Jurnal Masyarakat Informatika, 16(1), 2025
DOI:10.14710/jmasif.16.1.73127
ISSN: 2777-0648

Systematic Literature Review on Medical Image Captioning Using CNN-LSTM and Transformer-Based Models

Husni Fadhilah*, Nugraha Priya Utama
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia
* *Corresponding author: 23523034@std.stei.itb.ac.id*

Abstract
Medical image captioning aims to automatically generate descriptive textual reports from medical images, serving as a crucial task in the development of computer-aided diagnosis systems. This paper presents a systematic literature review (SLR) of medical image captioning models, focusing specifically on CNN-LSTM and Transformer-based architectures published between 2019 and 2024. The study analyzes 40 selected articles based on their architectures, datasets, evaluation metrics, and clinical applicability. We provide a comparative synthesis based on five research questions (RQ1–RQ5), highlighting advancements, challenges, and future directions. A taxonomy of model types is introduced, and summary tables covering datasets, evaluation scores, and concept integration are presented. Our findings reveal the growing importance of multimodal learning, domain-specific metrics (e.g., BERTScore, RadGraph F1), and the integration of medical ontologies such as UMLS. The paper identifies key gaps in interpretability, clinical validation, and real-world applicability, providing guidance for future research.

Keywords : *Medical image captioning, convolutional neural network, transformer, healthcare ai, automatic report generation*

1 Introduction

In the fields of artificial intelligence (AI) and healthcare, medical image captioning is a critical task that aims to generate accurate and informative textual descriptions from medical images such as computed tomography (CT) scans, magnetic resonance imaging (MRI), and X-rays [1]. The exponential growth of imaging data has placed increasing pressure on radiologists [2], prompting the need for automated and standardized reporting systems to reduce workload and diagnostic variability [3]. Traditional image analysis relied heavily on handcrafted feature extraction, which requires expert knowledge and is prone to inconsistency. Over time, captioning methods have evolved from rigid retrieval and template-based approaches to more adaptive deep learning models, most notably encoder-decoder architectures enhanced with attention mechanisms and transformer-based designs [1]. These architectures integrate vision and language processing, enabling the transformation of complex visual inputs into coherent textual descriptions [4].

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have shown strong performance in medical imaging tasks such as classification, segmentation, and captioning by effectively extracting spatial hierarchies from images [4]. While CNN-LSTM-based models extract spatial features effectively, their limited capacity for long-range dependencies has prompted the use of encoder-decoder and Transformer-based architectures [5]. Still, these models

Received: 10 May 2025 ; Revised: 20 May 2025; Accepted: 21 May 2025; Published: 31 May 2025

often fall short in handling clinical terminology, prompting the adoption of more advanced architectures like Transformers.

Originally developed for natural language processing (NLP) applications, transformer models have proven to be remarkably effective at representing contextual information and long-range dependencies across input sequences [5, 6]. Transformers, in contrast to CNNs, are able to generate text descriptions while focusing on pertinent portions of the input image thanks to self-attention processes [7]. Because they can better capture hierarchical feature representations than conventional CNNs, Vision Transformers (ViT) and their variations, such Swin Transformer, have been investigated in medical imaging applications [6]. By offering superior contextual comprehension and generalization across a variety of medical datasets, these models have outperformed traditional methods.

Despite these developments, the intricacy of medical terminology and the requirement for sizable, annotated datasets make medical image captioning a difficult undertaking [7]. Domain specialists must put in a lot of work to annotate medical images with matching captions, which frequently leaves little data available for building strong models. Furthermore, the model's capacity to generalize to rare diseases is limited by the strong bias of available datasets like IU-Xray, MIMIC-CXR, and ROCO towards common ailments [8]. The interpretability of reports produced by AI is another difficulty, raising questions about acceptance and trust in therapeutic practice. For these models to be successfully used in real-world contexts, it is essential that the generated captions adhere to clinical principles and standards.

To overcome the difficulty of integrating clinical knowledge into the generated reports, it has been suggested that medical image captioning models use external knowledge bases, such as the Unified Medical Language System (UMLS) [8]. Medical terminology is standardized with the use of Concept Unique Identifiers (CUI), which also enhances the correctness and consistency of the generated descriptions. According to studies, adding CUIs to the training process improves the model's capacity to generate interpretable and clinically significant captions, increasing the output's utility for radiologists and other healthcare professionals [9, 10].

Three main architecture families e.g CNN-LSTM, Transformer-based, and hybrid encoder-decoder models, are highlighted in this systematic literature review (SLR), which offers a targeted examination of deep learning-based methods for medical image captioning. CNN-LSTM pipelines served as the foundation for many early medical captioning systems because of their interpretability, ease of architecture, and suitability for use in clinical settings with low resources [3]. However, recent research has shown that Transformer-based models, perform noticeably better than their CNN-based counterparts in producing semantically coherent and clinically correct reports [23].

Key elements including model architecture design, dataset utilization (e.g., IU X-Ray, MIMIC-CXR, ROCO), assessment metrics (BLEU, ROUGE, CIDEr, BERTScore), and clinical deployment issues are methodically synthesized in the review. Although previous reviews of the research on image captioning have generally addressed either natural language generation or general medical vision tasks [2, 5], they either did not compare the performance of different model types directly or did not look at the architectural implications for clinical relevance. Furthermore, a lot of earlier research only used traditional n-gram overlap criteria, like BLEU or ROUGE, which have been demonstrated to be inconsistent with radiologists' assessments and to miss factual correctness [26].

By addressing these drawbacks, this paper provides a comparative synthesis of CNN-LSTM, Transformer, and hybrid techniques, thereby solving a significant gap in the field. In addition to

highlighting the language and clinical aspects of model performance, it also points out unresolved research issues such as explainability, factual consistency, and domain generalization. This review is limited in scope to peer-reviewed studies published in English between 2019 and 2024, focusing exclusively on deep learning-based image captioning models. Studies that are purely retrieval-based, GAN-driven, or outside the medical imaging domain were excluded. By clearly defining these boundaries, the study ensures targeted synthesis while offering practical guidance for future research on the development of reliable and clinically integrated medical image captioning systems.

2 Research Method

In order to gather, examine, and synthesize the body of research on medical image captioning using deep learning techniques, specifically Convolutional Neural Networks (CNNs) - Long Short Term Memory (LSTM) and Transformer-based models, this study uses a systematic literature review (SLR) methodology. Following the recommendations made by Kitchenham and Charters [11], the approach consists of formulating research questions, creating a search strategy, implementing inclusion and exclusion criteria, extracting data, and synthesizing findings.

2.1 Research Questions

The following research questions (RQs) have been developed in order to gain a thorough grasp of the body of existing literature and pinpoint areas in need of further investigation:

1. RQ1: Using CNN-LSTM and Transformer-based architectures, what are the most advanced techniques currently employed in medical image captioning?
2. RQ2: Which evaluation metrics and datasets are frequently employed in this field?
3. RQ3: What is the performance and clinical applicability comparison between Transformer-based and CNN-LSTM based approaches?
4. RQ4: What difficulties arise when using automatic captioning for medical images in actual clinical settings?
5. RQ5: How could automated medical image captioning systems be improved in the future?

2.2 Search Strategy

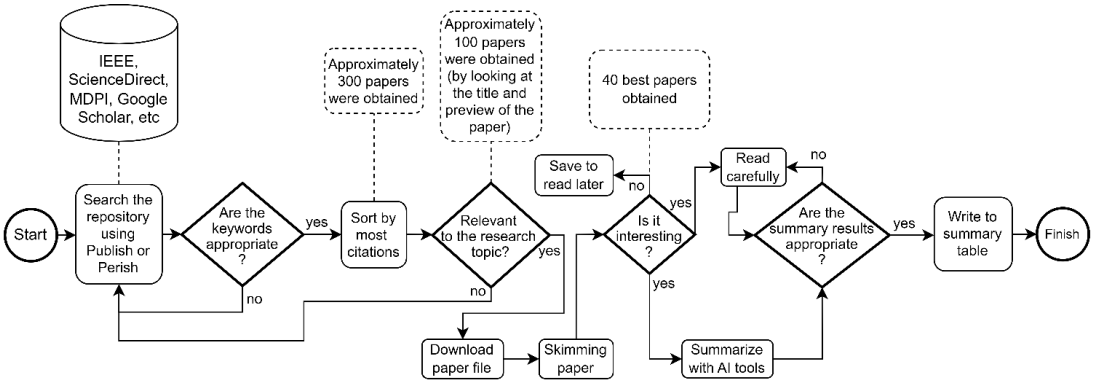


Figure 1 Search strategy workflow

To guarantee thorough coverage of pertinent studies, the literature search was carried out across several reliable digital databases. Among the databases that were searched are:

1. IEEE Xplore

2. PubMed
3. ScienceDirect
4. SpringerLink
5. Google Scholar

To optimize the retrieval of pertinent articles, Boolean operators and keywords were combined. The following search terms were developed in accordance with the research focus.

```
"CNN" AND "medical imaging" AND ("captioning" OR "description generation")
"Transformer" AND "medical image analysis" AND ("diagnosis" OR "automatic report generation")
"Deep learning" AND "medical image captioning" AND ("X-ray" OR "CT" OR "MRI")
"Hybrid model" AND ("CNN" OR "Transformer") AND "medical image"
"Automatic report generation" AND "medical imaging" AND ("self-supervised learning" OR "transfer learning")
```

Figure 2 Deep learning search terms for diagnosing and medical image captioning

The following filters were used to narrow down the search results:

1. Publications from 2019 to 2024, making sure to incorporate the most recent developments.
2. Papers on AI in healthcare that have been published at conferences and peer-reviewed journals.
3. Publications in English exclusively.

