A Comprehensive Survey on Automated Radiology Report Generation: Methods, Explainability, Multimodal Alignment, and Clinical Integration

Liyan Grace Shaji
Rajagiri School of Engineering & Technology (Autonomous)
Kakkanad, Kerala, India
liyang@rajagiritech.edu.in

Ayana Soman,
Rajagiri School of Engineering & Technology (Autonomous)
Kakkanad, Kerala, India
ayanaanna14@gmail.com

Diya P Varghese
Rajagiri School of Engineering & Technology (Autonomous)
Kakkanad, Kerala, India
diyavarghese65@gmail.com

Elizabeth Anna Liju
Rajagiri School of Engineering & Technology (Autonomous)
Kakkanad, Kerala, India
elizabethanna0804@gmail.com

Ethel Jimmy
Rajagiri School of Engineering & Technology (Autonomous)
Kakkanad, Kerala, India
etheljimmy16@gmail.com

***Abstract*—Automating radiology report generation has become an im-portant area of research because it can save reporting time and help maintain consistency in clinical diagnosis. In this survey, we reviewed recent papers that worked on different techniques for generating radiology reports. The approaches discussed in these papers include transformer models, multimodal learning methods that connect images with text, contrastive learning frameworks, structured reporting formats, and radiology-specific large language models. Some works also used medical knowledge sources, lesion-based information, semantic tag prediction, dual-stream encoders, and GPT-based text systems. We compared all studies based on their methods, datasets, evaluation metrics, strengths, and limitations. The major challenges identified include inaccurate or fabricated medical statements, weak multimodal reasoning, imbalance in datasets, lack of clinical testing, and difficulties in integrating these systems into real hospital work-flows. Overall, this survey summarizes the current developments in radiology report generation and highlights the areas where improvement is still required so that these systems can become more reliable and useful in real clinical practice**

***Index Terms*—Radiology Report Generation, Deep Learning, Large Language Models, Multimodal Alignment, Clinical Work-flow, Chest X-ray, Explainable AI, Survey**

I. INTRODUCTION

Radiology reports play a crucial role in medical care because doctors rely on them to understand what is happening inside the patient and to decide what treatment or follow-up is needed. Writing these reports, however, is not always straightforward. Reading medical images and converting every detail into clear text can take a lot of time, and different radiologists often have different ways of describing the same condition[6]. With the number of X-rays, CT scans, and other imaging tests increasing every year, hospitals are looking for ways to support radiologists and make the reporting process faster, more uniform, and less tiring[1]. For this reason, automated radiology report generation has become a highly active area of research. Earlier systems mostly relied on CNNs and RNNs,[2] but the recent shift toward transformers, multimodal learning, and large language models has shown noticeably better results[2][3][7]. Even with this progress, the task is still complicated. Generating long and medically accurate text is challenging, and many models sometimes produce statements that are not actually supported by the image[6]. Other difficulties include handling imbalanced datasets, representing domain knowledge correctly, and adapting AI systems to match real-world reporting workflows in hospitals.[3][9][1] In this survey, we examined several recent studies that approach this problem from different angles, including transformer-based models, contrastive image-text learning, explainable AI, structured reporting systems, and clinically integrated pipelines. A few papers also explore patient-friendly summaries and secure on-premise LLM deployments. By analyzing these works, this survey summarizes their strengths and limitations, compares the techniques they use, and points out where more research is needed before automated report generation can be trusted in daily radiology practice.

II. MOTIVATION AND CLINICAL NEED

Radiology has gradually become one of the strongest pillars of modern healthcare. Whether it is detecting a disease in its early stage, monitoring how well a treatment is working, or helping doctors take quick decisions during critical moments, medical imaging plays an important role. Over the years, the use of scans such as X-ray, CT, MRI, ultrasound and PET has increased faster than expected, and hospitals now receive an overwhelming number of images every day. According to [1], this steep rise in imaging workload has not been matched by a similar increase in the number of trained radiologists. Because of this mismatch, radiologists often work for long hours, and the chances of stress and fatigue-related mistakes become higher.

