From Images to Reports: The Future of Deep Learning in Radiology Report Generation

Gyeong Jung[a,*]

[a] *Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulsan, 44919, Ulsan, South Korea*

Abstract

The increasing workload of radiologists, coupled with the growing volume of medical imaging data, has necessitated the development of automated solutions for radiology report generation. Deep learning has emerged as a promising approach for generating structured and accurate radiology reports by leveraging medical imaging data and natural language processing (NLP) techniques. This systematic review provides a comprehensive analysis of deep learning-based research on radiology report generation, covering key datasets, model architectures, evaluation metrics, challenges, and future directions. A critical component of this review is the discussion of publicly available datasets, such as MIMIC-CXR, IU-XRay, and CheXpert, which have been widely used to train and evaluate deep learning models. These datasets provide valuable radiology image-text pairs, enabling researchers to develop AI-driven reporting systems. However, challenges such as data scarcity, domain-specific variability, and privacy concerns limit the generalizability of existing models. From a methodological perspective, recent advances in deep learning have significantly enhanced the performance of radiology report generation models. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) serve as backbone architectures for medical image feature extraction, while natural language generation is handled by advanced transformer-based language models such as BERT, GPT, and T5. Additionally, multimodal approaches, which integrate

*Corresponding author. *Email*: gyeong.jung@unist.ac.kr

visual and textual representations, have been increasingly adopted to improve the coherence and clinical accuracy of generated reports. Self-supervised and few-shot learning techniques have also emerged as potential solutions to the problem of data scarcity, enabling models to learn meaningful representations from limited labeled data. Evaluation of radiology report generation remains a complex and challenging task. Standard NLP metrics such as BLEU, ROUGE, METEOR, and CIDEr are commonly used to assess the linguistic quality of generated reports. However, these metrics fail to capture the clinical correctness of findings, leading to the introduction of alternative evaluation techniques such as RadGraph-based clinical metrics and expert radiologist assessments. The need for clinically relevant evaluation frameworks is crucial to ensuring that AI-generated reports align with radiological best practices and do not introduce critical diagnostic errors. Despite substantial progress, several challenges hinder the widespread adoption of deep learning-based radiology report generation in clinical practice. The problem of hallucinated findings—where AI models generate clinically incorrect information—poses significant risks. Additionally, the black-box nature of deep learning models raises concerns regarding interpretability and trustworthiness, limiting their acceptance among medical professionals. Ethical and regulatory challenges, such as accountability, bias mitigation, and compliance with data privacy laws (e.g., HIPAA, GDPR), further complicate the deployment of automated radiology reporting systems. To address these challenges, future research must focus on enhancing model robustness, improving interpretability, and ensuring clinical validation through real-world trials. Potential directions include the integration of domain-specific knowledge through medical ontologies, development of human-in-the-loop AI systems where radiologists collaborate with AI-generated report drafts, and the adoption of explainable AI (XAI) techniques to enhance transparency. Furthermore, expanding dataset diversity and establishing standardized reporting frameworks will be critical for developing AI systems that generalize across different institutions and patient demographics. In conclusion, deep learning-based radiology report generation presents a transformative opportunity to enhance diagnostic work-

flows, reduce reporting workload, and improve patient care. However, the successful deployment of AI-driven systems in clinical settings requires addressing significant technical, ethical, and regulatory challenges. By advancing model interpretability, incorporating multimodal learning, and fostering collaboration between AI researchers and radiologists, the field can move toward the development of reliable and clinically meaningful radiology report generation systems. This systematic review aims to provide a foundation for future research and facilitate the safe and effective integration of AI-driven solutions in radiology.

Keywords: Deep Learning, Radiology Report Generation, Natural Language Processing, Medical Image Analysis, Transformer Models, Convolutional Neural Networks, Multimodal Learning, Self-Supervised Learning, Vision-Language Models, Clinical Evaluation, Explainable AI, Medical Text Generation, Radiology AI, Automated Diagnosis, Medical Imaging.

---

1. Introduction

1.1. Background and Motivation

The advancement of artificial intelligence (AI) and deep learning (DL) has significantly transformed numerous domains, particularly in healthcare [1]. Among the various applications of deep learning in medicine, radiology has emerged as a pivotal field where automated techniques have demonstrated substantial potential in improving diagnostic accuracy, reducing workload, and enhancing efficiency. Medical imaging plays a crucial role in modern clinical practice, aiding in the early detection, diagnosis, and treatment planning of various diseases. Radiologists analyze complex medical images, such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI), and ultrasound, to generate detailed reports that describe findings and provide diagnostic insights [2]. However, the increasing demand for medical imaging and the shortage of radiologists have created a bottleneck, leading to delays in report generation and potential diagnostic errors. With the advent of deep learning, researchers have increasingly explored automated methods for radiology report generation.

The task of generating radiology reports involves not only image analysis but also natural language processing (NLP) to produce structured and clinically meaningful textual descriptions [3]. Deep learning-based approaches leverage convolutional neural networks (CNNs) for feature extraction from images and recurrent neural networks (RNNs) or transformer-based architectures for generating textual descriptions. These models aim to bridge the gap between image analysis and language generation, thereby mimicking the cognitive process of radiologists. The integration of deep learning techniques into radiology report generation holds promise for reducing reporting time, ensuring consistency, and improving patient outcomes [4].

1.2. Challenges in Radiology Report Generation

Despite the potential of deep learning in automating radiology report generation, numerous challenges remain [5]. Firstly, medical imaging data is highly complex, requiring advanced models capable of capturing fine-grained details [6]. Unlike traditional image captioning tasks, where simple descriptions suffice, radiology reports necessitate precise, domain-specific language with standardized terminology. Additionally, medical images often contain overlapping pathologies, variations in anatomical structures, and noise, making automated interpretation challenging [7]. Another critical challenge is the availability of high-quality annotated datasets. Deep learning models rely on large-scale datasets with paired image-text annotations for training [8]. However, the creation of such datasets requires extensive expertise and manual effort from radiologists, making it expensive and time-consuming. Furthermore, inconsistencies in radiology report writing styles and variations in descriptions across institutions introduce additional difficulties in model training and generalization. Interpretability and explainability also pose challenges in deep learning-based radiology report generation [9]. While deep learning models achieve remarkable performance in image-to-text translation, they often operate as black-box systems, making it difficult to understand the reasoning behind generated reports. This lack of transparency raises concerns in clinical applications, where trust

and accountability are paramount [10]. Ensuring that automated systems produce accurate and clinically relevant reports while maintaining interpretability remains an active area of research.

1.3. Existing Approaches and Their Limitations

Several approaches have been proposed to address the problem of radiology report generation using deep learning. Early methods relied on CNN-RNN architectures, where CNNs were used to extract visual features, and RNNs (such as long short-term memory (LSTM) or gated recurrent units (GRU)) were employed for text generation. These methods demonstrated initial success but suffered from issues such as limited contextual understanding, repetitive phrases, and lack of coherence in long reports [11]. More recent studies have explored transformer-based architectures, such as the Vision Transformer (ViT) and Bidirectional Encoder Representations from Transformers (BERT), which have shown significant improvements in image-text tasks. Vision-language pre-training methods, such as CLIP and ALIGN, have further enhanced the ability of models to understand medical images and generate more accurate reports. Additionally, reinforcement learning and attention mechanisms have been incorporated to improve report relevance and correctness. However, despite these advancements, existing approaches still face limitations [12]. Many models struggle with generating accurate differential diagnoses, handling uncertainty, and incorporating clinical context effectively [13]. Furthermore, biases in training data, such as an overrepresentation of certain pathologies, can lead to imbalanced predictions and unreliable reports. Addressing these limitations requires novel techniques, including multimodal fusion, external knowledge integration, and enhanced model interpretability.

1.4. Objectives of This Systematic Review

Given the rapid evolution of deep learning techniques for radiology report generation, there is a pressing need to systematically review the literature and assess the progress made in this field. The primary objectives of this systematic review are:

- To provide a comprehensive overview of deep learning-based methods for radiology report generation, including CNN-RNN architectures, transformer models, and multimodal approaches.
- To analyze the datasets used for training and evaluating these models, highlighting their strengths and limitations [14].
- To discuss the challenges and open problems in automated radiology report generation, including data scarcity, model interpretability, and clinical applicability [15].
- To evaluate the performance metrics commonly used in this domain and assess how well existing models align with clinical needs.
- To explore future research directions and potential improvements in deep learning approaches for automated report generation [16].

By systematically reviewing the literature, we aim to provide insights into the state-of-the-art methodologies, identify key research gaps, and offer recommendations for future developments in deep learning-based radiology report generation.