2.3 Criteria for Inclusion and Exclusion

The following set of inclusion and exclusion criteria was established in order to guarantee the quality and applicability of the included studies:

Table 1 Criteria for inclusion and exclusion used in this study

No	Inclusion Criteria	Exclusion Criteria
1	Research that was published between 2019 and 2024 in indexed journals and prestigious conferences.	Studies that are not written in English.
2	Studies that explicitly apply CNN and Transformer models to the captioning of medical images.	Research lacking sufficient empirical validation or methodological clarity.
3	Studies that use assessment metrics like BLEU, ROUGE, CIDEr, and METEOR to present empirical data.	Review articles and opinion papers that do not present novel experimental findings.
4	Articles addressing the difficulties and prospects for further study in medical image captioning.	Duplicate studies or preprints without final publication.

2.4 Extraction and Synthesis of Data

Relevant information was gathered from the chosen studies using a systematic data extraction procedure. Among the elements of the extracted data are:

1. Study details: title, authors, year of publication, and location.
2. Model Architecture: The kind of encoder-decoder models that are employed (Hybrid methods, CNN, LSTM, Transformer, etc.).
3. Dataset Used: Open-IU, ROCO, and MIMIC-CXR are examples of public datasets.
4. Evaluation Metrics: Performance reported in terms of BLEU, ROUGE, CIDEr, METEOR, and BertScore.
5. Challenges Identified: Issues related to dataset quality, interpretability, and generalization.
6. Future Directions: Suggestions for model improvements and clinical integration.

Thematic analysis was utilized to synthesize the data, classifying the results according to patterns, similarities, and variations in the methods employed in the various research.

2.5 Evaluation of Quality

A quality evaluation checklist that comprises the following criteria was used to evaluate each chosen study in order to guarantee the validity and reliability of the results:

1. Relevance to the Research Questions: How well the study responds to the research questions that were posed.
2. Methodological Rigor: The lucidity of the research design and evaluation metrics, among other aspects.
3. Replicability: Whether enough information is given to enable the experiments to be repeated.
4. Findings and Results: The importance and influence of the stated results in the medical domain.
5. Bias Assessment: Finding possible biases, like imbalances in the dataset or restricted generalizability.

To guarantee that only top-notch research was incorporated into the final synthesis, each paper was graded according to these standards.

2.6 Analysis of Data and Comparative Assessment

The gathered data was analyzed using both quantitative and qualitative methods. Key performance metrics were evaluated between CNN-LSTM and Transformer-based models, including:

1. Accuracy and Efficiency: Performance is assessed across various datasets using common measures.
2. Interpretability: Evaluation of the generated captions usefulness and clarity.
3. Computational Complexity: Evaluation of each model's resource needs.
4. Clinical Relevance: Assessing the degree to which the generated captions correspond with reports from human experts.

To conduct a meta-analysis and find trends among the examined research, statistical methods were employed.

2.7 The Methodical Review Procedure

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [12] was adhered to throughout the review procedure, guaranteeing an open and organized methodology. The following Figure 3 is an overview of the procedure:

1. Identification: Looking through several databases to find pertinent articles.
2. Screening: Using predetermined criteria, duplicate and unnecessary investigations are eliminated.
3. Eligibility: Examining the appropriateness of chosen publications through a full-text examination.
4. Data Extraction: Noting important findings from relevant research.
5. Synthesis: Condensing research results and coming to insightful conclusions.

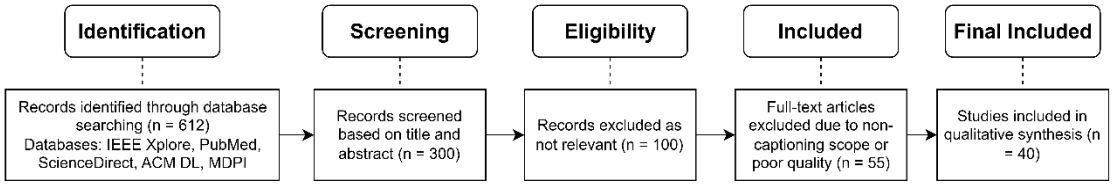
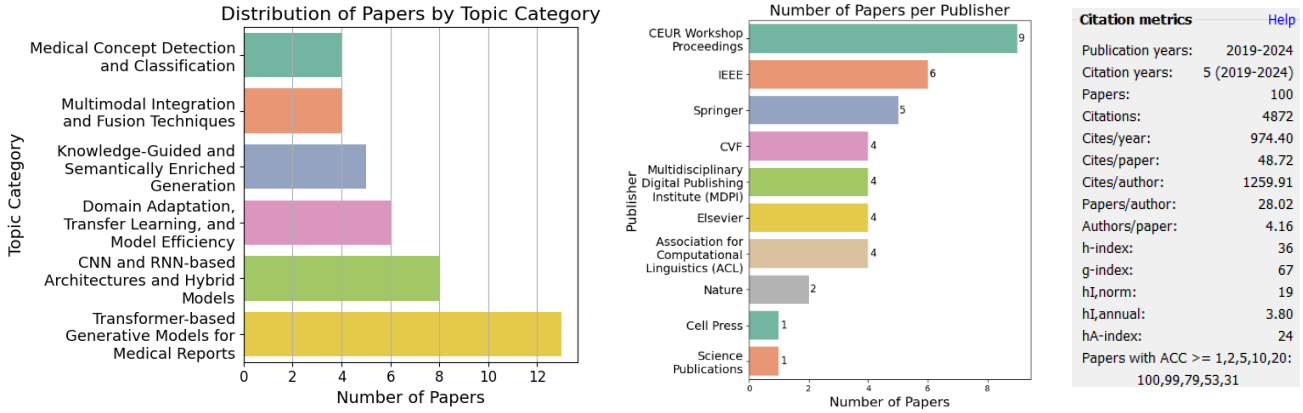


Figure 3 PRISMA diagram of 612 records, with 40 studies included in the final synthesis

3 Result and Discussion

The systematic literature review's findings offer insightful information about the state-of-the-art methods, difficulties, and potential paths in medical image captioning. Figure 4 explains several criteria, including model designs, datasets, performance measures, and important discoveries, that were used to examine the chosen papers.



(a) (b) (c)

Figure 4 Searching result of paper: (a) Distribution of Papers by Topic Category, (b) Number of Papers per Publisher, (c) Citation Metrics Overview from Publish or Perish.

3.1 Summary of Models for Medical Image Captioning (RQ1)

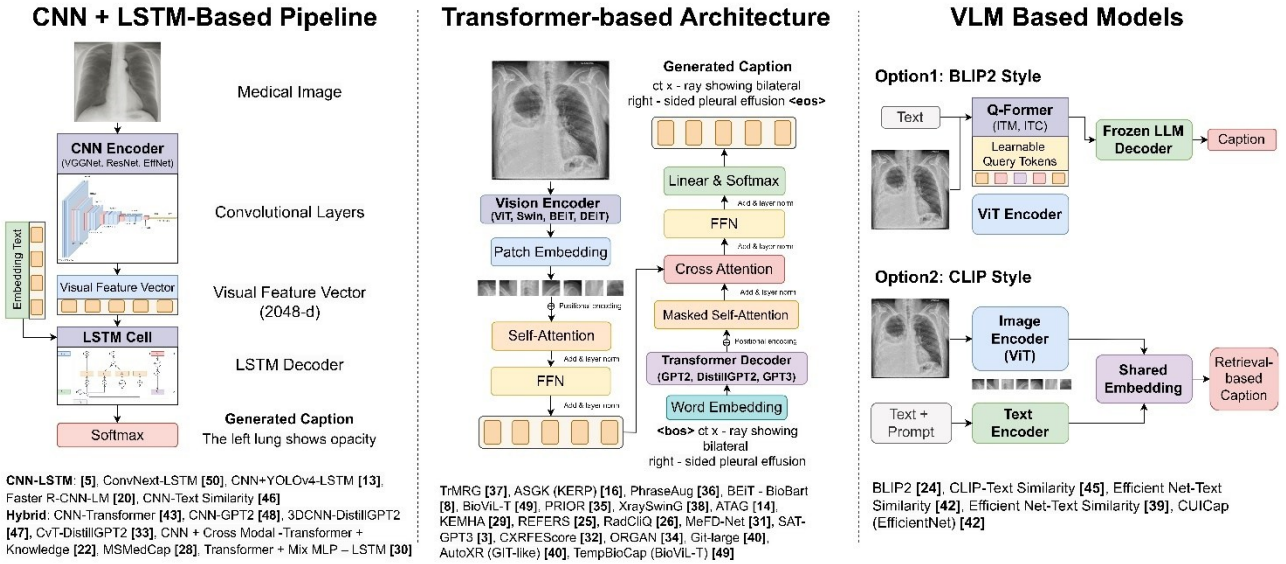
From early CNN-RNN pipelines to more sophisticated vision-language foundation models, the development of medical image captioning models has undergone significant change. Convolutional neural networks (CNNs) like VGG16 or DenseNet-121 were the main tool used in the early methods to extract spatial information from medical images. Decoders based on LSTM or GRU were then used to provide descriptive reports. Notable examples include region-focused designs using YOLOv4 and attention-based LSTM models [13], as well as modifications of the Show-Attend-and-Tell framework for radiology [3].