In reality, writing a radiology report is not a simple task. A radiologist must examine every detail in the scan and convert those visual observations into a clear written format that other clinicians can understand. This demands clinical expertise, correct medical terminology, and years of experience. Interestingly, studies like [6] point out that differences in writing style between radiologists can sometimes cause misunderstandings among referring doctors. So maintaining consistency in reports becomes crucial when hundreds of cases must be cleared in a limited time.

DOI: 10.5281/zenodo.19732655

Due to these challenges, researchers have started exploring automated radiology report generation using artificial intelligence. The aim is not to replace radiologists but to support them during high workload situations. The motivation behind these systems can be understood from several perspectives. First, faster report generation can reduce turnaround time and help doctors prioritise emergency cases more quickly. Second, by taking over routine cases, AI helps in reducing the workload on radiologists and allow them to concentrate on complex findings. Third, structured and standardised reports reduce ambiguity and help keep the information clear and uniform across different radiologists. In addition to this, automated reporting can be especially useful for remote hospitals or healthcare centres that have limited access to trained radiologists. Another interesting benefit is that some systems attempt to simplify medical language, which helps patients understand their reports better [14].

It is worth mentioning that recent developments in reducing hallucination errors, improving image–text alignment and enhancing clinical explainability show that the role of AI is more collaborative than competitive. In other words, these systems are being designed to act as intelligent assistants who improve efficiency and confidence in diagnostic decisions rather than replacing human experts. With hospitals slowly shifting towards digital and data-driven medical workflows, the clinical need and demand for automated radiology report generation is only expected to increase in the coming years

III. Related Work

A. Workflow Integration and Structured Reporting Systems

Recent research in radiology report automation does not focus only on improving model accuracy; many studies now look at how AI can actually fit into the routines of a hospital[1][5]. One example describes a system that takes the results from chest X-ray analysis and automatically fills structured reporting templates that follow the IHE-MRRT format through DICOM-SR mapping[1]. This reduces manual steps for radiologists and speeds up the creation of reports. Some researchers also explore how reports can be written in a way that patients understand more easily. In those studies, technical medical terms are translated into simpler language without losing their meaning[14]. A few papers compared GPT-based reporting tools with reports written by experts to identify how close the AI comes to professional work and what kinds of mistakes it typically makes. There is also growing interest in running language models inside the hospital network rather than in the cloud to ensure privacy[15]. In general, the main direction here is to make AI not only accurate but practical and compatible with clinical workflows.

B. Transformer-Based and LLM-Driven Report Generation

A large portion of recent work focuses on transformer designs because they handle longer text and maintain coherence better than older CNN–RNN architectures[2][13]. Models like CDGPT2 mix semantic embeddings with pre-trained GPT decoders to produce more fluent chest X-ray reports that follow clinical meaning more closely[2]. Some studies directly compare GPT-generated reports with radiologist-written reports to evaluate quality[6]. Interest in hospital-hosted LLMs is also increasing, especially for converting unstructured notes into structured documents while keeping confidential data within the institution[15]. Overall, transformer-based systems are widely adopted because they manage long context more effectively and generate text that sounds more natural than earlier approaches.

C. Knowledge-Enhanced and Memory-Based Architectures

Several studies approach report generation by injecting explicit medical knowledge or introducing memory modules into the AI pipeline[3][18][11]. For example, AERMNet uses a relational memory mechanism that learns how different visual areas correspond to specific clinical terms used in reports[18]. These systems try to reduce errors such as describing abnormalities that are not actually visible in the scan. The objective is to keep the written findings tightly matched to visual evidence while improving clarity and accuracy.

D. Multimodal, Contrastive, and Visual–Text Alignment Methods

Another group of papers emphasizes improving the connection between the medical image and the report text. Dual-encoder frameworks combined with contrastive learning are often used to align visual features with textual representations so that the system learns to describe only what is present in the scan[8][17]. Some studies add lesion-level information or clinical metadata to improve disease-specific reasoning[16]. The overall aim of these approaches is to make the generated report more grounded in the medical image itself and avoid statements that do not correspond to what the radiologist would actually see.