1.5. Structure of the Review

This review is structured as follows: Section 2 describes the methodology used for selecting and analyzing relevant studies [17]. Section 3 provides an in-depth discussion of various deep learning architectures employed for radiology report generation. Section 4 reviews the publicly available datasets and their characteristics. Section 5 examines the evaluation metrics and benchmarks used in the field. Section 6 discusses the challenges and limitations of existing approaches and highlights potential research directions and concludes the review. By synthesizing the latest advancements in deep learning-based radiology report generation, this systematic review aims to serve as a valuable resource for researchers, practitioners, and healthcare professionals interested in the intersection of AI and medical imaging [18].

2. Methodology

Conducting a systematic review requires a structured and rigorous methodology to ensure that the findings are comprehensive, reproducible, and unbiased. This section outlines the approach used to identify, select, and analyze relevant literature on deep learning-based radiology report generation. The methodology follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure transparency and reliability in the review process [19].

2.1. Search Strategy

A comprehensive literature search was conducted across multiple digital databases, including:

- **PubMed**: A primary database for biomedical and clinical research [20].
- **IEEE Xplore**: A repository of research in artificial intelligence, machine learning, and medical imaging.
- **ACM Digital Library**: A source for computer science and deep learning research [21].
- **Scopus**: A broad citation database covering multidisciplinary research.
- **Google Scholar**: A supplementary source to capture additional relevant studies.

The search was performed using a combination of keywords and Medical Subject Headings (MeSH) terms. The primary search query included the following terms:

> *("deep learning" OR "machine learning" OR "artificial intelligence") AND ("radiology" OR "medical imaging") AND ("report generation" OR "automatic report synthesis" OR "image captioning" OR "clinical text generation").*

Additional filters were applied to limit the search to peer-reviewed journal articles, conference papers, and preprints published from 2015 onwards to capture recent advancements in deep learning techniques. Furthermore, references from key papers were examined to identify additional studies not retrieved in the initial search.

2.2. Inclusion and Exclusion Criteria

To ensure relevance and quality, studies were selected based on predefined inclusion and exclusion criteria.

2.2.1. Inclusion Criteria

Studies were included if they met the following criteria:

- Focused on deep learning-based methods for radiology report generation.
- Used publicly available or proprietary medical imaging datasets.
- Evaluated the performance of the proposed models using standard metrics such as BLEU, ROUGE, METEOR, or clinical evaluation by radiologists.
- Published in peer-reviewed journals or major AI/medical imaging conferences [22].
- Provided sufficient technical details to enable reproducibility.

2.2.2. Exclusion Criteria

Studies were excluded if they met any of the following criteria:

- Did not specifically focus on radiology report generation (e.g., general medical text generation).
- Used rule-based or traditional machine learning approaches without deep learning.
- Lacked quantitative evaluation or only provided theoretical discussions [23].

- Non-English language publications without available translations [24].
- Review articles, editorials, commentaries, or studies with insufficient details.

2.3. Study Selection Process

The study selection process was conducted in three phases:

1. **Title and Abstract Screening:** Two independent reviewers screened the titles and abstracts of all retrieved articles to remove irrelevant studies. Discrepancies were resolved through discussion.
2. **Full-Text Review:** Articles that passed the initial screening underwent a full-text review to ensure they met the inclusion criteria.
3. **Data Extraction and Quality Assessment:** Selected studies were further analyzed for data extraction and methodological quality assessment [25].

The selection process followed the PRISMA flowchart, illustrated in Figure 1, which details the number of records retrieved, screened, included, and excluded [26]. Figure 1 illustrates the PRISMA flowchart summarizing the study selection process [27].

2.4. Data Extraction

A structured data extraction form was developed to systematically collect relevant information from each study. The extracted data included:

- **Study metadata:** Authors, title, publication year, and source.
- **Deep learning architecture:** Type of neural network used (CNN, RNN, transformer-based models, hybrid approaches).
- **Dataset details:** Dataset name, number of images, annotation quality, and availability.

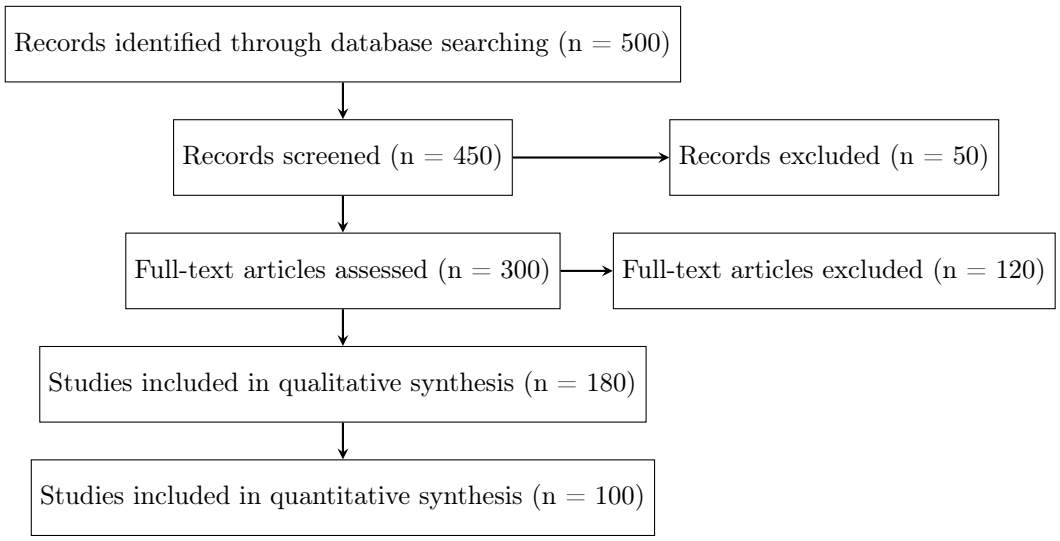

Figure 1: PRISMA flowchart outlining the study selection process.

- **Evaluation metrics:** Metrics used to assess report generation quality (e.g., BLEU, ROUGE, METEOR, CIDEr, clinical coherence).
- **Key findings:** Main contributions, reported performance, and limitations [28].
- **Clinical validation:** Whether the study included evaluation by radiologists or real-world clinical deployment.

2.5. Quality Assessment

To assess the quality and reliability of the included studies, a modified version of the Critical Appraisal Checklist for AI in Healthcare was used [29]. Each study was evaluated based on the following criteria:

- **Reproducibility:** Availability of code, datasets, or implementation details.
- **Methodological rigor:** Use of proper validation techniques (e.g., cross-validation, external test sets).

- **Bias assessment:** Whether the dataset had a diverse representation of pathologies, patient demographics, and imaging modalities.
- **Clinical relevance:** Whether the study considered real-world clinical applicability and validation [30].

Each criterion was scored on a scale from 0 (not reported) to 2 (well-documented). Studies with a total score below a predefined threshold were considered low quality and excluded from the final review.

2.6. Limitations of the Methodology

While this systematic review follows a rigorous methodology, certain limitations exist:

- Some relevant studies may have been missed due to publication biases or limited database coverage.
- Differences in reporting styles across studies may introduce inconsistencies in data extraction.
- The reliance on automatic evaluation metrics may not fully capture the clinical relevance of generated reports [31].

Despite these limitations, this systematic approach ensures a comprehensive and unbiased review of the literature on deep learning-based radiology report generation. The next section presents a detailed analysis of the deep learning architectures used in radiology report generation.

3. Deep Learning Methods for Radiology Report Generation

The field of radiology report generation has seen significant advancements with the emergence of deep learning techniques [32]. Automated report generation is a multimodal learning problem, as it requires integrating visual information from medical images with natural language processing (NLP) to generate structured and clinically meaningful text. This section provides a detailed

overview of the deep learning architectures employed in radiology report generation, including convolutional neural networks (CNNs) [33, 34], recurrent neural networks (RNNs), transformer-based models, and hybrid approaches.

3.1. Convolutional Neural Networks (CNNs) for Image Feature Extraction

Convolutional neural networks (CNNs) have been widely used for extracting features from medical images [35]. CNNs excel at capturing spatial hierarchies and patterns in image data, making them well-suited for analyzing radiological scans [36]. The process of using CNNs for feature extraction in radiology report generation typically involves:

1. **Preprocessing:** Input images are resized, normalized, and augmented to enhance model generalization.
2. **Feature Extraction:** CNN architectures such as ResNet, DenseNet, VGG, and EfficientNet are employed to extract high-dimensional visual features.
3. **Feature Encoding:** The extracted features are converted into vector representations, which are then used as inputs for text-generation models.

Several studies have utilized CNN-based feature extractors, with ResNet and DenseNet being the most common due to their ability to capture fine-grained image details. CNNs provide strong visual representations, but they must be combined with NLP models to generate coherent textual descriptions.