A major change was brought about by the development of Transformer-based architectures, which used self-attention mechanisms to represent long-range interactions between text sequences and visual areas. Clinical knowledge graphs and structured medical representations were incorporated into transformer-based models like R2GenCMN [14], PPKED [15], and KERP [16] to improve the factual correctness and structural coherence of the reports that were produced. In order to align vision and language embeddings, more recent developments, such as BLIP-2-based models like MedBLIP [17] and MAKEN [18], adopted frozen large language models (like OPT-2.7B, T5), and trained intermediate Q-Former modules, frequently with concept-level supervision from medical ontologies (like UMLS CUIs). Other innovations, such as RGRG [19], introduced region-guided captioning using

Faster R-CNN and GPT-2 decoders for sentence-level anatomical grounding, while longitudinally-aware models like that of Serra et al. [20] incorporated prior and current X-rays to support progression-aware report generation using sentence-anatomy dropout mechanisms. Moreover, few-shot GPT-3.5 prompting and RadGraph serialization were employed by prompt-based models such as Yan et al. [21] to generate style-aware and institution-adaptable reports.

Hybrid architectures have become viable solutions for striking a compromise between global semantic modeling and local detail extraction. For instance, MSMedCap [22] used a dual-encoder configuration with CLIP and SAM (Segment Anything Model) to capture both fine-grained anatomical information and generic semantics, while ViT-GPT2 [23] integrated Vision Transformer encoders with GPT-2 decoders. Two separate Q-Former modules were employed to combine these features into an OPT decoder, allowing for precise and comprehensible caption creation. Hybrid models provide a useful compromise between representational power and computational economy, especially in clinical settings with constrained hardware or data.

Current developments place a strong emphasis on knowledge-enhanced generation, cross-supervision, and domain-specific fine-tuning in order to adapt large-scale foundation models to the medical field. Zhou et al. [24] combined concept-level and caption-level fine-tuning to match BLIP-2 with medical data, while REFERS [25] used cross-supervision between free-text reports and radiographs without the need for structured labels. At the same time, knowledge-aware techniques such as RadGraph-based pipelines [26] and ATAG [14] have shown better explainability and factual basis. Frameworks for style-controlled generation have also made it possible to customize outputs to suit the tastes of radiologists or other institutions. In general, the area is heading toward architectures that are more easily integrated into actual healthcare settings since they are cmodular, generalizable, and clinically interpretable. Medical image captioning has evolved through several architectural paradigms, from early CNN-LSTM models to more recent Transformer-based and Vision-Language Models. Figure 5 provides a comparative overview of these architectures, highlighting their key components and data flow.



CNN-LSTM: [5], ConvNext-LSTM [50], CNN+YOLOv4-LSTM [13], Faster R-CNN-LM [20], CNN-Text Similarity [46]
Hybrid: CNN-Transformer [43], CNN-GPT2 [48], 3DCNN-DistillGPT2 [47], CvT-DistillGPT2 [33], CNN + Cross Modal -Transformer + Knowledge [22], MSMedCap [28], Transformer + Mix MLP – LSTM [30]

TrMRG [37], ASGK (KERP) [16], PhraseAug [36], BEiT - BioBart [8], BioViL-T [49], PRIOR [35], XraySwinG [38], ATAG [14], KEMHA [29], REFERS [25], RadCliQ [26], MeFD-Net [31], SAT-GPT3 [3], CXRFEScore [32], ORGAN [34], Git-large [40], AutoXR (GIT-like) [40], TempBioCap (BioViL-T) [49]

BLIP2 [24], CLIP-Text Similarity [45], Efficient Net-Text Similarity [42], Efficient Net-Text Similarity [39], CUICap (EfficientNet) [42]

Figure 5 Comparison of three medical image captioning architectures (left-to-right): CNN-LSTM, Transformer-based, and Vision-Language Models, showing key components and data flow

3.2 Datasets and Metrics Evaluation (RQ2)

a. Datasets in Medical Image Captioning Research

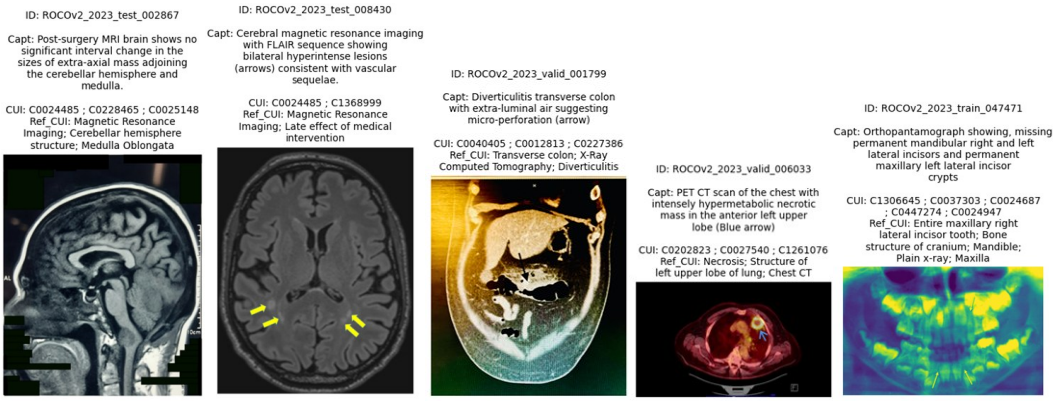


Figure 6. Example of datasets image captioning (Radiology Objects in Context - ROCOv2 [27])

The reviewed papers primarily utilize a handful of publicly available medical datasets, most commonly chest X-ray datasets, but also extend to multi-modal collections including CT, MRI, ultrasound, and PET scans. Below is a structured breakdown:

Table 2 Overview of commonly used datasets for medical image captioning, detailing modalities, dataset size, descriptions, and representative citations

Dataset Name	Modality	Description	# Images / Captions	Used in (Paper Citations)
MIMIC-CXR	Chest X-ray (frontal/ lateral)	Large-scale clinical dataset with DICOM images and free-text reports	~377,110 images / ~227,835 reports	[3], [14], [15], [19], [21], [23], [26], [28], [29], [30], [31], [32], [33], [34], [35], [36]
IU X-Ray	Chest X-ray	Small dataset with paired findings and impressions	7,470 images / 3,955 reports	[14], [15], [19], [23], [28], [29], [30], [31], [33], [34], [36], [37], [38]
ROCO	Multi-modal (X-ray, CT, MRI, PET, etc)	Educational image dataset with UMLS CUIs and captions	81,342 images / 81,342 captions	[17], [22], [39], [40], [41], [42], [43]
ImageCLEF medical Caption	Multi-modal (ROCO-derived)	Competition dataset (2020–2024), enriched with concepts	~90,000 images / ~90,000 captions	[5], [18], [24], [44], [45], [46]
Chest ImaGenome	Chest X-ray	Region-annotated dataset with phrase-level descriptions	~240,000 images / ~100,000 sentences	[20]
MedICaT	CT, MRI, + text	Biomedical publications with images and textual alignment	~217,000 figures / ~1.2M sentences	[22], [41]
ICH (Kaggle)	Brain CT	Labeled CT scans for intracranial hemorrhage detection	~25,000 scans / 25,000 binary labels	[47], [48]
MS-CXR-T	Temporal Chest X-ray	Paired current–prior CXR scans with longitudinal structure	~140,000 images / ~70,000 pairs	[49]
PEIR-Gross	Histopathology	Educational pathology dataset with gross specimen captions	7,442 images / 7,442 captions	[13]
Stomach Histo-pathology Image Captioning	Histopathology (H&E)	Patch-level dataset of H&E-stained gastric biopsy images for stomach adenocarcinoma	34,000 train / 5,700 test images + captions	[50]

b. Natural Language Generation (NLG) Metrics

Natural language generation (NLG) metrics and clinical or semantic accuracy measurements are the two types of metrics that are most frequently employed in evaluation. Surface-level fluency and overlap between generated and reference text are measured using NLG measures. The most commonly reported of these, BLEU (particularly BLEU-1 to BLEU-4) [51] captures accuracy in n-gram overlap. CIDEr [52] rewards rare but informative content with a TF-IDF weighted n-gram similarity score, while ROUGE-L [53] assesses sequence-level recall using the longest common subsequence. METEOR [54] enhances these by adding stemming and synonym matching, which increases its sensitivity to linguistic variance. A learning regression-based metric called BLEURT, which frequently represents human judgment more accurately than BLEU or ROUGE, was also used in some recent works. Various Natural Language Generation-based evaluation metrics have been employed to assess the quality of medical image captioning. These commonly used metrics are summarized in Table 3.

