privacy concerns in making healthcare data publicly available. We also observe a general lack of research interest in combining image information with time-series (or sequences) such as genetic signatures, ECGs, EEGs, etc. This could either be attributed to the lack of clinical necessity to scan this information simultaneously or just due to lack of data. In the next sections, we propose research directions that could circumvent these shortcomings.

B. Limitations

Out of the corpus of reviewed works, we selected publications that best represented the challenges of multimodal machine learning. While we aimed to give a comprehensive review in the field of multimodal representation learning, our review of the fusion and image + unstructured non-imaging translation literature is not exhaustive. The subjective nature

of our selection process constitutes a limitation. Since positive results are usually disproportionately reported, publication bias is an additional limitation of our work. A publication bias may lead to overestimating the benefit of analyzing multimodal over unimodal data. Inspired by the clinical practice of radiologists, we solely considered works dealing with image, text, and tabular data. Hence, the absence of additional modalities like audio, video, genomic sequences, etc. is a potential limitation as well as the fact that we did not include multimodal work without medical imaging. Lastly, from an information theoretic perspective, it is debatable if multiple modalities always help in a radiological setting. For example, a radiology report emphasizes on the *findings* while ignoring the *normal* parts of the image, thus containing lesser information compared to the image. However, a laboratory test could contain information not seen in the image, thereby helping in a multimodal setting. Our review does not discuss this feasibility of an improvement.

C. Future Research

This work showed that multimodal machine learning in radiology is a promising, and hence growing, discipline that emulates the information processing of radiologists. We categorized the existing work into a taxonomy that mirrors the diverse technical challenges present in multimodal machine learning and evaluated the state-of-the-art therein. By imposing an established taxonomy that captures the methodological variety of the works conducted in the medical specialty of radiology, we hope to foster multimodal research through guiding researchers whose research involves multimodal data. In the following, we identify research gaps that could be addressed by future research.

Considering the image modalities, there is a pronounced imbalance towards radiographs. While it can be explained by the availability of open hospital-scale datasets, the release of multimodal datasets comprising other medical imaging (e.g. CT or MRI) could lead to works that complement the current research in order to better reflect clinical routine. Another shortcoming in current open datasets is the lack of text data that are not image-centric as it is being the case with radiology reports. We hypothesize, that such non image-centric data could add more complementary information to the multimodal analyses. The scarcity of suitable open accessible data might also cause the shortage of works dealing with medical imaging and clinical time series.

We expect that future research will not restrict itself to bimodal combinations. Instead, leveraging image, tabular, and text data could lead to a holistic view of a patient that is

able to facilitate an improvement in patient outcome through personalized medicine. Furthermore, this calls for research towards learning patient-level representation from multimodal data, accounting for missing modalities. Lastly, considering the evaluation of fusion approaches and downstream tasks used in the works in representation learning, we encourage researchers to report unimodal (self-supervised) baselines to see whether multimodal approaches turn out to be more performant and label-efficient with possibly sparse and noisy data inherent to real-world applications.

V. CONCLUSION

Motivated by the fact that radiologists incorporate modalities other than medical imaging into their decision making, multimodal deep learning is an area of research that is gaining increasing attention within radiology. In this review, we surveyed multimodal deep learning and structured the works into two bimodal combinations: imaging + unstructured non-imaging data and imaging + structured non-imaging data. Within each combination, we present the works in a common taxonomy that encompasses datasets along with the technical challenges of fusion, representation learning, and translation. By means of the applied taxonomy, we introduce researchers entering the field to the methodological diversity and the state-of-the-art in the technical challenges inherent in multimodal machine learning. Furthermore, we highlight existing research gaps and unsolved challenges to pave the way for future research.

ACKNOWLEDGEMENT

We acknowledge the REACT-EU project KITE (Plattform für KI-Translation Essen) and the Helmut Horten Stiftung (University of Zurich).