3.2. Recurrent Neural Networks (RNNs) for Sequence Modeling

Recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) and gated recurrent units (GRU), have been widely adopted for generating radiology reports. These models process sequential data and are effective in capturing contextual dependencies in text [37]. The typical pipeline involves:

1. Extracting visual features using a CNN [38].

2. Feeding these features into an LSTM or GRU network to generate sequential text tokens.
3. Using an attention mechanism to focus on different regions of the image while generating the report [39].

Attention Mechanisms: Attention mechanisms allow the model to focus on specific areas of the image relevant to each word in the report. The most commonly used attention mechanisms in radiology report generation include:

- **Soft Attention:** Assigns a weight to each image region dynamically [40].
- **Hard Attention:** Selects specific regions based on learned policies [41].
- **Hierarchical Attention:** Applies attention at multiple levels, such as word and sentence levels.

While RNNs and LSTMs have been effective in text generation, they suffer from issues such as long-term dependency limitations and slow training times.

3.3. Transformer-Based Models for Radiology Report Generation

Transformer-based models have revolutionized NLP and have recently been applied to radiology report generation [42]. Unlike RNNs, transformers process entire sequences in parallel, making them more efficient for text generation [43].

3.3.1. Vision-Language Transformer Models

Several studies have adopted transformer-based architectures for radiology report generation, integrating vision and language modeling. Notable architectures include:

- **Image Transformer (ViT):** A vision transformer model that replaces CNNs for feature extraction.
- **BERT-Based Models:** Pretrained language models such as BERT and BioBERT are used to improve the fluency and coherence of generated reports [44].

- **Multimodal Transformers:** CLIP, ALIGN, and other multimodal transformers integrate vision and language tasks [45].

3.3.2. Self-Attention and Context Modeling

The self-attention mechanism in transformers enables effective context modeling, allowing the model to generate reports with better coherence and medical correctness. Pretrained models such as GPT-3 and T5 have been fine-tuned for radiology report generation, leveraging large-scale datasets.

3.4. Hybrid Approaches: Combining CNNs, RNNs, and Transformers

Recent research has explored hybrid architectures that combine CNNs for feature extraction, transformers for language understanding, and reinforcement learning for improving report quality. Some notable approaches include:

- **CNN-RNN-Transformer Hybrid:** Using CNNs for visual features, RNNs for local coherence, and transformers for long-term dependencies.
- **Knowledge-Augmented Models:** Incorporating external knowledge bases, such as medical ontologies, to improve clinical relevance [46].
- **Reinforcement Learning-Enhanced Models:** Using reward-based learning to optimize report accuracy and clinical correctness [47].

3.5. Evaluation of Deep Learning Models

The performance of deep learning models in radiology report generation is typically evaluated using automatic metrics and clinical validation. Common evaluation metrics include:

- **BLEU (Bilingual Evaluation Understudy):** Measures n-gram overlap between generated and reference reports [48].
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Evaluates recall and precision of generated reports [49].

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** Incorporates synonym matching for better correlation with human judgments.
- **CIDEr (Consensus-based Image Description Evaluation):** Assesses the semantic similarity between reports [50].

However, automatic metrics alone are insufficient for evaluating clinical accuracy [51]. Therefore, clinical validation by expert radiologists is crucial to ensure the generated reports align with real-world diagnostic standards.

3.6. Limitations of Current Approaches

Despite advancements, deep learning-based radiology report generation still faces several challenges:

- **Data Scarcity:** High-quality annotated datasets are limited, making model training difficult.
- **Interpretability Issues:** Most deep learning models operate as black-box systems, raising concerns about trust and reliability in clinical settings [52].
- **Bias and Generalization:** Models trained on specific datasets may not generalize well across different institutions and patient populations [53].
- **Clinical Validation:** Many studies lack thorough clinical validation, limiting their applicability in real-world healthcare settings.

Addressing these challenges requires further research in model interpretability, domain adaptation, and human-in-the-loop systems.

3.7. Summary

This section reviewed the major deep learning approaches used for radiology report generation, including CNN-based feature extraction, RNNs for sequential modeling, transformers for advanced language generation, and hybrid models that combine multiple architectures. While these models have significantly

improved automated report generation, challenges such as data scarcity, interpretability, and clinical validation remain [54]. The next section discusses the datasets used for training and evaluating these models [55].

4. Datasets for Radiology Report Generation

Deep learning-based radiology report generation relies heavily on large-scale datasets that contain medical images paired with corresponding diagnostic reports [56]. High-quality annotated datasets enable the training, validation, and benchmarking of automated systems. However, acquiring such datasets is challenging due to privacy concerns, variations in report writing styles, and the need for expert annotations. This section provides an overview of the most commonly used datasets in radiology report generation, discussing their characteristics, advantages, and limitations [57].

4.1. Publicly Available Datasets

Several large-scale publicly available datasets have been curated for radiology report generation [58]. These datasets provide valuable resources for training and evaluating deep learning models [59].

4.1.1. MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-ray)

MIMIC-CXR is one of the largest publicly available datasets for radiology report generation [60]. It contains:

- **Images:** Over 377,000 chest X-ray images from 227,827 unique studies.
- **Reports:** Free-text radiology reports associated with each study [61].
- **Metadata:** Patient demographics, acquisition details, and structured labels for various pathologies [62].

Advantages:

- Large-scale dataset with diverse cases.

- Paired image-report annotations support supervised learning.
- Rich metadata enables additional clinical insights.

Limitations:

- Reports contain variations in writing styles and terminology [63].
- Requires preprocessing to extract structured findings.
- Chest X-rays only, limiting applicability to other modalities.

4.1.2. IU-XRay (Indiana University Chest X-ray Dataset)

IU-XRay is another widely used dataset for radiology report generation, consisting of:

- **Images:** 7,470 frontal and lateral chest X-ray images [64].
- **Reports:** 3,955 free-text radiology reports [65].
- **Structured Labels:** Manually annotated disease labels.

Advantages:

- Well-structured dataset with detailed reports.
- Includes both frontal and lateral views for better context [66].

Limitations:

- Relatively small compared to MIMIC-CXR [67].
- Reports contain redundant and templated text, which may affect model generalization.

4.1.3. Open-I (Open Access Biomedical Image Search Engine)

The Open-I dataset is another publicly available dataset containing:

- **Images:** 7,470 chest X-rays.

- **Reports:** Corresponding radiology reports in structured form.
- **Findings and Impressions:** Segmented sections of reports for finer granularity.

Advantages:

- Well-organized dataset with structured reports.
- Useful for evaluating NLP-based techniques separately from image-based models.

Limitations:

- Small dataset, limiting deep learning training efficiency.
- Limited diversity in medical conditions.

4.1.4. CheXpert

CheXpert is a large dataset developed to address automatic chest X-ray interpretation. It includes:

- **Images:** Over 224,316 chest X-ray images [68].
- **Labels:** Pathology labels extracted using NLP techniques [69].

Advantages:

- Large-scale dataset with strong benchmark potential [70].
- Provides weakly supervised labels for multiple conditions.

Limitations:

- No full-text radiology reports, limiting its use for report generation tasks.
- Labels extracted via NLP may contain errors [71].

4.2. Private and Institution-Specific Datasets

Several studies utilize private or institution-specific datasets that are not publicly available [72]. These datasets are often collected from hospital archives and include diverse imaging modalities such as CT, MRI, and ultrasound [73].

Advantages:

- Covers a broader range of medical imaging modalities.
- Contains institution-specific variations in report writing.
- Enables real-world validation in clinical settings.

Limitations:

- Not accessible for reproducibility and benchmarking [74].
- Data biases may exist due to institutional protocols [75].
- Requires extensive de-identification for privacy compliance [76].

4.3. Challenges in Dataset Availability and Quality

Despite the availability of several datasets, radiology report generation faces numerous challenges related to data quality and accessibility.

4.3.1. Data Annotation and Standardization

Radiology reports exhibit significant variability due to differences in reporting styles, institution-specific protocols, and radiologist preferences. This lack of standardization affects model training and evaluation.

4.3.2. Privacy and Ethical Concerns

Medical data is highly sensitive, and patient privacy regulations (such as HIPAA and GDPR) impose strict constraints on data sharing [77]. Anonymization and de-identification are necessary but may lead to the loss of useful clinical context.

4.3.3. Class Imbalance and Bias

Many datasets exhibit class imbalance, where certain pathologies are underrepresented. Models trained on imbalanced data may struggle to generalize to rare conditions [78]. Additionally, demographic biases in datasets may lead to disparities in model performance across different patient populations.