Table 3 Summary of Natural Language Generation based evaluation metrics commonly used in medical image captioning

Metric	Purpose	What it Measures	Observed in
BLEU (1–4)	Fluency & n-gram overlap	Measures precision of n-grams (e.g., unigrams for BLEU-1, 4-grams for BLEU-4); higher = more literal match	Used in nearly all papers; e.g., PPKED [15], MAKEN [18], MedBLIP [17]
ROUGE-L	Recall-oriented match	The longest common subsequence (LCS) between the reference and generated text	Common in all Transformer-based studies
CIDEr	Consensus-based informativeness	Uses TF-IDF weighting of n-grams across multiple references; rewards rare but informative words	MAKEN [18], MedBLIP [17], R2GenCMN [14]
METEOR	Semantic alignment	Considers synonyms, stemming, and word order; more sensitive to linguistic variation	Used in RGRG, PPKED [15], R2GenCMN [14]
SPICE	Scene graph overlap	Compares semantic graphs (objects, attributes, relationships); seldom used in medical setting due to domain mismatch	Limited use
BLEURT	Learned scoring of quality	Pretrained regression metric to model human judgments; measures fluency + semantics	ImageCLEF submissions

c. Semantic and Clinical Relevance Metrics

Newer measures have gained popularity to assess clinical integrity and semantic accuracy. Models like MAKEN and MedBLIP frequently use BERTScore [55], which assesses similarity at the phrase embedding level using contextual embeddings (typically from BioBERT or PubMedBERT). CheXbert F1 measures the degree of alignment between generated reports and reference illness labels for 14 chest disorders in order to conduct clinical evaluation. The more modern RadGraph F1 is particularly helpful for grounding accuracy since it assesses the overlap of medical entities and their relationships (e.g., "opacity located_at left lung"). RadCliQ, a composite metric that combines BLEU, CheXbert, BERTScore, and RadGraph F1, was developed in 2023 and was shown to have a strong correlation with radiologist preferences. Other specialized metrics include RadRQI-F1 (focusing on abnormality-attribute alignment) and Anatomy Sensitivity Ratio (AS-Ratio), which measures how well model-generated sentences align with visual regions of interest. A summary of these semantic and clinical relevance metrics is provided in Table 4.

Table 4 Summary of semantic and clinical relevance metrics commonly used in medical image captioning

Metric	Purpose	What it Measures	Used in
BERTScore	Semantic similarity	Uses contextual embeddings (e.g., BERT, PubMedBERT) to align predicted vs. reference captions	MAKEN [18], MedBLIP [17], MSMedCap [22]
RadGraph F1	Clinical entity/relation match	Measures overlap of radiographic findings and relations (e.g., "opacity located_at left lung")	ATAG [14]
CheXbert F1	Clinical concept match	Matches presence/absence of 14 diseases; evaluates disease detection consistency	KERP [16]
RadCliQ	Correlated with radiologist judgment	Combines BLEU, CheXbert, BERTScore, RadGraph F1 into one regression metric	Proposed by Yu [26]
RadRQI-F1	Abnormality + attribute overlap	Evaluates presence and modifiers (e.g., size, location) of clinical findings	ATAG [14]
Anatomy Sensitivity Ratio (AS-Ratio)	Grounding quality	Proportion of sentences that correctly reference image regions	RGRG [19]

3.3 Performance Comparison (RQ3)

The systematic literature review's findings offer insightful information about the state-of-the-art methods, difficulties, and potential paths in medical image captioning. A number of criteria, such as model designs, datasets, performance measures, and important discoveries, were used to examine the chosen papers. In terms of evaluation criteria like BLEU, ROUGE, CIDEr, and METEOR, Transformer-based techniques perform better than CNN-LSTM based designs, according to a comparative examination of several models [9][10]. A summary of the performance comparison of several models across widely used datasets is shown in the Table 5.

Table 5 Performance comparison of medical image captioning models across various datasets and evaluation metrics

Model	Dataset	BLEU	ROUGE	CIDEr	METEOR	BERT-Score	Notes	Strength	Limitation
MAKEN [18]	ROCO (CLEF 2023)	0.226	0.252	0.287	–	0.639	Adapter tuning + MKE	High semantic fidelity; efficient tuning	Needs external medical knowledge
MedBLIP [17]	ROCO (CLEF 2023)	0.221 (BLEU-1)	0.247	0.220	0.098	0.617	BLIP-2 + CUI-based fine-tuning	Grounded medical concept representation	Complex architecture; resource-intensive
ViT-GPT2 [23]	IU X-Ray	0.226 (BLEU-4)	0.433	–	0.385	–	Vision Transformer + GPT-2	Fluent generation with autoregressive decoding	Lacks clinical concept alignment
PPKED [15]	IU X-Ray	0.168 (BLEU-4)	0.376	0.351	-	–	Prior/posterior KG + distillation (CNN-RNN)	Structured planning enhances factuality	Depends on historical data
R2GenCMN [21]	MIMIC-CXR	0.184	-	-	-	0.378	RadGraph (Clinical graph + multi-view)	Good for handling radiographic complexity	Missing metrics; limited dataset validation
RGRG [19]	MIMIC-CXR	0.126 (BLEU-4)	0.264	0.495	0.168	–	Region-guided GPT-2 decoder	Strong alignment with anatomical regions	Low BLEU; long inference time

The MAKEN model [18], based on a BLIP-2 backbone with adapter tuning and Medical Knowledge Enhancement (MKE) loss, achieved consistently high performance on the

ImageCLEFmedical 2023 Caption Prediction task. It reported strong scores across BLEU (0.226), ROUGE (0.252), and CIDEr (0.287), with the highest observed BERTScore (0.639). The MKE loss, which reinforces learning of clinically significant terms based on concept frequency, contributes to the model's strong semantic alignment with ground-truth reports. Its adapter-based design allows efficient domain adaptation without updating the frozen LLM decoder (OPT-2.7B).

MedBLIP [17], another BLIP-2-derived model, also demonstrated competitive performance: BLEU-1 (0.221), ROUGE (0.247), CIDEr (0.220), METEOR (0.098), and BERTScore (0.617). It introduces a two-phase training pipeline, starting with CUI-based concept supervision followed by fine-tuning on full captions, which improves its ability to generate factually grounded and domain-aware descriptions. Despite slightly lower CIDEr and BERTScore values compared to MAKEN, MedBLIP's overall performance and clinical relevance remain strong.

ViT-GPT2 [23], a hybrid model combining a Vision Transformer encoder and a GPT-2 decoder, achieved BLEU-4 (0.226) and ROUGE (0.433) on IU X-Ray. Its autoregressive decoder ensures fluent generation, while the Transformer encoder preserves high-resolution spatial detail. However, the model lacks medical concept alignment or knowledge-aware modules, which may limit its factual specificity compared to medically tuned alternatives.

PPKED [15] employs a cross-domain Transformer framework that integrates prior and posterior knowledge via two modules e.g PoKE (posterior knowledge explorer) and PrKE (prior knowledge explorer). On IU X-Ray, it achieved BLEU-4 (0.168), ROUGE (0.376), and CIDEr (0.351). Although PPKED predates modern Q-Former or BLIP architectures, its strength lies in explicitly structured content planning and knowledge distillation, which enhances its performance on datasets with structured reports.

R2GenCMN [21], designed for the MIMIC-CXR dataset, incorporates a clinical knowledge graph and multi-view fusion (frontal and lateral views). It reported a BLEU score of 0.184 and BERTScore of 0.378, although ROUGE and CIDEr values were not reported. The model's use of entity-aware attention and structural supervision allows for more context-aware generation, particularly in handling radiographic complexity and spatial alignment.

RGRG (Region-Guided Radiology Generation) [19] represents a unique anatomy-grounded captioning approach. Using a Faster R-CNN to detect anatomical ROIs and GPT-2 to generate one sentence per region, it scored BLEU-4 (0.126), ROUGE (0.264), CIDEr (0.495), and METEOR (0.168) on MIMIC-CXR. Despite the relatively low BLEU score, it produced the highest CIDEr and METEOR values among these models, reflecting strong alignment with human-readable and clinically informative text. Its modular design supports interpretability and region-level traceability, which is particularly useful for clinician-end usability, though potentially at the cost of holistic fluency.

Transformers typically require large datasets to perform well, which poses a challenge in the medical domain due to privacy and limited annotations. In contrast, CNNs are more effective with smaller datasets, especially when pre-trained on natural images and fine-tuned for medical use. Recent hybrid models like ConTrans and CTranS combine the strengths of both architectures, using CNNs for local feature extraction and Transformers for capturing global context, resulting in improved performance in medical image segmentation and captioning tasks. [11, 15].

In a variety of medical image processing applications, hybrid models have been demonstrated to perform better than both Transformer and pure CNN models. For example, Transformer-based models were able to capture long-range dependencies between various breast views in mammograms used for

breast cancer diagnosis, and Transformer-based models produced more accurate and comprehensible captions for multimodal tasks such as combining retinal fundus images and OCT scans [58], [59].