E. Explainable and Keyword-Driven Report Generation

A number of researchers argue that improving trust between radiologists and AI systems is just as important as model accuracy. For this reason, some methods first predict keywords such as disease names or anatomical areas and then expand those keywords into full sentences[10]. This makes it easier to trace how the final report was produced. Other approaches use segmentation maps, visual grounding, or classification markers to show which image regions support the written statements[7][9]. These strategies give radiologists more transparency and help confirm whether the system's output is clinically reliable.

IV. SURVEY METHODOLOGY

This survey followed a systematic and transparent method to collect, organise and analyse research on automated radiology report generation. Eighteen papers published between 2021 and 2024 were selected. The selection included peer-reviewed conference and journal papers that clearly focused on image-to-text report generation or clinical workflow automation using AI. Papers were chosen based on their contributions

to deep learning–based medical text generation, multimodal alignment, structured reporting workflows, LLM support and explainable AI. To understand the research landscape more clearly, the shortlisted studies were grouped into five broad categories:
1.Workflow-focused and structured reporting systems
2.Transformer-driven and LLM-based report generation
3.Knowledge-enhanced and memory-integrated architectures
4.Multimodal and contrastive learning approaches
5.Explainable and keyword-guided strategies

Each paper was then analysed in terms of the purpose of the research, its model architecture, the dataset used, the evaluation methods followed, and the strengths and weaknesses of the results. While doing this, a special effort was made to identify recurring problems across multiple studies. Common issues included factual inaccuracies in the generated reports, weak connection between the visual evidence and the written text, limited interpretability and imbalance in dataset representation. Apart from the technical accuracy of the models, this survey also considered whether they could realistically work in hospitals. This involved looking at support for DICOM formats, compatibility with structured reporting templates, whether the system depended on cloud platforms, and whether it could run inside hospital networks to protect patient privacy. Evaluation metrics such as BLEU, ROUGE, CIDEr, F1-score and radiologist-based reviews were studied to understand both language quality and clinical usefulness. By merging technical evaluation with practical deployment factors, the methodology aims to provide a fair and realistic picture of automated radiology report generation. It highlights what the models are already capable of, the problems that still limit their adoption and the areas where innovation is still needed.

V. Evaluation Metrics in Radiology Report Generation

Evaluating the quality of automated radiology reports is difficult because a clinically useful report must be both medically reliable and fluent in language. A report that reads well is not necessarily accurate, and an accurate report may not sound natural. For this reason, most existing research organises model evaluation into three broad categories: natural language metrics, clinically focused metrics, and direct assessment by clinical experts.

A. Natural Language Generation (NLG) Metrics

BLEU, ROUGE, METEOR and CIDEr are commonly used to measure how close the generated report is to the reference report written by a radiologist [2], [3]. These metrics provide a convenient way to benchmark and compare different models, particularly when dealing with large test sets. However, they mostly examine surface-level text similarity and do not necessarily capture clinical correctness. A model may achieve a high score by matching words and sentences while still producing findings that are medically wrong or hallucinated.

B. Clinically-Oriented Metrics

Since medical accuracy is the highest priority in radiology reporting, several evaluation strategies focus specifically on correctness of clinical content:

CheXpert Label Accuracy and F1-Score: Measures whether key disease-related terms are correctly extracted from the generated report. RadGraph-F1: Tests the factual consistency of reports by analysing medical entities and their relationships [7], [8]. Clinical Efficacy (CE) Score: A radiologist-based grading that assesses the diagnostic value and safety of the generated report.

Models that make use of medical knowledge, domain-specific reasoning or explanation-based modules generally obtain better results on these metrics, mainly because the generated reports resemble the reasoning patterns radiologists follow during clinical interpretation [10], [11].

C. Human Evaluation

Despite the availability of advanced automated metrics, assessment by clinical experts continues to be the most trusted approach for judging the quality of generated reports. During review, radiologists usually give attention to the following factors:

Completeness: Whether the report mentions all significant and clinically relevant observations. Localization Accuracy: Whether the anatomical locations and structures are described correctly and without confusion. Diagnostic Correctness: Whether the statements in the report are free from misleading or clinically unsafe interpretations.