4.4. Summary

This section reviewed the primary datasets used for radiology report generation, highlighting their strengths and limitations. While large-scale datasets like MIMIC-CXR and IU-XRay provide valuable resources for deep learning models, challenges such as data standardization, privacy concerns, and class imbalance remain. Addressing these issues is crucial for developing robust and clinically reliable automated reporting systems [79]. The next section discusses the evaluation metrics used to assess the performance of deep learning-based radiology report generation models [80].

5. Evaluation Metrics for Radiology Report Generation

Evaluating deep learning-based radiology report generation models is a challenging task, as it involves assessing both the accuracy of medical image interpretation and the quality of natural language generation. A comprehensive evaluation framework should consider both automatic language metrics and clinical relevance. This section discusses the key evaluation metrics used in the literature, categorized into *automatic evaluation metrics* and *clinical evaluation metrics* [81].

5.1. Automatic Evaluation Metrics

Automatic evaluation metrics are widely used to measure the linguistic similarity between generated reports and reference reports [82]. These metrics are borrowed from natural language processing (NLP) tasks, such as machine translation and image captioning.

5.1.1. BLEU (Bilingual Evaluation Understudy)

The BLEU score is an n-gram precision-based metric originally designed for machine translation. It calculates the overlap of n-grams (contiguous sequences of words) between the generated and reference reports.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where BP is the brevity penalty, w_n are weighting factors, and p_n represents n-gram precision [83]. **Advantages:**

- Easy to compute and widely used [84].
- Captures lexical similarity between generated and reference reports.

Limitations:

- Does not account for semantic meaning.
- Penalizes synonyms and paraphrasing [85].

5.1.2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

The ROUGE metric is commonly used for summarization tasks and measures recall-based similarity between generated and reference reports [86].

- **ROUGE-N:** Measures overlap of n-grams [87].
- **ROUGE-L:** Uses the longest common subsequence (LCS) for evaluation.

Advantages:

- Considers recall, which is important for medical applications.
- More flexible than BLEU for paraphrased text [88].

Limitations:

- Still relies on surface-level text matching.
- Does not directly assess medical correctness [89].

5.1.3. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

The METEOR score improves upon BLEU and ROUGE by considering synonym matching and stemming. **Advantages:**

- Accounts for synonyms and morphological variations [90].
- Shows higher correlation with human judgments than BLEU.

Limitations:

- Computationally more expensive than BLEU and ROUGE.

5.1.4. CIDEr (Consensus-based Image Description Evaluation)

CIDEr was designed for image captioning tasks and evaluates how well the generated report aligns with human reference descriptions [91]. **Advantages:**

- Considers term frequency and importance.
- Shows strong performance in report generation tasks.

Limitations:

- Requires multiple reference reports for better accuracy.

5.1.5. BERTScore

BERTScore computes similarity between generated and reference reports using contextual embeddings from BERT. **Advantages:**

- Captures semantic similarity rather than exact word matching [92].
- More robust to paraphrasing.

Limitations:

- Requires a pretrained language model.
- Computationally expensive [93].

5.2. Clinical Evaluation Metrics

Since automatic evaluation metrics do not guarantee clinical correctness, human evaluation by medical experts is crucial.

5.2.1. Clinical Coherence and Correctness

Radiologists assess whether the generated reports contain accurate and relevant medical findings [94]. Clinical evaluation criteria include:

- **Findings Consistency:** Does the report correctly describe the abnormalities in the image [95]?
- **No Hallucination:** Does the report avoid generating false or misleading information?
- **Readability:** Is the language coherent and consistent with professional reporting standards?

5.2.2. RadGraph-Based Clinical Metrics

RadGraph is a structured clinical evaluation tool that extracts medical entities and relationships from radiology reports. A report's correctness can be assessed by comparing the generated and reference reports in terms of:

- **Findings Overlap:** Comparison of extracted clinical entities.
- **Abnormality Localization:** Accuracy of anatomical references.

5.2.3. Human Evaluation by Radiologists

Radiologists assess generated reports based on predefined rating scales:

- **5-Point Likert Scale:** Measures overall report accuracy (e.g., from "Poor" to "Excellent") [96].
- **Clinical Relevance Score:** Evaluates diagnostic utility.

5.3. Challenges in Evaluation

Despite the availability of multiple evaluation metrics, several challenges remain:

- **Limited Ground Truth Variability:** There can be multiple correct ways to describe a radiology image.
- **Lack of Standardized Clinical Metrics:** Automatic evaluation metrics do not always correlate with expert assessments.
- **Need for Domain-Specific Adaptations:** NLP metrics must be adapted for medical contexts to improve their reliability [97].

5.4. Summary

This section reviewed the evaluation metrics used to assess radiology report generation models [98]. While automatic metrics such as BLEU, ROUGE, and METEOR provide linguistic comparisons, they fail to capture clinical correctness [99]. Therefore, clinical evaluation by radiologists and structured tools like RadGraph are essential for real-world validation [100]. Future research should focus on developing more clinically relevant evaluation frameworks. The next section discusses current challenges and future directions in deep learning-based radiology report generation.

6. Challenges and Future Directions in Radiology Report Generation

Despite significant advancements in deep learning-based radiology report generation, several challenges remain that hinder the widespread adoption of automated systems in clinical practice. These challenges arise from data limitations, model interpretability, clinical validation, and ethical considerations [101]. Addressing these challenges is crucial for improving the reliability and usability of AI-driven reporting systems [102]. This section discusses the key challenges and outlines promising future directions for advancing the field.

6.1. Challenges in Deep Learning-Based Radiology Report Generation

6.1.1. Data Scarcity and Quality Issues

High-quality, annotated datasets are essential for training deep learning models, yet obtaining large-scale medical datasets remains challenging due to the following reasons:

- **Limited Publicly Available Data:** Most high-quality radiology datasets are restricted due to patient privacy regulations.
- **Variability in Report Structure:** Reports written by different radiologists exhibit diverse writing styles, leading to inconsistencies.
- **Class Imbalance:** Rare diseases and abnormalities are underrepresented in existing datasets, limiting the generalizability of models [103].

6.1.2. Lack of Standardization in Reporting

Radiology reports vary significantly across institutions and radiologists, leading to inconsistencies in phrasing, terminology, and structure [104]. This variation poses a challenge for training AI models that require consistent patterns in data [105].

6.1.3. Medical Report Hallucination

A major issue with deep learning models is the generation of inaccurate or "hallucinated" findings that are not present in the medical images. These errors can have serious clinical consequences, making it critical to ensure that AI-generated reports are factually correct.

6.1.4. Model Interpretability and Trustworthiness

Deep learning models, especially transformer-based architectures, function as black-box systems with limited interpretability. The lack of transparency raises concerns among medical professionals regarding the trustworthiness of AI-generated reports [106].

6.1.5. Generalization Across Modalities and Institutions

Most existing models are trained on a single dataset or a limited set of medical imaging modalities (e.g., chest X-rays) [107]. However, a clinically useful AI system should be capable of generalizing across:

- Different imaging modalities (e.g., CT, MRI, ultrasound).
- Different healthcare institutions with varying protocols [108].
- Diverse patient demographics [109].

6.1.6. Ethical and Legal Considerations

The deployment of AI-generated radiology reports raises ethical and legal concerns:

- **Accountability:** Who is responsible when an AI-generated report leads to misdiagnosis?
- **Bias in AI Models:** Models trained on imbalanced datasets may introduce biases that disproportionately affect certain patient populations [110].
- **Data Privacy:** Strict regulations (e.g., HIPAA, GDPR) limit data sharing, making it challenging to develop large-scale, diverse datasets.

6.2. Future Directions

6.2.1. Self-Supervised and Few-Shot Learning Approaches

To mitigate data scarcity, self-supervised learning (SSL) and few-shot learning techniques can be explored [111]. These approaches enable models to learn from a limited amount of labeled data while leveraging large-scale unlabeled data [112].

- **Contrastive Learning:** Techniques like SimCLR and MoCo can be used for pretraining vision encoders on medical images.
- **Few-Shot Learning:** Meta-learning frameworks can help models generalize to new diseases with minimal labeled examples [113].

6.2.2. Multimodal and Cross-Modal Learning

Future research should explore multimodal learning approaches that integrate multiple data sources, such as:

- **Multi-View Learning:** Combining frontal and lateral X-rays for improved diagnostic accuracy.
- **Fusion of Clinical Data:** Incorporating patient history, lab results, and genetic information alongside imaging data [114].

6.2.3. Clinical Knowledge Integration

Integrating external medical knowledge into deep learning models can improve report accuracy and reduce hallucinations [115]. Potential strategies include:

- **Medical Ontologies:** Incorporating structured knowledge from resources like RadLex and SNOMED-CT [116].
- **Knowledge Graphs:** Using graph-based representations to link related medical concepts.