3.4 Summary Table of Reviewed Study

This table summarizes the key characteristics of the 40 reviewed studies, including model architectures (CNN-LSTM, Transformer-based, or Hybrid), datasets utilized, evaluation metrics applied, and primary findings. It offers a comparative overview that supports the synthesis of results and directly addresses Research Questions RQ1 to RQ3, concerning architectural trends, dataset selection, and model performance outcomes can see at Table 6

Table 6 Summary of Reviewed Studies in Medical Image Captioning (2019–2024)

Model	Dataset	Notes	Result	Strength	Limitation
CLIP-Text Similarity [45]	ImageCLEF medical Caption	Multi-stage approach combining image classification and CLIP for medical image captioning.	BERT: 0.5776, ROUGE: 0.15392, BLUERT: 0.27136, METEOR: 0.15404, CIDEr: 0.06970, CLIP Score: 0.10482.	Combines classification with CLIP to improve accuracy and text relevance.	Initial image classification is limited; struggles with hard-to-classify images.
Efficient Net-Text Similarity [42]	ROCOv2	Uses CUI-based multi-label classification and feature similarity to generate captions.	BERT: 0.5673 (Run 5, ViT32, 10 groups).	Enhances consistency with CUI-based classification; easier for medical professionals to interpret.	Complex model with high computational needs; requires CUI-labeled data; difficult to scale across diverse medical imaging modalities.
BioViL-T [49]	MS-CXR-T	Uses temporal info from medical images to improve tasks like disease progression classification/reporting.	BLEU-4: 9.2 ± 0.3, ROUGE-L: 29.6 ± 0.1, CHEXBERT: 31.7 ± 1.0.	Effectively uses temporal features to enhance clinical report quality.	Requires historical data; complex model.
CNN-Transformer [43]	ROCO	Combines CNN and Transformer for generating medical image descriptions across multiple modalities.	Accuracy: 0.7628, BLEU-1: 0.5387.	Strong integration of CNN and Transformer; high accuracy and cross-modality support.	High computational demand; some descriptions lack detail due to data diversity limitations.
CNN-LSTM [5]	ImageCLEF medical Caption	Uses visual and semantic features (via CNN & multi-label classifier) with LSTM for medical image captions.	BLEU: 42.28% (beam search k=5). Outperformed ImageSem (BLEU: 25.70%).	Combines visual & semantic features effectively; beam search improves output; good ablation study.	Struggles with long/complex captions; performance sensitive to image quality; rare terms poorly handled.
Multimodal Masked Auto Encoder (M3AE) [41]	ROCO, VQA-RAD, SLACK, VQA-2019	Self-supervised learning using multi-modal masked autoencoders to align vision and language representations.	VQA-RAD: 77.01, SLACK: 83.25, VQA-2019: 79.87, ROCO: 78.50.	No need for large labeled data; strong cross-modal understanding; relatively simple architecture.	Needs further optimization for complex modalities; limited by smaller datasets.
PRIOR [35]	MIMIC-CXR	Joint learning of global-local representations from medical images and reports for accurate and granular clinical understanding.	CIDEr: 82.1, BLEU-4: 20.5.	Improves granular clinical representation.	High complexity in global-local alignment.
CNN-LSTM [50]	Histo-pathological Dataset	Combines ConvNext-Large + PVT_v2_b5 for vision, and BioLinkBERT-Large for captioning histopathological images.	BLEU1: 0.576, BLEU2: 0.455, BLEU3: 0.411, BLEU4: 0.309.	High accuracy and consistency in histopathological image captions.	Requires heavy computation; performance may drop with highly varied image datasets.

Model	Dataset	Notes	Result	Strength	Limitation
ORGAN [34]	IU-Xray, MIMIC-CXR	Uses tree-based reasoning and observation graphs to generate consistent radiology reports.	+12.4% report consistency, +10.5 CIDEr over baseline.	Improves report consistency significantly.	Tree-based reasoning adds complexity and longer training time.
XraySwinG [38]	NIH Chest X-ray, PT-BR (derived from IU X-ray) dataset	Combines Swin Transformer (vision) and GPT-2 (language) with cross-attention for multilingual reporting.	ROUGE-L: 0.748, METEOR: 0.741 *(PT-BR)*; ROUGE-L: 0.404, METEOR: 0.393 *(NIH)*.	Strong multimodal integration; supports multilingual reporting; captures global image features well.	Limited availability of high-quality, multilingual medical report datasets.
CvT-DistillGPT2 [33]	IU-Xray, MIMIC-CXR	Uses CvT as image encoder and DistillGPT2 as text decoder for automated CXR report generation.	BLEU-1: 0.750, BLEU-2: 0.500, BLEU-3: 0.632, BLEU-4: 0.707, METEOR: 0.768, ROUGE-L: 0.643, CIDEr: 0.785.	Efficient transformer architecture; performs well in resource-limited settings.	Limited to CXR; requires optimization for broader generalization.
GIT-large [40]	ROCOv2	Part of ImageCLEFmedical 2024: concept detection using CNN ensemble; caption prediction using GIT models.	*Concept Detection*: F1-score 0.6374; *Caption Prediction*: BERTScore 0.5769	Strong ensemble for concept detection; GIT performs reasonably for captioning.	Hierarchical model underperformed; ensemble is resource-intensive; GIT still lacks medical precision.
3DCNN-DistillGPT2 [47]	ICH (Kaggle)	Captioning CT images for intracerebral hemorrhage using 3D CNNs and DistillGPT-2 for spatially-aware descriptions.	BLEU: 0.35, METEOR: 0.23, ROUGE-L: 0.56, Cosine Similarity (ClinicalBERT): 0.78.	Captures 3D spatial info well; efficient with DistillGPT-2; supports automated ICH diagnosis.	High computational cost; limited generalization to external datasets.
CNN-GPT2 [48]	ICH (Kaggle)	Uses pretrained CNN classifiers and GPT-2 for sequential brain CT image captioning for ICH.	BLEU-4: 0.17, ROUGE-L: 0.29, CIDEr: 0.27; Embedding Avg: 0.71, Greedy Matching: 0.63.	Strong CNN-GPT2 combo; DenseNet-121 yields best performance.	GPT-2 is resource-intensive; less effective on rare abnormalities.
ASGK (KERP) [16]	CX-CHR, IU X-Ray, COV-CTR	Uses auxiliary visual and linguistic signals to enhance encoder-decoder Transformer performance for report generation.	CIDEr: 89.4 (↑ 7.6 over baseline).	Improves benchmark scores using enriched auxiliary features.	High complexity; requires detailed auxiliary signals.
PhraseAug [36]	IU-Xray, MIMIC-CXR	Two-stage model using phrasebook and clinical history to improve medical report generation, addressing visual-linguistic bias.	*IU-Xray*: BLEU-4: 0.231, METEOR: 0.218, ROUGE-L: 0.431, CIDEr: 0.665. *MIMIC-CXR*: BLEU-4: 0.184, METEOR: 0.208, ROUGE-L: 0.353, CIDEr: 0.280.	Enhances cross-modal alignment with phrasebook; adapts to writing styles; uses clinical history for contextual relevance.	Phrase selection may miss rare but critical terms; handling underrepresented abnormalities is still a challenge; complex phrasebook construction.
CXRFEScore [32]	MIMIC-CXR	Leverages LLMs for fact extraction and uses BERT-based encoding to enhance radiology report representation.	Fact extraction F1: 75.8, fact encoding F1: 68.4.	Boosts factual representation of radiology text; effective use of LLM + BERT pipeline.	Dependent on LLM fact extraction quality.
TrMRG [37]	IU-Xray	Transformer-based model (ViT encoder + MiniLM decoder) for generating radiology reports from X-rays.	BLEU-1: 0.483, BLEU-4: 0.168, METEOR: 0.376, ROUGE-L: 0.351	Handles data imbalance better than CNN-RNN; better performance in ablation study	Struggles with rare abnormalities; requires large training data
Efficient Net-Text Similarity [39]	ImageCLEF medical 2024	Ensemble of 5 EfficientNet B0 models, each specialized in a subgroup of medical concepts.	F1-score: 0.59876 (run 1), 0.52921 (run 2);	Tackles data imbalance with subgroups; improves recall for certain classes;	Lower precision due to false positives; minimal improvement