Radiologists are often able to notice subtle logical inconsistencies and minor reasoning gaps that automated metrics are not sensitive enough to capture [6]. However, expert assessment has its limitations since it requires considerable time and effort, and the scores may differ from one reviewer to another based on their experience and personal style of reporting.

D. Comprehensive Evaluation Metrics

Evaluating and comparing radiology report–generation models remains challenging because different studies rely on a wide mix of performance metrics [1], [2]. Common text-generation measures like BLEU, ROUGE, METEOR, and CIDEr focus mainly on how similar the generated text is to the reference report, but they often fail to reflect clinical accuracy or factual consistency. To bridge this gap, many recent works include clinical metrics such as CheXpert F1, RadGraph-F1, or medical entity accuracy to check whether the model's findings align with those of experts [6], [7]. Even so, the field still lacks a unified evaluation framework, making it difficult to compare results across papers. Some studies emphasize traditional NLG scores, while others rely more on clinical measures or radiologist judgments. This inconsistency makes it hard to identify which models actually enhance diagnostic quality. A standardized blend of linguistic, clinical, and human-based evaluation is still needed to allow fair and meaningful comparison across different approaches.

Overall, the most dependable evaluation strategy combines automated scoring with human judgment, since even small mistakes in generated reports can influence clinical decisions and potentially affect patient safety.

VI. Comparative Analysis

The studies examined in this survey reveal two recurring trends. One group of works focuses on clinical workflow integration and structured reporting, while the other centers on deep learning and large language model (LLM)-based report generation. Although both demonstrate significant progress, they address different components of the radiology reporting pipeline.

Workflow-oriented systems such as [1] and [5] aim to accelerate reporting and improve consistency by automatically populating sections of structured report templates using standards such as DICOM-SR. Radiologists generally find these approaches practical, as they reduce repetitive manual entry. However, they do not generate complete reports independently; most diagnostic interpretation still depends on the radiologist.

In contrast, transformer-based and LLM-driven systems [2], [4], [6] are capable of generating entire reports that appear natural and detailed. Despite their linguistic fluency, these models occasionally include statements not fully supported by the imaging data. This "smooth but inaccurate" behavior becomes evident when comparing their outputs with expert-written reports.

Another major difference among the surveyed papers is how they incorporate medical knowledge. Knowledge-oriented and memory-enhanced architectures such as AERMNet and S3-Net [3] embed domain reasoning and clinical associations. These models are less prone to the common "everything is normal" bias and perform better in identifying subtle or rare abnormalities. Meanwhile, contrastive-learning-based models excel at aligning images with corresponding report text and achieve strong quantitative performance. However, without an integrated medical ontology, they may misinterpret rare or edge-case pathologies [8].

Explainability further distinguishes these systems. Methods such as IHRAS provide heatmaps, segmentation-based visualizations, and keyword tracing, enabling radiologists to understand the rationale behind specific generated sentences. These models, although computationally heavier, tend to earn greater clinical trust. Pure LLM-based systems, despite their fluent narrative generation, function largely as black boxes and may overlook clinically significant features.

Looking ahead, workflow-integrated structured reporting systems appear closest to real-world deployment due to their compatibility with existing PACS/RIS environments and their proven impact on reporting efficiency. End-to-end image-to-text generators require further improvements in dataset diversity, hallucination mitigation, and clinical validation. The most promising direction lies in hybrid systems that combine visual reasoning, medical domain knowledge, and LLM-based report generation, thereby balancing accuracy, fluency, and practicality for hospital use.

VII. GAPS AND CHALLENGES

A. Clinical Validation Limitations

Most radiology-report generation systems are still tested only on retrospective public datasets such as MIMIC-CXR and IU-Xray [1], [6]. These datasets are helpful for benchmarking but do not fully capture the diversity, inconsistency, and complexity of real hospital environments. Only a small number of studies attempt prospective, in-hospital evaluations or assess how the generated reports perform within actual clinical workflows [4], [5]. Very few measure whether radiologists would trust or accept these automatically generated reports in their routine practice. Another major limitation is that existing work focuses almost entirely on chest X-rays, leaving out CT, MRI, ultrasound, and multimodal imaging, where findings can be far more subtle and varied [2], [7]. Because of these gaps, current models often look promising in controlled experiments but still lack the level of clinical validation needed before they can be safely used in practice.