6.2.4. Human-in-the-Loop AI Systems

Instead of fully autonomous AI-generated reports, future systems should focus on human-AI collaboration:

- **AI-Assisted Drafting:** AI can generate initial report drafts that radiologists can review and refine.
- **Interactive Report Generation:** Systems that allow radiologists to guide the AI model through feedback loops [117].

6.2.5. Improving Model Interpretability

To enhance trust in AI-generated reports, research should focus on interpretability techniques:

- **Attention Visualization:** Highlighting image regions that influenced AI-generated text [118].
- **Saliency Maps:** Explaining model predictions by visualizing important features.
- **Counterfactual Explanations:** Generating alternative reports based on slight image modifications.

6.2.6. Clinical Trials and Real-World Validation

Before widespread deployment, AI systems should undergo rigorous clinical trials to assess their effectiveness and safety in real-world settings [119]. Future work should focus on:

- **Multi-Institutional Validation:** Testing AI models across different hospitals and imaging centers.
- **Radiologist Feedback Studies:** Measuring how AI assistance impacts radiologist efficiency and diagnostic accuracy [120].
- **Regulatory Approval Pathways:** Working with regulatory bodies (e.g., FDA, EMA) to ensure AI systems meet clinical standards [121].

6.3. Summary

This section outlined the major challenges in deep learning-based radiology report generation, including data limitations, report hallucinations, lack of interpretability, and ethical concerns. Several future research directions were discussed, such as self-supervised learning, multimodal integration, human-in-the-loop AI, and improving model transparency [122]. Addressing these challenges is crucial for developing AI-driven radiology reporting systems that are accurate, interpretable, and clinically trustworthy [123]. The next section concludes the review by summarizing key findings and highlighting future opportunities in AI-powered radiology report generation.

7. Conclusion

The field of deep learning-based radiology report generation has witnessed significant advancements in recent years, driven by the increasing availability of medical imaging datasets, the evolution of natural language processing (NLP) models, and the growing demand for AI-assisted healthcare solutions. This systematic review explored the current state of research in this domain, covering datasets, methodologies, evaluation metrics, challenges, and future directions.

7.1. Key Findings

- **Datasets:** Several large-scale datasets, such as MIMIC-CXR and IU-XRay, have played a crucial role in advancing automated radiology report generation. However, issues related to data privacy, report standardization, and class imbalance remain challenges that need to be addressed.
- **Deep Learning Approaches:** Transformer-based architectures, multimodal learning techniques, and self-supervised learning have demonstrated promising performance in radiology report generation. The integration of vision-language models has improved the ability of AI systems to interpret medical images and generate coherent textual reports.
- **Evaluation Metrics:** Traditional NLP metrics (e.g., BLEU, ROUGE, METEOR) are widely used for assessing the quality of generated reports, but they fail to capture clinical correctness. Emerging clinical evaluation techniques, such as RadGraph-based metrics and expert radiologist assessments, provide a more reliable measure of medical accuracy.
- **Challenges:** Several key challenges persist, including data scarcity, report hallucination, lack of model interpretability, and generalization across different institutions and imaging modalities. Ethical and regulatory concerns, such as accountability, bias, and patient privacy, must also be carefully considered before deploying AI-based reporting systems in real-world clinical practice.

- **Future Directions:** Promising research avenues include self-supervised learning, multimodal fusion of clinical data, human-in-the-loop AI systems, and enhanced interpretability techniques. The development of clinically validated AI models, combined with real-world testing and regulatory approval, is essential for ensuring the safe and effective adoption of automated radiology report generation in healthcare settings.

7.2. Final Remarks

Deep learning-based radiology report generation holds immense potential for improving radiologists' workflow efficiency, reducing reporting errors, and enhancing the overall quality of healthcare services. However, significant challenges remain in ensuring that AI-generated reports are clinically accurate, explainable, and trustworthy. Future research efforts should focus on integrating AI-driven automation with human expertise to develop hybrid systems that complement rather than replace radiologists.

With continued advancements in AI, increased availability of high-quality medical datasets, and rigorous validation through clinical trials, deep learning-powered radiology report generation systems may soon become an integral part of modern radiology practice. A collaborative effort between AI researchers, radiologists, and regulatory authorities will be essential to achieving this goal and ensuring the safe deployment of AI in medical imaging.

This concludes the systematic review of deep learning-based research on radiology report generation. Future work should continue exploring novel methodologies, addressing existing challenges, and pushing the boundaries of AI-driven medical imaging analysis.

References

[1] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. 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, in: AAAI 2019, 2019, pp. 590–597.

[2] J. D. López-Cabrera, R. Orozco-Morales, J. A. Portal-Diaz, O. Lovelle-Enríquez, M. Pérez-Díaz, Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging, Health and Technology 11 (2) (2021) 411–424. doi:10.1007/s12553-021-00520-2.

[3] P. Michael, H.-J. Yoon, Survey of image denoising methods for medical image classification, in: Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, 2020, p. 132. doi:10.1117/12.2549695.

[4] Y. Wang, K. Wang, X. Liu, T. Gao, J. Zhang, G. Wang, Self Adaptive Global-Local Feature Enhancement for Radiology Report Generation, 2023 IEEE International Conference on Image Processing (ICIP) (2022) 2275–2279.

[5] J. P. Cohen, M. Hashir, R. Brooks, H. Bertrand, On the limits of cross-domain generalization in automated x-ray prediction, Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR (2020) 121:136–155arXiv:2002.02497.

[6] D. Nukrai, R. Mokady, A. Globerson, Text-Only Training for Image Captioning using Noise-Injected CLIP, in: Findings of EMNLP 2022, Abu Dhabi, United Arab Emirates, 2022, pp. 4055–4063.

[7] A. Madani, M. Moradi, A. Karargyris, T. Syeda-Mahmood, Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1038–1042. doi:10.1109/ISBI.2018.8363749.

[8] C.-Y. Su, T.-Y. Tsai, C.-Y. Tseng, K.-H. Liu, C.-W. Lee, A Deep Learning Method for Alerting Emergency Physicians about the Presence of Sub-

phrenic Free Air on Chest Radiographs, Journal of Clinical Medicine 10 (2) (2021) 254. doi:10.3390/jcm10020254.

[9] 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: CVPR 2018, 2018, pp. 9049–9058.

[10] F. Narváez, G. Díaz, C. Poveda, E. Romero, An Automatic BI-RADS Description of Mammographic Masses by Fusing Multiresolution Features, Expert Systems with Applications 74 (2017) 82–95.

[11] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in: EMNLP 2021, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528.

[12] J. Ureta, O. Aran, J. P. Rivera, Detecting pneumonia in chest radiographs using convolutional neural networks, in: Twelfth International Conference on Machine Vision (ICMV 2019), SPIE, 2020, p. 116. doi:10.1117/12.2559527.

[13] Z. Li, Z. Zhong, Y. Li, T. Zhang, L. Gao, D. Jin, Y. Sun, X. Ye, L. Yu, Z. Hu, J. Xiao, L. Huang, Y. Tang, From community-acquired pneumonia to COVID-19: a deep learning–based method for quantitative analysis of COVID-19 on thick-section CT scans, Eur. Radiol. 30 (12) (2020) 6828–6837. doi:10.1007/s00330-020-07042-x.

[14] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 640–651. doi:10.1109/tpami.2016.2572683.

[15] J. Islam, Y. Zhang, Towards robust lung segmentation in chest radiographs with deep learning (2018). arXiv:1811.12638.

[16] S. Pathan, P. Siddalingaswamy, T. Ali, Automated Detection of Covid-19 from Chest X-ray scans using an optimized CNN architecture, Appl. Soft Comput. 104 (2021) 107238. doi:10.1016/j.asoc.2021.107238.

[17] M. P. Shah, S. N. Merchant, S. P. Awate, MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Vol. 11073, Springer, 2018, pp. 379–387. doi:10.1007/978-3-030-00937-$3_4$4.

[18] J. S. Suri, S. Agarwal, S. K. Gupta, A. Puvvula, M. Biswas, L. Saba, A. Bit, G. S. Tandel, M. Agarwal, A. Patrick, G. Faa, I. M. Singh, R. Oberleitner, M. Turk, P. S. Chadha, A. M. Johri, J. Miguel Sanches, N. N. Khanna, K. Viskovic, S. Mavrogeni, J. R. Laird, G. Pareek, M. Miner, D. W. Sobel, A. Balestrieri, P. P. Sfikakis, G. Tsoulfas, A. Protogerou, D. P. Misra, V. Agarwal, G. D. Kitas, P. Ahluwalia, J. Teji, M. Al-Maini, S. K. Dhanjil, M. Sockalingam, A. Saxena, A. Nicolaides, A. Sharma, V. Rathore, J. N. Ajuluchukwu, M. Fatemi, A. Alizad, V. Viswanathan, P. Krishnan, S. Naidu, A narrative review on characterization of acute respiratory distress syndrome in COVID-19-infected lungs using artificial intelligence, Comput. Biol. Med. 130 (2021) 104210. doi:10.1016/j.compbiomed.2021.104210.