Model	Dataset	Notes	Result	Strength	Limitation
Vision Diagnostor-BioBART [44]	ImageCLEF medical 2024	Transformer-based models (BioBART, ClinicalT5, Q-BioMistral) used for diagnostic caption generation.	BERTScore: 0.6267, ROUGE: 0.2452, CIDEr: 0.2243;	High scores on BERTScore and CIDEr; advanced Transformer architecture; object-level feature exploration	Large models (e.g., Q-BioMistral) don't always outperform smaller ones; weak on long captions
BEiT - BioBart [8]	ImageCLEF medical 2024	Combines concept detection (Swin-V2) and enhanced attention for medical captioning.	F1 *(concept)*: 0.58944 (val), 0.61998 (test); BERTScore *(caption)*: 0.60589 (val), 0.5794 (test);	Concept integration improves relevance; significant boost in concept detection and captioning performance	Struggles with complex image context; post-processing not always effective
CNN-Text Similarity [46]	ROCOv2	Uses DenseNet-121, MobileNetV2, and ResNet-50 for multi-label concept detection in medical images.	Best: ResNet-50 (F1: 0.181); MobileNetV2 (F1: 0.178); DenseNet-121 (F1: 0.114)	ResNet-50 strong in concept detection; MobileNetV2 efficient for low-resource settings	DenseNet-121 prone to overfitting; more regularization needed
MeFD-Net [31]	IU X-ray, MIMIC-CXR	Mimics clinical multi-expert diagnosis with fusion modules integrating visual and text features.	*IU X-ray*: BLEU-4: 0.190, ROUGE-L: 0.413, METEOR: 0.248; *MIMIC-CXR*: BLEU-4: 0.148, ROUGE-L: 0.336, METEOR: 0.155	Accurate via multi-expert modeling; modular and extendable.	Performance drops with imbalanced multimodal data; high computational demand.
CNN + YOLOv4 - LSTM [13]	PEIR	YOLOv4 for feature extraction and soft attention-based LSTM for generating medical image captions.	BLEU: 81.78%, METEOR: 78.56%;	Efficient object detection with YOLOv4; accurate descriptive captions.	Less suitable for complex medical images with subtle or multiple anomalies.
SAT-GPT3 [3]	MIMIC-CXR, IU-Chest	Combines Show-Attend-Tell (SAT) with GPT-3 for generating cohesive and clinically meaningful captions.	Accuracy: 0.861, Precision: 0.445, Recall: 0.351, F1: 0.369. CIDEr: 1.989, ROUGE-L: 0.480, BLEU-4: 0.418	Combines powerful models for rich captions; integrates 2D heatmaps for visual support.	Slightly lower CIDEr due to weak n-gram tuning; complex architecture with high computational cost.
Faster R-CNN-LM [20]	Chest ImaGenome	Uses longitudinal X-ray representations (past & present) via CNN-RNN to generate anatomy-controllable reports.	BLEU: 21.4, CIDEr: 77.3;	Fine-grained control over anatomical focus in reports.	Requires longitudinal image data, which is not always available in practice.
Transformer + Mix MLP – LSTM [30]	IU-Xray, MIMIC-CXR	Combines label information with co-attention and hierarchical LSTM for structured report generation.	BLEU-4: 0.227 *(IU)*, 0.155 *(MIMIC)*; METEOR: 0.244 *(IU)*, 0.201 *(MIMIC)*; ROUGE-L: 0.435 *(IU)*	Label-guided attention improves report accuracy; good performance on multi-label classification.	Needs high-quality labeled data; struggles with rare outliers.
ATAG [14]	IU-Xray, MIMIC-CXR	Uses Attributed Abnormality Graph and Graph Attention Networks to model spatial relationships between abnormalities.	Clinical accuracy: 84.3% vs SOTA 79.5%.	Captures abnormality-attribute relations effectively; improves clinical accuracy.	High computational demand due to complex graph modeling.
KEMHA [29]	IU-Xray, MIMIC-CXR	Integrates general and specific knowledge into multi-head attention using knowledge graphs for richer report generation.	BLEU-4: 24.6, CIDEr: 87.2; improves clinical understanding and terminology use.	Enhances clinical accuracy and medical language in reports.	Requires large, well-structured external knowledge sources that are hard to integrate.
RadCliQ [26]	MIMIC-CXR	Evaluates alignment between automated metrics and radiologist judgment; introduces RadGraph F1 and RadCliQ metrics.	RadCliQ Kendall's tau: 0.615 (highest), RadGraph F1: 0.531, BERTScore: 0.518; BLEU worst at 0.462.	Better alignment with clinical evaluation than traditional metrics; captures critical clinical errors.	BLEU still underperforms; gap remains between automated and radiologist-level report quality.

Model	Dataset	Notes	Result	Strength	Limitation
CNN + Cross Modal - Transformer + Knowledge [22]	COCO, ROCO, MedICaT	Uses dual encoders (CLIP + SAM) with mixed semantic pre-training; incorporates Q-Former for aligning vision and language features.	BLEU-1: 48.1, ROUGE: 15.4, CIDEr: 57.5;	Strong global-local feature extraction; effective cross-modal representation; substantial performance boost.	Sensitive to noisy captions; lacks deep evaluation in strictly medical domains; vulnerable to localized errors.
MSMedCap [28]	IU-Xray, MIMIC-CXR	Combines teacher-student model with encoder-decoder network; enhances semantic alignment and output fluency via decoding supervision.	BLEU-4: 0.116, METEOR: 0.150, ROUGE-L: 0.292 *(MIMIC-CXR)*;	Better handling of small lesion detail and long report fluency; improved output structure.	Computationally intensive; less adaptable across highly varied datasets.
REFERS [25]	-	Uses cross-supervision between radiograph images and associated free-text reports to improve representation learning.	CIDEr improved by 9.3 over baseline on multiple radiograph datasets; better generalization to unseen domains.	Reduces reliance on manual labels; supports domain shift handling.	Relies on quality of free-text reports used as supervision.
BLIP2 [24]	ImageCLEF 2023	Utilizes BLIP-2 with ViT-g encoder and OPT-2.7B decoder; two-stage fine-tuning and post-processing applied for better captions.	BERTScore: 0.6281, ROUGE: 0.2401, BLEURT: 0.3209, CIDEr: 0.2377, BLEU: 0.1846, METEOR: 0.0873	Strong across multiple metrics; post-processing improves quality; well adapted to medical imaging tasks.	Struggles with complex anomaly detection; prone to misclassifying anatomical regions in some cases.

3.5 Challenges Identified (RQ4)

Despite the notable advancements driven by deep learning in medical image captioning, significant challenges persist, chief among them being the limited availability of annotated datasets. This issue is especially pronounced for imaging modalities beyond chest X-rays, such as MRI, CT, and histopathology, where public, captioned datasets are scarce. Annotating medical images demands specialized domain expertise and is both time-consuming and costly, often resulting in small, institution-specific corpora that hinder model generalization. For instance, Elbedwehy et al. [50] resorted to generating synthetic descriptions for histopathology images due to the absence of publicly available captioned datasets.

Another persistent issue is clinical interpretability and explainability. While Transformer-based models offer superior fluency, it is often unclear how predictions align with medical findings. Several models attempt to bridge this gap through region-grounded generation [19], RadGraph serialization [21], or anatomical attention maps [22], but these remain limited in clinical usability without extensive validation.

Computational complexity also poses a major barrier. High-performing models like BLIP-2 [17] and MAKEN [18] require large memory and inference time, which limits their deployability in resource-constrained clinical environments. Moreover, these models are data-hungry, with optimal performance only achieved when pre-trained on massive general-domain datasets and then carefully adapted, a step not feasible for many healthcare institutions.

Lastly, domain shifts between training and target data present a challenge for generalization. Variability in imaging protocols, language styles, and institutional reporting conventions often causes performance degradation in cross-domain evaluation, as noted in studies on transfer learning [24] and few-shot prompting [21].

3.6 Key Findings

From the review of recent literature, several consistent findings emerge. First, Transformer-based models, particularly those using vision-language pretraining (e.g., BLIP-2), outperform traditional CNN-RNN architectures across all major metrics e.g BLEU, ROUGE, CIDEr, and BERTScore. Their strength lies in modeling global dependencies and integrating rich semantics.

Second, the introduction of domain knowledge, whether through CUIs (e.g., MedBLIP [17]), prior/posterior report reasoning (PPKED) [15], or structured knowledge graphs (e.g., RadGraph [21], ATAG [14]), significantly improves factual accuracy and clinical relevance. These models better identify abnormalities and produce consistent terminology.

Third, hybrid and region-aware approaches demonstrate that modular interpretability can be embedded without sacrificing overall fluency. For instance, RGRG [19] achieved high METEOR and CIDEr scores while maintaining sentence-level alignment with anatomical ROIs, offering a path toward explainable AI.

Fourth, evaluation strategies are maturing. While BLEU and ROUGE remain ubiquitous, metrics like RadGraph F1, CheXbert F1, and RadCliQ are increasingly adopted to reflect clinical correctness rather than just n-gram overlap. This shift aligns with findings from Yu et al. [26], which showed weak correlation between BLEU and radiologist-assessed report quality.

3.7 Potential Applications (RQ5)

a. Automated Radiology Report Generation

The automated creation of radiology reports, which can save time and minimize human mistake, is one promising use of medical image captioning. Medical images can be analyzed by vision-language models such as 3D-CT-GPT, which can produce precise text summaries that include the location and severity of diseases [18, 19]. Additionally, memory-driven transformer models have demonstrated promise in producing comprehensive reports that include image-text attention mappings and medically important phrases, supporting radiologist training and guaranteeing consistent reporting.

b. Clinical Decision Support Systems

Vision-language models (VLMs) have shown significant promise in enhancing Clinical Decision Support Systems (CDSS) by interpreting medical images and generating relevant textual insights. These models integrate visual data (e.g., X-rays, MRIs, CT scans) with natural language understanding to deliver accurate, context-aware responses that assist clinicians in making informed decisions. One notable application is Medical Visual Question Answering (Med-VQA), where models like VQA-RAD, PathVQA, and more recently MedBLIP and Med-VQA are trained to answer complex clinical questions grounded in radiology images. By combining image features with domain-specific language models, they provide precise and explainable answers that support both diagnostic reasoning and treatment planning [62].

To improve diagnostic accuracy, these models can also be tailored for particular tasks, including recognizing anatomical features in multi-modal pictures or spotting anomalies in mammograms [19, 20]. These systems can offer contextually relevant information by utilizing sophisticated attention processes and sizable databases, which lowers the possibility of diagnostic errors and enhances patient outcomes.

c. Enhanced Patient Communication

Through the creation of comprehensible summaries of radiological results, medical image captioning can help enhance patient communication. Patients can have more informed conversations with clinicians if they have a better understanding of their diseases thanks to these summaries. For instance, models such as those reported in recent studies might produce easily understandable explanations of observations, such as the location and severity of diseases [65].