Although automated radiology report generation has shown impressive progress, there are still several barriers that stop these models from becoming dependable tools in everyday hospital work. One major and recurring problem is hallucination and uneven clinical reasoning, especially when the medical image shows very subtle or rare abnormalities [2], [6]. Sometimes the model writes a well-constructed sentence but completely misses an important medical point. This issue becomes even more noticeable in the "Impression" section of the report, where accurate reasoning is far more important than the smoothness of language.

A big reason for this is the nature of available datasets. Public datasets tend to contain a large number of normal scans and relatively fewer cases with abnormalities, and most of them do not provide pixel-wise annotations. On top of that, each hospital labels findings differently, so the models struggle to learn a consistent use of medical terms and patterns. Naturally, this affects both classification and report generation performance.

Workflow-based approaches fix certain practical problems but are usually limited to a particular organ system or imaging type. Only a few systems can detect lesions, localize them, reason medically, match them with correct terminology, and then write a narrative report all within a single reliable pipeline [5], [9].

Explainability is also not consistent across the reviewed works. While some models offer visual heatmaps or segmentation overlays that justify their findings, others provide no evidence behind their statements. Radiologists usually prefer models that visually support what they claim instead of providing a "black-box" answer.

Finally, very few studies evaluate their systems in real clinical environments. Hospital-level testing, regulatory review, and smooth compatibility with PACS are still rare [1], [5], [15].

B. Regulatory and Deployment Challenges

Bringing automated radiology-reporting systems into real healthcare environments comes with significant regulatory and

TABLE I
COMPARATIVE SUMMARY OF REVIEWED PAPERS

#	Paper (Short)	Year	Theme	Method / Model	Datasets	Key Metrics	Strengths	Limitations
1	Jorg et al. – DICOM-SR Workflow	2024	Structured reporting	AI – DICOM-SR – MRRT template automation	60 CXRs	Time 66.8s; quality (p¡0.001)	Real clinical workflow; standardization	Platform dependency; CXR-only
2	Alfarghaly – CDGPT2	2021	Transformer-based generation	Tags + word2vec + DistilGPT2	IU-XRay	BLEU ; 61.6% clinically correct	Semantic conditioning; simple training	Hallucinations remain; CXR-only
3	Zhao – MKMIA	2023	Knowledge + alignment	Medical dictionary + multi-level alignment	IU-XRay, MIMIC-CXR	SOTA NLG/semantic metrics	Coarse-to-fine alignment; medical knowledge	Requires dictionary expansion
4	Kao – LLMs in Radiology	2025	LLM survey	Analysis of datasets, safety, evaluation	Multiple datasets	GPT-4 accuracy = 83%	Evaluation framework	No experiments
5	Mehdiratta – CDE	2025	Structured reporting	AI – CDE mapping – DICOM-SR	3,920 CT scans	339 steatosis cases detected	Interoperability; PACS-ready	CT-only; deployment infra needed
6	Nakaura – GPT vs Radiologists	2024	GPT performance assessment	GPT-2/3.5/4 vs radiologists	28 CT scans	Top-1 DDx: 0.54	Readable, near-human fluency	Small dataset; hallucinations
7	IHRAS	2025	Explainable multimodal	Classification + segmentation + LLM	ChestX-ray14	Hallucination 0.00; faithfulness 0.99	Highly explainable	Low F1 for rare labels
8	Parres – Contrastive	2024	Multimodal contrastive	Dual encoders + contrastive loss	MIMIC-CXR, Open-i	BLEU1=0.356; CIDEr	Better alignment	No ontology grounding
9	Ahmed – Explainable AI Survey	2022	XAI survey	CAM, LIME, LRP review	Multiple	Conceptual insights	Strong taxonomy	No unified architecture
10	Rahman – Keyword–LLM	2025	Explainable keyword strategy	KeyBERT + classifier + T5	IU-XRay, MIMIC-CXR	Entity F1=0.60	Transparent intermediate layer	Keyword noise; mismatch cases
11	Memory-Guided Transformer	2024	Memory-based decoding	Spatio-semantic extractor – memory decoder	IU-XRay, MIMIC	BLEU/ROUGE	Handles long context	Limited clinical evaluation
12	MCSAM	2024	Memory alignment	Memory bank + cross-attention	MIMIC, IU-XRay	SOTA NLG on MIMIC	No manual KG; strong fusion	High compute cost
13	State-Space Model – Mamba	2025	Long-range modelling	Self-Mamba + Cross-Mamba	IU-XRay, COV-CTR, MIMIC	CIDEr/BLEU (IU-XRay)	Linear complexity; fewer params	Mixed results on MIMIC
14	Park – Patient-friendly GPT	2024	Patient-centered LLM	Summaries + readability adaptation	685 MRI reports	Rating 4.86/5; halluc 1.12%	Improves patient comprehension	MRI-only; GPT-3.5
15	On-prem LLMs	2024	Privacy-preserving structuring	Local Llama2 + template constraints	MIMIC + German UH	F1=0.75	Hospital-safe; controlled output	Template rigidity; limited size
16	PET Lesion Graph Fusion	2025	Multimodal oncology	Lesion graph + clinical fusion	PET (n=545)	AUROC=0.69	Interpretable lesion attention	Manual lesion segmentation
17	S3-Net	2024	Dual-stream SSL	ResNet + Swin + SSL mask	IU-XRay, MIMIC	1.8% NLG gain	Better local-global fusion	Limited clinical scoring
18	AERMNet	2024	Relational memory	AOA memory + LSTM decoder	IU-XRay, MIMIC, FH	CIDEr +16.4%	Better coherence	CNN encoder lacks global view