[19] J. S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, 2011, pp. 2546–2554.

[20] B. Unnikrishnan, C. M. Nguyen, S. Balaram, C. S. Foo, P. Krishnaswamy, Semi-supervised Classification of Diagnostic Radiographs with NoTeacher: A Teacher that is Not Mean, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Vol. 12261, Springer, 2020, pp. 624–634. doi:10.1007/978-3-030-59710-$8_6$1.

[21] M. D. Li, N. T. Arun, M. Gidwani, K. Chang, F. Deng, B. P. Little, D. P. Mendoza, M. Lang, S. I. Lee, A. O'Shea, A. Parakh, P. Singh, J. Kalpathy-Cramer, Automated assessment of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks, medRxiv (may 2020). doi:10.1101/2020.05.20.20108159.

[22] K. G. van Leeuwen, S. Schalekamp, M. J. Rutten, B. van Ginneken, M. de Rooij, Artificial intelligence in radiology; 100 commercially available products and their scientific evidence, 2021, European Radiology (in press) (2021).

[23] A. S. Al-Waisy, S. Al-Fahdawi, M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. S. Maashi, M. Arif, B. Garcia-Zapirain, COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images, Soft Computing (Nov. 2020). doi:10.1007/s00500-020-05424-3.

[24] S. Schalekamp, N. Karssemeijer, A. M. Cats, B. De Hoop, B. H. J. Geurts, O. Berger-Hartog, B. van Ginneken, C. M. Schaefer-Prokop, The Effect of Supplementary Bone-Suppressed Chest Radiographs on the Assessment of a Variety of Common Pulmonary Abnormalities: Results of an Observer Study, Journal of Thoracic Imaging 31 (2) (2016) 119–125. doi:10.1097/RTI.0000000000000195.

[25] S. Candemir, S. Jaeger, W. Lin, Z. Xue, S. K. Antani, G. R. Thoma, Automatic heart localization and radiographic index computation in chest x-rays, in: Medical Imaging 2016: Computer-Aided Diagnosis, Vol. 9785, SPIE, 2016, p. 978517. doi:10.1117/12.2217209.

[26] X.-L. Zou, Y. Ren, D.-Y. Feng, X.-Q. He, Y.-F. Guo, H.-L. Yang, X. Li, J. Fang, Q. Li, J.-J. Ye, L.-Q. Han, T.-T. Zhang, A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study, PLOS ONE 15 (7) (2020) e0236378. doi:10.1371/journal.pone.0236378.

[27] N. Dong, M. Xu, X. Liang, Y. Jiang, W. Dai, E. Xing, Neural Architecture Search for Adversarial Medical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Vol. 11769, Springer, 2019, pp. 828–836. doi:10.1007/978-3-030-32226-$7_9$2.

[28] Siemens (2021).

[29] T. Khatibi, A. Shahsavari, A. Farahani, Proposing a novel multi-instance learning model for tuberculosis recognition from chest X-ray images based on CNNs,

complex networks and stacked ensemble, Physical and Engineering Sciences in Medicine (Feb. 2021). doi:10.1007/s13246-021-00980-w.

[30] K. Canas, B. Ubiera, X. Liu, Y. Liu, Scalable biomedical image synthesis with GAN, in: Proceedings of the Practice and Experience on Advanced Research Computing, ACM, 2018. doi:10.1145/3219104.3229261.

[31] A. Altan, S. Karasu, Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique, Chaos, Solitons & Fractals 140 (2020) 110071. doi:10.1016/j.chaos.2020.110071.

[32] J. E. McManigle, R. R. Bartz, L. Carin, Y-Net for Chest X-Ray Preprocessing: Simultaneous Classification of Geometry and Segmentation of Annotations, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1266–1269. doi:10.1109/EMBC44109.2020.9176334.

[33] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, TPAMI 39 (6) (2017) 1137–1149.

[34] V. T. Pham, T. P. Nguyen, Identification and localization covid-19 abnormalities on chest radiographs, in: The International Conference on Artificial Intelligence and Computer Vision, Springer, 2023, pp. 251–261.

[35] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, X. Sun, Contrastive Attention for Automatic Chest X-ray Report Generation, in: Findings of ACL-IJCNLP 2021, Online, 2021, pp. 269–280.

[36] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov,

P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, CoRR abs/2307.09288 (2023). arXiv:2307.09288.

[37] H. Alshazly, C. Linse, E. Barth, T. Martinetz, Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning, Sensors 21 (2) (2021) 455. doi:10.3390/s21020455.

[38] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating Radiology Reports via Memory-driven Transformer, in: EMNLP 2020, Online, 2020, pp. 1439–1449.

[39] H. Nguyen, D. Nie, T. Badamdorj, Y. Liu, Y. Zhu, J. Truong, L. Cheng, Automated Generation of Accurate & Fluent Medical X-ray Reports, in: EMNLP 2021, Online and Punta Cana, Dominican Republic, 2021, pp. 3552–3569.

[40] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang, Scaling Up Vision-language Pretraining for Image Captioning, in: CVPR 2022, 2022, pp. 17959–17968.

[41] R. D. S. Portela, J. R. G. Pereira, M. G. F. Costa, C. F. F. C. Filho, Lung Region Segmentation in Chest X-Ray Images using Deep Convolutional Neural Networks, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1246–1249. doi:10.1109/EMBC44109.2020.9175478.

[42] A. Nicolson, J. Dowling, B. Koopman, Improving Chest X-ray Report Generation by Leveraging Warm Starting, Artificial Intelligence in Medicine 144 (2023) 102633.

[43] D. Han, M. Heuvelmans, M. Rook, M. Dorrius, L. van Houten, N. W. Price, L. C. Pickup, P. Novotny, M. Oudkerk, J. Declerck, F. Gleeson, P. van Ooi-

jen, R. Vliegenthart, Evaluation of a novel deep learning–based classifier for perifissural nodules, Eur. Radiol. (Dec. 2020). doi:10.1007/s00330-020-07509-x.

[44] A. J. DeGrave, J. D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, medRxiv (sep 2020). doi:10.1101/2020.09.13.20193565.

[45] J. Sun, D. Wei, L. Wang, Y. Zheng, Lesion Guided Explainable Few Weak-Shot Medical Report Generation, in: MICCAI 2022, Vol. 13435, 2022, pp. 615–625.

[46] H. Ravishankar, R. Venkataramani, S. Anamandra, P. Sudhakar, P. Annangi, Feature Transformers: Privacy Preserving Lifelong Learners for Medical Imaging, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Vol. 11767, Springer, 2019, pp. 347–355. doi:10.1007/978-3-030-32251-$9_3$8.

[47] M. Nash, R. Kadavigere, J. Andrade, C. A. Sukumar, K. Chawla, V. P. Shenoy, T. Pande, S. Huddart, M. Pai, K. Saravu, Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India, Scientific Reports 10 (1) (2020) 210. doi:10.1038/s41598-019-56589-3.

[48] L. M. Prevedello, S. S. Halabi, G. Shih, C. C. Wu, M. D. Kohli, F. H. Chokshi, B. J. Erickson, J. Kalpathy-Cramer, K. P. Andriole, A. E. Flanders, Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions, Radiology: Artificial Intelligence 1 (1) (2019) e180031. doi:10.1148/ryai.2019180031.

[49] S. S. Habib, S. Rafiq, S. M. A. Zaidi, R. A. Ferrand, J. Creswell, B. Van Ginneken, W. Z. Jamal, K. S. Azeemi, S. Khowaja, A. Khan, Evaluation of computer aided detection of tuberculosis on chest radiography among people with diabetes in Karachi Pakistan, Scientific Reports 10 (1) (2020) 6276. doi:10.1038/s41598-020-63084-7.

[50] A. Bayat, A. Sekuboyina, J. C. Paetzold, C. Payer, D. Stern, M. Urschler, J. S. Kirschke, B. H. Menze, Inferring the 3D Standing Spine Posture from 2D Radiographs, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Vol. 12266, Springer, 2020, pp. 775–784. doi:10.1007/978-3-030-59725-2$_7$5.

[51] R. Bigolin Lanfredi, J. D. Schroeder, C. Vachet, T. Tasdizen, Adversarial Regression Training for Visualizing the Progression of Chronic Obstructive Pulmonary Disease with Chest X-Rays, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Vol. 11769, Springer, 2019, pp. 685–693. doi:10.1007/978-3-030-32226-7$_7$6.