Furthermore, visual aids like heatmaps or highlighted places of interest can be produced using vision-language models to go along with text-based reports. This multimodal strategy can improve patient comprehension and involvement in their treatment.

d. Medical Education and Training

Vision-language models are useful resources for medical education because they can produce precise, in-depth descriptions of medical images. Radiology residents can investigate a variety of cases and abnormalities by using these models to generate training datasets. For instance, models such as those reported in recent research can produce captions for vast collections of medical images, assisting learners in accurately recognizing and characterizing diseases [66].

Additionally, trainees can practice picture interpretation and report generation in a controlled setting by using these models to mimic clinical events [13]. This can assist novice radiologists become more proficient in image interpretation and lower their learning curve.

e. Research and Public Health Applications

Additionally, vision-language models can be very important for public health and medical research. These models can assist researchers in finding patterns and trends that might not be visible through manual analysis by automating the investigation of big databases of medical images. For instance, multi-modal medical imaging, like MRI and CT scans, can be analyzed using models similar to those reported in recent studies to find disease biomarkers or track the course of a disease over time [63].

By producing textual summaries at scale, medical image captioning models may extract structured insights from millions of archival images. In large health systems, this facilitates phenotypic discovery, risk profiling, and retrospective study of illness trends. For instance, REFERS [25] opened the door for automated population health surveillance and public health research by using cross-supervision to extract radiographic ideas from large CXR corpora.

4 Conclusion

This systematic literature review has comprehensively addressed five research questions (RQ1–RQ5) related to medical image captioning using deep learning. First, regarding RQ1, the review classified and compared three primary architecture families: CNN-LSTM models, Transformer-based models, and hybrid approaches. Transformer-based models, particularly those incorporating cross-modal alignment and large-scale pretraining, have emerged as the dominant paradigm due to their superior semantic understanding and flexibility. In response to RQ2, the study identified key datasets, such as IU X-Ray, MIMIC-CXR, ROCOv2, and CLEF 2023, as the most widely used benchmarks, detailing their imaging modalities, annotation formats, and scale diversity. Addressing RQ3, performance metrics including BLEU, ROUGE, CIDEr, METEOR, and BERTScore were systematically compiled and compared, revealing that Transformer-based models consistently

outperform CNN-LSTM baselines in generating clinically coherent and linguistically fluent reports. For RQ4, this review synthesized major challenges such as the lack of factual grounding, domain shift issues, limited explainability, and dataset imbalance, which often compromise the clinical reliability of generated captions. Lastly, in response to RQ5, the review proposed future directions emphasizing explainable and knowledge-grounded captioning, few-shot and cross-institutional learning, and the integration of structured clinical knowledge into generative pipelines.

Overall, this review contributes an updated synthesis of the state-of-the-art, a taxonomy of model architectures, a comparative performance framework, and a detailed analysis of clinical challenges and future needs. These insights are expected to inform and guide future research toward safer, more accurate, and clinically relevant applications of image captioning in medical AI.

Acknowledgement

The author would like to thank Lembaga Pengelola Dana Pendidikan (LPDP) for helping to make this review possible.

References

[1] G. Reale-Nosei, E. Amador-Domínguez, and E. Serrano, "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," *Med. Image Anal.*, vol. 97, no. August 2023, 2024, doi: 10.1016/j.media.2024.103264.

[2] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, "Diagnostic captioning: a survey," *Knowl. Inf. Syst.*, vol. 64, no. 7, pp. 1691–1722, 2022, doi: 10.1007/s10115-022-01684-7.

[3] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V Dylov, "Medical image captioning via generative pretrained transformers," *Sci. Rep.*, pp. 1–12, 2023, doi: 10.1038/s41598-023-31223-5.

[4] H. Sharma and D. Padha, "Domain-specific image captioning : a comprehensive review," *Int. J. Multimed. Inf. Retr.*, vol. 13, no. 2, pp. 1–27, 2024, doi: 10.1007/s13735-024-00328-6.

[5] D. R. Beddiar, M. Oussalah, T. Seppänen, and R. Jennane, "ACapMed: Automatic Captioning for Medical Imaging," *Appl. Sci.*, vol. 12, no. 21, pp. 1–24, 2022, doi: 10.3390/app122111092.

[6] R. Thirunavukarasu and E. Kotei, "A comprehensive review on transformer network for natural and medical image analysis," *Comput. Sci. Rev.*, vol. 53, no. April, 2024, doi: 10.1016/j.cosrev.2024.100648.

[7] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 5999–6009, 2017.

[8] N. N. Y. Nguyen, H. L. Tu, P. D. Nguyen, T. N. Do, T. M. Thai, and T. B. Nguyen-tat, "DS @ BioMed at ImageCLEFmedical Caption 2024 : Enhanced Attention Mechanisms in Medical Caption Generation through Concept Detection Integration ⋆," 2024.

[9] F. A. Zahra and R. J. Kate, "Obtaining clinical term embeddings from SNOMED CT ontology," *J. Biomed. Inform.*, vol. 149, no. November 2023, 2024, doi: 10.1016/j.jbi.2023.104560.

[10] H. Zhou *et al.*, "Complementary and Integrative Health Information in the literature: Its lexicon and named entity recognition," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 2, pp. 426–434, 2024, doi: 10.1093/jamia/ocad216.

[11] C. Wohlin and R. Prikladniki, "Systematic literature reviews in software engineering," *Inf. Softw. Technol.*, vol. 55, no. 6, pp. 919–920, 2013, doi: 10.1016/j.infsof.2013.02.002.

[12] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010, doi: 10.1016/j.ijsu.2010.02.007.

[13] P. Ravinder and S. Srinivasan, "Automated Medical Image Captioning with Soft Attention-

Based LSTM Model Utilizing YOLOv4 Algorithm," *J. Comput. Sci.*, vol. 20, no. 1, pp. 52–68, 2024, doi: 10.3844/jcssp.2024.52.68.

[14] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation," *IEEE Trans. Med. Imaging*, vol. 42, no. 8, pp. 2211–2222, 2023, doi: 10.1109/TMI.2023.3245608.

[15] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13748–13757. doi: 10.1109/CVPR46437.2021.01354.

[16] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *World Wide Web*, vol. 26, no. 1, pp. 253–270, 2023, doi: 10.1007/s11280-022-01013-6.

[17] V. T. Phan and K. T. Nguyen, "EasyChair Preprint MedBLIP : Multimodal Medical Image Captioning Using BLIP MedBLIP : Multimodal medical image captioning using," 2024.

[18] B. Yang, Z. Ye, H. Wang, H. Zheng, and S. International, "MAKEN : IMPROVING MEDICAL REPORT GENERATION WITH ADAPTER TUNING AND KNOWLEDGE ENHANCEMENT IN VISION-LANGUAGE FOUNDATION MODELS ADSPLAB , School of Electronic and Computer Engineering , Peking University , Shenzhen , China Shenzhen Institute of Advanced Te," *2024 IEEE Int. Symp. Biomed. Imaging*, pp. 1–5, 2024, doi: 10.1109/ISBI56570.2024.10635421.

[19] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and Explainable Region-guided Radiology Report Generation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2023-June, pp. 7433–7442, 2023, doi: 10.1109/CVPR52729.2023.00718.

[20] F. Dalla Serra, C. Wang, F. Deligianni, J. Dalton, and A. Q. O'Neil, "Controllable Chest X-Ray Report Generation from Longitudinal Representations," *Find. Assoc. Comput. Linguist. EMNLP 2023*, pp. 4891–4904, 2023, doi: 10.18653/v1/2023.findings-emnlp.325.

[21] B. Yan *et al.*, "Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting," *Find. Assoc. Comput. Linguist. EMNLP 2023*, pp. 14676–14688, 2023, doi: 10.18653/v1/2023.findings-emnlp.977.

[22] Z. Zhang *et al.*, "Sam-Guided Enhanced Fine-Grained Encoding with Mixed Semantic Learning for Medical Image Captioning," *ICASSP 2024 - 2024 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1731–1735, 2024, doi: 10.1109/icassp48485.2024.10446878.

[23] S. Raminedi, S. Shridevi, and D. Won, "Multi-modal transformer architecture for medical image analysis and automated report generation," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-69981-5.

[24] W. Zhou *et al.*, "Transferring Pre-Trained Large Language-Image Model for Medical Image Captioning," *CEUR Workshop Proc.*, vol. 3497, pp. 1776–1784, 2023.

[25] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nat. Mach. Intell.*, vol. 4, no. 1, pp. 32–40, 2022.

[26] F. Yu, M. Endo, R. Krishnan, and C. P. Langlotz, "Article Evaluating progress in automatic chest X-ray radiology report generation Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, p. 100802, 2023, doi: 10.1016/j.patter.2023.100802.

[27] J. Rückert *et al.*, "ROCOv2 : Radiology Objects in COntext Version 2 , an Updated Multimodal Image Dataset," pp. 1–15, 2024, doi: 10.1038/s41597-024-03496-6.

[28] S. Zhang, Q. Han, J. Li, Y. Sun, and Y. Qin, "A medical report generation method integrating teacher–student model and encoder–decoder network," *Biomed. Signal Process. Control*, vol. 94, no. March, 2024, doi: 10.1016/j.bspc.2024.106251.