REFERENCES

- [1] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, D. F. Kallmes, The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload, *Academic radiology* 22 (9) (2015) 1191–1198.
- [2] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, *Annual review of biomedical engineering* 19 (2017) 221–248.
- [3] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Rumviboonsuk, L. M. Vardoulakis, A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [4] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (brats), *IEEE transactions on medical imaging* 34 (10) (2014) 1993–2024.
- [5] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence* 41 (2) (2018) 423–443.
- [6] 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 digital medicine* 3 (1) (2020) 1–9.
- [7] 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: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.
- [8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [9] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (8) (1997) 1735–1780.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).
- [12] 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)*, IEEE, 2020, pp. 1999–2004.
- [13] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, *arXiv preprint arXiv:1908.07490* (2019).
- [14] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, *arXiv preprint arXiv:1908.03557* (2019).
- [15] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Learning universal image-text representations (2019).
- [16] Z. Huang, Z. Zeng, B. Liu, D. Fu, J. Fu, Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, *arXiv preprint arXiv:2004.00849* (2020).
- [17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [18] E. Alsentner, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.03323* (2019).
- [19] T.-M. H. Hsu, W.-H. Weng, W. Boag, M. McDermott, P. Szolovits, Unsupervised multimodal representation learning across medical images and reports, *arXiv preprint arXiv:1811.08615* (2018).
- [20] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [21] 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, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 529–539.
- [22] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, C. P. Langlotz, Contrastive learning of medical visual representations from paired images and text, *arXiv preprint arXiv:2010.00747* (2020).
- [23] M. Endo, R. Krishnan, V. Krishnan, A. Y. Ng, P. Rajpurkar, Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model, in: *Machine Learning for Health*, PMLR, 2021, pp. 209–219.
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, *arXiv preprint arXiv:2103.00020* (2021).
- [25] 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: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3942–3951.
- [26] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpankaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 590–597.
- [27] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.
- [28] 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, et al., Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia, *Radiology: Artificial Intelligence* 1 (1) (2019) e180041.
- [29] P. Müller, G. Kaissis, C. Zou, D. Rückert, Joint learning of localized representations from medical images and reports, *arXiv preprint arXiv:2112.02889* (2021).
- [30] X. Wang, Z. Xu, L. Tam, D. Yang, D. Xu, Self-supervised image-text pre-training with mixed data in chest x-rays, *arXiv preprint arXiv:2103.16022* (2021).

- [31] W. Boag, T.-M. H. Hsu, M. McDermott, G. Berner, E. Alesentzer, P. Szolovits, Baselines for chest x-ray report generation, in: *Machine Learning for Health Workshop*, PMLR, 2020, pp. 126–140.
- [32] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest x-ray report generation, in: *Machine Learning for Healthcare Conference*, PMLR, 2019, pp. 249–269.
- [33] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, D. Xu, When radiology report generation meets knowledge graph, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, pp. 12910–12917.
- [34] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven transformer, *arXiv preprint arXiv:2010.16056* (2020).
- [35] F. Liu, C. You, X. Wu, S. Ge, X. Sun, et al., Auto-encoding knowledge graph for unsupervised medical report generation, *Advances in Neural Information Processing Systems* 34 (2021).
- [36] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 6666–6673.
- [37] T. Syeda-Mahmood, K. C. Wong, Y. Gur, J. T. Wu, A. Jadhav, S. Kashyap, A. Karargyris, A. Pillai, A. Sharma, A. B. Syed, et al., Chest x-ray report generation through fine-grained label learning, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 561–571.
- [38] I. Rodin, I. Fedulova, A. Shelmanov, D. V. Dylov, Multitask and multimodal neural network model for interpretable analysis of x-ray images, in: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2019, pp. 1601–1604.
- [39] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019., in: *CLEF (Working Notes)*, 2019.
- [40] X. Yang, N. Gireesh, E. Xing, P. Xie, Xraygan: Consistency-preserving generation of x-ray images from radiology reports, *arXiv preprint arXiv:2006.10552* (2020).
- [41] G. Spinks, M.-F. Moens, Evaluating textual representations through image generation, in: *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, 2018, pp. 30–39.
- [42] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, S. Horng, Mimic-cxr database, *PhysioNet* <https://doi.org/10.13026/C2JT1Q> (2019).
- [43] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* 6 (1) (2019) 1–8.
- [44] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, *circulation* 101 (23) (2000) e215–e220.
- [45] A. E. 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 preprint arXiv:1901.07042* (2019).
- [46] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, Z. Lu, Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, *AMIA Summits on Translational Science Proceedings* 2018 (2018) 188.
- [47] 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, *Journal of the American Medical Informatics Association* 23 (2) (2016) 304–310.
- [48] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, *Medical image analysis* 66 (2020) 101797.
- [49] B. Ionescu, H. Müller, M. Villegas, A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, Y. D. Cid, V. Liauchuk, V. Kovalev, S. A. Hasan, et al., Overview of imageclef 2018: Challenges, datasets and evaluation, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2018, pp. 309–334.
- [50] A. G. S. De Herrera, S. Bromuri, R. Schaer, H. Müller, Overview of the medical tasks in imageclef 2016, *CLEF Working Notes*. Evora, Portugal (2016).
- [51] 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, 2018, pp. 180–189.
- [52] S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, H. Hajishirzi, Mediat: A dataset of medical images, captions, and textual references, *arXiv preprint arXiv:2010.06000* (2020).
- [53] S. Feng, D. Azzollini, J. S. Kim, C.-K. Jin, S. P. Gordon, J. Yeoh, E. Kim, M. Han, A. Lee, A. Patel, et al., Curation of the candid-ptx dataset with free-text reports, *Radiology: Artificial Intelligence* 3 (6) (2021) e210136.
- [54] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, *Molecular psychiatry* 19 (6) (2014) 659–667.
- [55] A. Di Martino, D. O’connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggato, S. Bernaerts, et al., Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii, *Scientific data* 4 (1) (2017) 1–15.
- [56] T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultonwood, R. Collins, M. C. Conroy, et al., The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions, *Nature communications* 11 (1) (2020) 1–12.
- [57] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. Amrhein, M. K. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, et al., Radfusion: Benchmarking performance and fairness for multi-modal pulmonary embolism detection from ct and enr (2021).
- [58] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., The cancer imaging archive (tcia): maintaining and operating a public information repository, *Journal of digital imaging* 26 (6) (2013) 1045–1057.
- [59] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, *Scientific data* 3 (1) (2016) 1–9.
- [60] 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, *Scientific reports* 10 (1) (2020) 1–9.
- [61] S.-C. Huang, T. Kothari, I. Banerjee, C. Chute, R. L. Ball, N. Borus, A. Huang, B. N. Patel, P. Rajpurkar, J. Irvin, et al., Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging, *NPJ digital medicine* 3 (1) (2020) 1–9.
- [62] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the royal statistical society: series B (statistical methodology)* 67 (2) (2005) 301–320.
- [63] 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: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3294–3303.
- [64] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.
- [65] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al., Big self-supervised models advance medical image classification, *arXiv preprint arXiv:2101.05224* (2021).
- [66] Y. N. T. Vu, R. Wang, N. Balachandar, C. Liu, A. Y. Ng, P. Rajpurkar, Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation, *arXiv preprint arXiv:2102.10663* (2021).
- [67] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [68] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, *arXiv preprint arXiv:2003.04297* (2020).
- [69] A. B. Qasim, I. Ezhov, S. Shit, O. Schoppe, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Lipkova, H. Li, B. Menze, Red-gan: Attacking class imbalance via conditioned generation. yet another medical imaging perspective., in: *Medical Imaging with Deep Learning*, PMLR, 2020, pp. 655–668.
- [70] K. Oh, Y.-C. Chung, K. W. Kim, W.-S. Kim, I.-S. Oh, Classification and visualization of alzheimer’s disease using volumetric convolutional neural network and transfer learning, *Scientific Reports* 9 (1) (2019) 1–16.