[52] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, M. Reyes, Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer, 2018, pp. 580–588. doi:10.1007/978-3-030-00934-2$_6$5.

[53] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, Journal of Machine Learning Research 12 (2011) 2121–2159.

[54] O. Gozes, H. Greenspan, Deep Feature Learning from a Hospital-Scale Chest X-ray Dataset with Application to TB Detection on a Small-Scale Dataset, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 4076–4079. doi:10.1109/EMBC.2019.8856729.

[55] J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 7263–7271. doi:10.1109/cvpr.2017.690.

[56] V. Madaan, A. Roy, C. Gupta, P. Agrawal, A. Sharma, C. Bologa, R. Prodan, XCOVNet: Chest X-ray Image Classification for COVID-19 Early Detection

Using Convolutional Neural Networks, New Generation Computing (Feb. 2021). doi:10.1007/s00354-021-00121-7.

[57] M. Woźniak, D. Połap, G. Capizzi, G. L. Sciuto, L. Kośmider, K. Frankiewicz, Small lung nodules detection based on local variance analysis and probabilistic neural network, Computer Methods and Programs in Biomedicine 161 (2018) 173–180. doi:10.1016/j.cmpb.2018.04.025.

[58] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift (2015). arXiv:1502.03167.

[59] M. F. Rahman, T.-L. B. Tseng, M. Pokojovy, W. Qian, B. Totada, H. Xu, An automatic approach to lung region segmentation in chest x-ray images using adapted U-Net architecture, in: Medical Imaging 2021: Physics of Medical Imaging, Vol. 11595, International Society for Optics and Photonics, 2021, p. 115953I. doi:10.1117/12.2581882.

[60] F.-F. Xue, J. Peng, R. Wang, Q. Zhang, W.-S. Zheng, Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Vol. 11769, Springer, 2019, pp. 846–854. doi:10.1007/978-3-030-32226-$7_9$4.

[61] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1125–1134. doi:10.1109/cvpr.2017.632.

[62] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum Learning, in: ICML, New York, NY, USA, 2009, p. 41–48.

[63] X. Zhang, G. Kumar, H. Khayrallah, K. Murray, J. Gwinnup, M. J. Martindale, P. McNamee, K. Duh, M. Carpuat, An Empirical Exploration of Curriculum Learning for Neural Machine Translation, CoRR abs/1811.00739 (2018). arXiv:1811.00739.

[64] S. Anis, K. W. Lai, J. H. Chuah, M. A. Shoaib, H. Mohafez, M. Hadizadeh, Y. Ding, Z. C. Ong, An overview of deep learning approaches in chest radiograph, IEEE Access (2020). doi:10.1109/access.2020.3028390.

[65] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, C.-N. Hsu, Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation, in: Findings of EMNLP 2021, Punta Cana, Dominican Republic, 2021, pp. 4009–4015.

[66] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, X. Liu, Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays, Pattern Recognition 110 (2021) 107613. doi:10.1016/j.patcog.2020.107613.

[67] T. Dyer, L. Dillard, M. Harrison, T. N. Morgan, R. Tappouni, Q. Malik, S. Rasalingham, Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm, Clin. Radiol. (2021) S0009926021000763doi:10.1016/j.crad.2021.01.015.

[68] X. Li, R. Cao, D. Zhu, Vispi: Automatic visual perception and interpretation of chest x-rays, in: Medical Imaging with Deep Learning, 2020, pp. 1–8.

[69] A. M. Fischer, A. Varga-Szemes, S. S. Martin, J. I. Sperl, P. Sahbaee, D. Neumann, J. Gawlitza, T. Henzler, C. M. Johnson, J. W. Nance, S. O. Schoenberg, U. J. Schoepf, Artificial Intelligence-based Fully Automated Per Lobe Segmentation and Emphysema-quantification Based on Chest Computed Tomography Compared With Global Initiative for Chronic Obstructive Lung Disease Severity of Smokers, Journal of Thoracic Imaging 35 Suppl 1 (2020) S28–S34. doi:10.1097/RTI.0000000000000500.

[70] V. Subramanian, H. Wang, J. T. Wu, K. C. L. Wong, A. Sharma, T. Syeda-Mahmood, Automated Detection and Type Classification of Central Venous Catheters in Chest X-Rays, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Vol. 11769, Springer, 2019, pp. 522–530. doi:10.1007/978-3-030-32226-$7_5$8.

[71] C. S. Zhu, P. F. Pinsky, B. S. Kramer, P. C. Prorok, M. P. Purdue, C. D. Berg, J. K. Gohagan, The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial and Its Associated Research Resource, JNCI Journal of the National Cancer Institute 105 (22) (2013) 1684–1693. doi:10.1093/jnci/djt281.

[72] P.-C. Kuo, C. C. Tsai, D. M. López, A. Karargyris, T. J. Pollard, A. E. W. Johnson, L. A. Celi, Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph, npj Digital Medicine 4 (1) (2021) 25. doi:10.1038/s41746-021-00393-9.

[73] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas, J. Barfett, Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs:, Investigative Radiology 52 (5) (2017) 281–287. doi:10.1097/RLI.0000000000000341.

[74] A. Karargyris, K. C. L. Wong, J. T. Wu, M. Moradi, T. Syeda-Mahmood, Boosting the Rule-Out Accuracy of Deep Disease Detection Using Class Weight Modifiers, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 877–881. doi:10.1109/ISBI.2019.8759532.

[75] B. van Ginneken, M. B. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, Medical Image Analysis 10 (1) (2006) 19–40. doi:10.1016/j.media.2005.02.002.

[76] S. Conjeti, A. G. Roy, A. Katouzian, N. Navab, Hashing with Residual Networks for Image Retrieval, in: Medical Image Computing and Computer Assisted Intervention MICCAI 2017, Vol. 10435, Springer, 2017, pp. 541–549. doi:10.1007/978-3-319-66179-$7_6$2.

[77] M. Owais, M. Arsalan, T. Mahmood, Y. H. Kim, K. R. Park, Comprehensive Computer-Aided Decision Support Framework to Diagnose Tuberculosis From Chest X-Ray Images: Data Mining Study, JMIR Medical Informatics 8 (12) (2020) e21790. doi:10.2196/21790.

[78] Y.-C. Liu, Y.-C. Lin, P.-Y. Tsai, O. Iwata, C.-C. Chuang, Y.-H. Huang, Y.-S. Tsai, Y.-N. Sun, Convolutional Neural Network-Based Humerus Segmentation and Application to Bone Mineral Density Estimation from Chest X-ray Images of Critical Infants, Diagnostics 10 (12) (2020) 1028. doi:10.3390/diagnostics10121028.

[79] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-text Models (2022). arXiv:2210.08402.

[80] Y. Balabanova, R. Coker, I. Fedorin, S. Zakharova, S. Plavinskij, N. Krukov, R. Atun, F. Drobniewski, Variability in interpretation of chest radiographs among russian clinicians and implications for screening programmes: observational study, BMJ 331 (7513) (2005) 379–382. doi:10.1136/bmj.331.7513.379.

[81] A. Keles, M. B. Keles, A. Keles, COV19-CNNet and COV19-ResNet: Diagnostic Inference Engines for Early Detection of COVID-19, Cognitive Computation (Jan. 2021). doi:10.1007/s12559-020-09795-5.

[82] P. Srinivasan, D. Thapar, A. Bhavsar, A. Nigam, Hierarchical x-ray report generation via pathology tags and multi head attention, in: ACCV 2020, Cham, 2021, pp. 600–616.

[83] D. Yu, K. Zhang, L. Huang, B. Zhao, X. Zhang, X. Guo, M. Li, Z. Gu, G. Fu, M. Hu, Y. Ping, Y. Sheng, Z. Liu, X. Hu, R. Zhao, Detection of peripherally inserted central catheter (PICC) in chest X-ray images: A multi-task deep learning model, Computer Methods and Programs in Biomedicine 197 (2020) 105674. doi:10.1016/j.cmpb.2020.105674.

[84] X. Liu, S. Wang, Y. Deng, K. Chen, Coronary artery calcification (CAC) classification with deep convolutional neural networks, in: Medical Imaging 2017: Computer-Aided Diagnosis, Vol. 10134, SPIE, 2017, p. 101340M. doi:10.1117/12.2253974.

[85] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, D. Jurafsky, Improving Factual Completeness and Consistency of Image-to-text Radiology Report Generation, in: NAACL 2021, Online, 2021, pp. 5288–5304.

[86] M. Long, H. Zhu, J. Wang, M. I. Jordan, Deep transfer learning with joint adaptation networks, in: International Conference on Machine Learning, 2017, pp. 2208–2217.