[29] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, 2022, doi:

10.1016/j.media.2022.102510.
[30] S. Sun, Z. Mei, X. Li, T. Tang, Z. Su, and Y. Wu, "A label information fused medical image report generation framework," *Artif. Intell. Med.*, vol. 150, no. October 2022, 2024, doi: 10.1016/j.artmed.2024.102823.
[31] R. Ran *et al.*, "MeFD-Net: multi-expert fusion diagnostic network for generating radiology image reports," *Appl. Intell.*, pp. 11484–11495, 2024, doi: 10.1007/s10489-024-05680-y.
[32] P. Messina, R. Vidal, D. Parra, A. Soto, and V. Araujo, "Extracting and Encoding: Leveraging Large Language Models and Medical Knowledge to Enhance Radiological Text Representation," *Find. Assoc. Comput. Linguist. ACL 2024*, no. Section 4, pp. 3955–3986, 2024, [Online]. doi: 10.18653/v1/2024.findings-acl.236.
[33] K. Kar, "Medical Image Captioning using CvT and," *2024 Second Int. Conf. Adv. Inf. Technol.*, vol. 1, pp. 1–6, 2024, doi: 10.1109/ICAIT61638.2024.10690339.
[34] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 8108–8122, 2023, doi: 10.18653/v1/2023.acl-long.451.
[35] P. Cheng, L. Lin, J. Lyu, Y. Huang, W. Luo, and X. Tang, "PRIOR: Prototype Representation Joint Learning from Medical Images and Reports," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 21304–21314, 2023, doi: 10.1109/ICCV51070.2023.01953.
[36] X. Mei, L. Yang, D. Gao, X. Cai, J. H. Fellow, and T. Liu, "PhraseAug : An Augmented Medical Report Generation Model with Phrasebook," *IEEE Trans. Med. Imaging*, vol. PP, no. Xx, p. 1, 2024, doi: 10.1109/TMI.2024.3416190.
[37] M. M. Mohsan, M. U. Akram, G. Rasool, N. S. Alghamdi, M. A. A. Baqai, and M. Abbas, "Vision Transformer and Language Model Based Radiology Report Generation," *IEEE Access*, vol. 11, no. December 2022, pp. 1814–1824, 2023, doi: 10.1109/ACCESS.2022.3232719.
[38] G. Veras Magalhães, R. L. de S. Santos, L. H. S. Vogado, A. Cardoso de Paiva, and P. de Alcântara dos Santos Neto, "XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model," *Heliyon*, vol. 10, no. 7, p. e27516, 2024, doi: 10.1016/j.heliyon.2024.e27516.
[39] A. Moncloa-Muro, G. Ramirez-Alonso, and F. Martinez-Reyes, "Automatic Medical Concept Detection on Images: Dividing the Task into Smaller Ones," *CEUR Workshop Proc.*, vol. 3740, pp. 1668–1680, 2024.
[40] H. Kauschke, K. Bogomasov, and S. Conrad, "Predicting Captions and Detecting Concepts for Medical Images: Contributions of the DBS-HHU Team to ImageCLEFmedical Caption 2024," *CEUR Workshop Proc.*, vol. 3740, pp. 1645–1655, 2024.
[41] Z. Chen *et al.*, "Multi-modal Masked Autoencoders for Medical Vision-and-Language Pre-training," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13435 LNCS, pp. 679–689, 2022, doi: 10.1007/978-3-031-16443-9_65.
[42] M. Aono, T. Asakawa, K. Shimizu, and K. Nomura, "Medical Image Captioning using CUI-based Classification and Feature Similarity," *CEUR Workshop Proc.*, vol. 3740, pp. 1490–1499, 2024.
[43] A. G. Barreto, J. M. de Oliveira, F. N. B. Gois, P. C. Cortez, and V. H. C. de Albuquerque, "A New Generative Model for Textual Descriptions of Medical Images Using Transformers Enhanced with Convolutional Neural Networks," *Bioengineering*, vol. 10, no. 9, 2023, doi: 10.3390/bioengineering10091098.
[44] Q. Van Nguyen, H. Q. Pham, D. Q. Tran, T. K. B. Nguyen, N. H. Nguyen-Dang, and T. B. Nguyen-Tat, "UIT-DarkCow team at ImageCLEFmedical Caption 2024: Diagnostic Captioning for Radiology Images Efficiency with Transformer Models," *CEUR Workshop Proc.*, vol. 3740, pp. 1695–1710, 2024.
[45] M. Aono, H. Shinoda, T. Asakawa, K. Shimizu, T. Togawa, and T. Komoda, "Multi-stage Medical Image Captioning using Classification and CLIP," *CEUR Workshop Proc.*, vol. 3497,

pp. 1387–1395, 2023.
[46] S. Ram, S. Vinoth, R. N. Gopalakrishnan, A. A. Balakumar, L. Kalinathan, and T. A. J. Velankanni, "Leveraging Diverse CNN Architectures for Medical Image Captioning: DenseNet-121, MobileNetV2, and ResNet-50 in ImageCLEF 2024," *CEUR Workshop Proc.*, vol. 3740, pp. 1720–1728, 2024.
[47] G. Y. Kim, B. D. Oh, C. Kim, and Y. S. Kim, "Convolutional Neural Network and Language Model-Based Sequential CT Image Captioning for Intracerebral Hemorrhage," *Appl. Sci.*, vol. 13, no. 17, 2023, doi: 10.3390/app13179665.
[48] J. W. Kong, B. D. Oh, C. Kim, and Y. S. Kim, "Sequential Brain CT Image Captioning Based on the Pre-Trained Classifiers and a Language Model," *Appl. Sci.*, vol. 14, no. 3, 2024, doi: 10.3390/app14031193.
[49] S. Bannur *et al.*, "Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15016–15027. doi: 10.1109/CVPR52729.2023.01442.
[50] S. Elbedwehy, T. M. Taher, and H. Mohammed, "Enhanced descriptive captioning model for histopathological patches," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 36645–36664, 2024, doi: 10.1007/s11042-023-15884-y.
[51] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2002-July, no. July, pp. 311–318, 2002.
[52] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4566–4575, 2015, doi: 10.1109/CVPR.2015.7299087.
[53] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004, [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85
[54] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, no. June, pp. 228–231, 2007.
[55] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating Text Generation With Bert," *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–43, 2020.
[56] Z. Gong, A. P. French, G. Qiu, and X. Chen, "CTranS: A Multi-Resolution Convolution-Transformer Network for Medical Image Segmentation," *Proc. - Int. Symp. Biomed. Imaging*, pp. 1–5, 2024, doi: 10.1109/ISBI56570.2024.10635192.
[57] T. T. Tran, D. T. Vu, T. H. Nguyen, and V. T. Pham, "A CNN-Transformer-based Approach for Medical Image Segmentation," *Proc. 2023 Int. Conf. Syst. Sci. Eng. ICSSE 2023*, pp. 22–27, 2023, doi: 10.1109/ICSSE58758.2023.10227162.
[58] X. Chen *et al.*, "Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, 2022, doi: 10.3390/diagnostics12071549.
[59] D. Abdal Hafeth and S. Kollias, "Insights into Object Semantics: Leveraging Transformer Networks for Advanced Image Captioning," *Sensors*, vol. 24, no. 6, p. 1796, 2024, doi: 10.3390/s24061796.
[60] H. Chen *et al.*, "3D-CT-GPT: Generating 3D Radiology Reports through Integration of Large Vision-Language Models," 2024, [Online]. Available: http://arxiv.org/abs/2409.19330
[61] Z. Chen, Y. Song, T. H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1439–1449, 2020, doi: 10.18653/v1/2020.emnlp-main.112.
[62] A. Kumar and A. Agrawal, "Towards Precision Healthcare: Leveraging Pre-trained Models and Attention Mechanisms in Medical Visual Question Answering for Clinical Decision Support," *2nd IEEE Int. Conf. Networks, Multimed. Inf. Technol. NMITCON 2024*, 2024, doi:

10.1109/NMITCON62075.2024.10699307.

[63] M. Kakkar, D. Shanbhag, C. Aladahalli, and M. Gurunath Reddy, "Language Augmentation in CLIP for Improved Anatomy Detection on Multi-modal Medical Images," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3–6, 2024, doi: 10.1109/EMBC53108.2024.10781689.

[64] L. V. De Moura, R. Ravazio, C. Mattjie, L. S. Kupssinsku, C. M. Dal Sasso Freitas, and R. C. Barros, "Unlocking The Potential Of Vision-Language Models For Mammography Analysis," *Proc. - Int. Symp. Biomed. Imaging*, pp. 5–8, 2024, doi: 10.1109/ISBI56570.2024.10635683.

[65] R. AlSaad *et al.*, "Multimodal Large Language Models in Healthcare: Applications, Challenges, and Future Outlook (Preprint)," *J. Med. Internet Res.*, vol. 26, p. e59505, 2024, doi: 10.2196/59505.

[66] A. Gopu, P. Nishchal, V. Mittal, and K. Srinidhi, "Image Captioning using Deep Learning Techniques," *Proc. IEEE InC4 2023 - 2023 IEEE Int. Conf. Contemp. Comput. Commun.*, vol. 1, pp. 1–5, 2023, doi: 10.1109/InC457730.2023.10263093.