- [71] P. Chang, J. Grinband, B. Weinberg, M. Bardis, M. Khy, G. Cadena, M.-Y. Su, S. Cha, C. Filippi, D. Bota, et al., Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas, *American Journal of Neuroradiology* 39 (7) (2018) 1201–1207.
- [72] H. Qu, M. Zhou, Z. Yan, H. Wang, V. K. Rustgi, S. Zhang, O. Gevaert, D. N. Metaxas, Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning, *NPJ precision oncology* 5 (1) (2021) 1–11.
- [73] D. Štern, C. Payer, M. Urschler, Automated age estimation from mri volumes of the hand, *Medical image analysis* 58 (2019) 101538.
- [74] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, et al., The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives, *Frontiers in Neuroinformatics* 7 (2013).
- [75] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, A. Eshaghi, T. Toni, et al., The alzheimer’s disease prediction of longitudinal evolution (tadpole) challenge: Results after 1 year follow-up, *arXiv preprint arXiv:2002.03419* (2020).
- [76] S. O. Arnk, T. Pfister, Tabnet: Attentive interpretable tabular learning, *arXiv* (2020).
- [77] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, *Information Fusion* 81 (2022) 84–90.
- [78] A. Kadra, M. Lindauer, F. Hutter, J. Grabocka, Regularization is all you need: Simple neural nets can excel on tabular data, *arXiv preprint arXiv:2106.11189* (2021).
- [79] A. Leslie, A. Jones, P. Goddard, The influence of clinical information on the reporting of ct by radiologists., *The British journal of radiology* 73 (874) (2000) 1052–1055.
- [80] M. D. Cohen, Accuracy of information on imaging requisitions: does it matter?, *Journal of the American College of Radiology* 4 (9) (2007) 617–621.
- [81] W. W. Boonn, C. P. Langlotz, Radiologist use of and perceived need for patient data access, *Journal of digital imaging* 22 (4) (2009) 357–362.