[87] A. Gooßen, H. Deshpande, T. Harder, E. Schwab, I. Baltruschat, T. Mabotuwana, N. Cross, A. Saalbach, Deep learning for pneumothorax detection and localization in chest radiographs (2019). arXiv:1907.07324.

[88] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: BioNLP, Florence, Italy, 2019, pp. 319–327.

[89] R. M. Neal, Connectionist learning of belief networks, Artificial Intelligence 56 (1) (1992) 71–113. doi:10.1016/0004-3702(92)90065-6.

[90] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification, Scientific Reports 9 (1) (2019) 6381. doi:10.1038/s41598-019-42294-8.

[91] A. Karargyris, S. Kashyap, J. T. Wu, A. Sharma, M. Moradi, T. Syeda-Mahmood, Age prediction using a large chest x-ray dataset, in: Medical Imaging 2019: Computer-Aided Diagnosis, SPIE, 2019, p. 66. doi:10.1117/12.2512922.

[92] Y. Song, Y. Tian, N. Wang, F. Xia, Summarizing Medical Conversations via Identifying Important Utterances, in: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 717–729.

[93] C. Raffel, D. P. W. Ellis, Feed-forward networks with attention can solve some long-term memory problems, CoRR abs/1512.08756 (2015). arXiv:1512.08756.

[94] F. A. Mettler, M. Bhargavan, K. Faulkner, D. B. Gilley, J. E. Gray, G. S. Ibbott, J. A. Lipoti, M. Mahesh, J. L. McCrohan, M. G. Stabin, B. R. Thomadsen, T. T. Yoshizumi, Radiologic and nuclear medicine studies in the united states and worldwide: Frequency, radiation dose, and comparison with other radiation sources—1950–2007, Radiology 253 (2) (2009) 520–531. doi:10.1148/radiol.2532082010.

[95] K. Kale, P. Bhattacharyya, A. Shetty, M. Gune, K. Shrivastava, R. Lawyer, S. Biswas, Knowledge Graph Construction and Its Application in Automatic Radiology Report Generation from Radiologist's Dictation, CoRR abs/2206.06308 (2022). arXiv:2206.06308.

[96] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, S. Horng, MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs, CoRR abs/1901.07042 (2019). arXiv:1901.07042.

[97] Y. Zhou, L. Huang, T. Zhou, H. Fu, L. Shao, Visual-textual Attentive Semantic Consistency for Medical Report Generation, in: ICCV 2021, 2021, pp. 3965–3974.

[98] R. Hermoza, G. Maicas, J. C. Nascimento, G. Carneiro, Region Proposals for Saliency Map Refinement for Weakly-Supervised Disease Localisation and Classification, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Vol. 12266, Springer, 2020, pp. 539–549. doi:10.1007/978-3-030-59725-2$_5$2.

[99] Y. Bar, I. Diamant, L. Wolf, H. Greenspan, Deep learning with non-medical training used for chest pathology identification, in: Medical Imaging 2015: Computer-Aided Diagnosis, Vol. 9414, SPIE, 2015, p. 94140V. doi:10.1117/12.2083124.

[100] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: The handbook of brain theory and neural networks, MIT Press, 1998, pp. 255–258.

[101] S. Sukhbaatar, a. szlam, J. Weston, R. Fergus, End-to-end Memory Networks, in: Advances in Neural Information Processing Systems, Vol. 28, 2015, pp. 1–9.

[102] S. R. Nayak, D. R. Nayak, U. Sinha, V. Arora, R. B. Pachori, Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study, Biomed. Signal Process. Control 64 (2021) 102365. doi:10.1016/j.bspc.2020.102365.

[103] V. Kovalev, S. Kazlouski, Examining the capability of GANs to replace real biomedical images in classification models training, in: Communications in Computer and Information Science, Springer, 2019, pp. 98–107. doi:10.1007/978-3-030-35430-5$_9$.

[104] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: A Neural Image Caption Generator, in: CVPR 2015, 2015.

[105] Z. Li, Z. Hou, C. Chen, Z. Hao, Y. An, S. Liang, B. Lu, Automatic cardiothoracic ratio calculation with deep learning, IEEE Access 7 (2019) 37749–37756. doi:10.1109/ACCESS.2019.2900053.

[106] J. von Berg, S. Young, H. Carolus, R. Wolz, A. Saalbach, A. Hidalgo, A. Giménez, T. Franquet, A novel bone suppression method that improves lung nodule detection, International Journal of Computer Assisted Radiology and Surgery 11 (4) (2015) 641–655. doi:10.1007/s11548-015-1278-y.

[107] G. Dhiman, V. Chang, K. Kant Singh, A. Shankar, ADOPT: automatic deep learning and optimization-based approach for detection of novel coronavirus COVID-19 disease using X-ray images, J. Biomol. Struct. Dyn. (2021) 1–13doi:10.1080/07391102.2021.1875049.

[108] F. Demir, DeepCoroNet: A deep LSTM approach for automated detection of COVID-19 cases from chest X-ray images, Appl. Soft Comput. 103 (2021) 107160. doi:10.1016/j.asoc.2021.107160.

[109] S. Onodera, Y. Lee, Y. Tanaka, Evaluation of dose reduction potential in scatter-corrected bedside chest radiography using U-net, Radiological Physics and Technology 13 (4) (2020) 336–347. doi:10.1007/s12194-020-00586-z.

[110] S. Toba, Y. Mitani, N. Yodoya, H. Ohashi, H. Sawada, H. Hayakawa, M. Hirayama, A. Futsuki, N. Yamamoto, H. Ito, T. Konuma, H. Shimpo, M. Takao, Prediction of Pulmonary to Systemic Flow Ratio in Patients With Congenital Heart Disease Using Deep Learning–Based Analysis of Chest Radiographs, JAMA Cardiology 5 (4) (2020) 449. doi:10.1001/jamacardio.2019.5620.

[111] J. Stubblefield, M. Hervert, J. L. Causey, J. A. Qualls, W. Dong, L. Cai, J. Fowler, E. Bellis, K. Walker, J. H. Moore, S. Nehring, X. Huang, Transfer learning with chest X-rays for ER patient classification, Sci. Rep. 10 (1) (2020) 20900. doi:10.1038/s41598-020-78060-4.

[112] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays, Comput. Methods Programs Biomed. 196 (2020) 105608. doi:10.1016/j.cmpb.2020.105608.

[113] C. J. McDonald, J. M. Overhage, M. Barnes, G. Schadow, L. Blevins, P. R. Dexter, B. Mamlin, The Indiana Network For Patient Care: A Working Local Health Information Infrastructure, Health Affairs 24 (5) (2005) 1214–1220, publisher: Health Affairs. doi:10.1377/hlthaff.24.5.1214.

[114] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual Instruction Tuning, CoRR abs/2304.08485 (2023). arXiv:2304.08485.

[115] C.-H. Wei, Y. Li, P. J. Huang, Mammogram Retrieval through Machine Learning within BI-RADS Standards, Journal of Biomedical Informatics 44 (4) (2011) 607–614.

[116] A. Shamsi, H. Asgharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebria, D. Nahavandi, S. Nahavandi, D. Srinivasan, An Uncertainty-Aware Transfer Learning-

Based Framework for COVID-19 Diagnosis, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–10doi:10.1109/TNNLS.2021.3054306.

[117] 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, Ann Arbor, Michigan, 2005, pp. 65–72.

[118] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2223–2232. doi:10.1109/iccv.2017.244.

[119] V.-T. Pham, C.-M. Tran, S. Zheng, T.-M. Vu, S. Nath, Chest x-ray abnormalities localization via ensemble of deep convolutional neural networks, in: 2021 International Conference on Advanced Technologies for Communications (ATC), IEEE, 2021, pp. 125–130.

[120] F. H. Chokshi, A. E. Flanders, L. M. Prevedello, C. P. Langlotz, Fostering a healthy AI ecosystem for radiology: Conclusions of the 2018 RSNA summit on AI in radiology, Radiology: Artificial Intelligence 1 (2) (2019) 190021. doi:10.1148/ryai.2019190021.

[121] V. Chernyak, K. J. Fowler, A. Kamaya, A. Z. Kielar, K. M. Elsayes, M. R. Bashir, Y. Kono, R. K. Do, D. G. Mitchell, A. G. Singal, A. Tang, C. B. Sirlin, Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients, Radiology 289 (3) (2018) 816–830.

[122] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, Q. Huang, Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation, IEEE Transactions on Multimedia 25 (2023) 167–178.

[123] F. Liu, S. Ge, X. Wu, Competence-based Multimodal Curriculum Learning for

Medical Report Generation, in: ACL-IJCNLP 2021, Online, 2021, pp. 3001–3012.