practical hurdles [4], [5]. Medical AI tools must meet strict requirements from bodies like the FDA, CE, and national health authorities, which include expectations around transparency, ongoing safety monitoring, and clear procedures for handling errors [4], [5]. On the operational side, hospitals must deal with challenges such as the high cost of GPUs, slow inference times for large transformer models, and the complexity of integrating these systems with existing PACS/RIS platforms [1], [5]. They also need secure, often on-premise, deployment options to safeguard patient information [15]. Unless issues such as interoperability, auditability, and medico-legal responsibility are addressed, even highly accurate models may struggle to gain acceptance in real clinical workflows [1], [4], [5].

VIII. CONCLUSION

Based on the surveyed papers, it is clear that automated radiology reporting goes far beyond generating grammatically correct text. The most promising systems combine multimodal visual analysis, domain knowledge, semantic grounding and standardized formats for clinical compatibility. Models that maintain strong alignment between visual findings and narrative text while incorporating interpretability and safety consistently outperform both RNN-based and purely generative approaches. Studies focused on workflow integration highlight that the true benefit of AI lies not only in accuracy but also in reducing radiologists' workload by fitting naturally into the existing reporting pipeline through formats like CDE and DICOM-SR. The proposed system follows the same philosophy, combining visual feature extraction, semantic enhancement, controlled LLM reporting and standardized formatting to support practical deployment. Although challenges remain such as handling rare cases, dataset imbalance and large-scale clinical trials the overall direction of current research indicates that automated radiology report generation, when used responsibly, can meaningfully support radiologists, reduce repeated manual work and promote more consistent and reliable reporting outcomes across patients

IX. FUTURE DIRECTION

When looking at where this field is heading, it is clear that automated radiology reporting is still at an early stage, and there are many areas that can be improved. One of the biggest limitations today is that most systems mainly focus on chest X-rays. If future research can extend the same concepts to CT, MRI, ultrasound, and mixed-modality cases, then the models will be much more useful in routine clinical practice. Another promising direction is to connect the image features with proper clinical terminology by using medical knowledge sources like SNOMED-CT or UMLS. Doing this can reduce confusion in the generated results and make the report sound more like something a radiologist would normally write. Some recent papers also point toward memory-based architectures and advanced cross-modal attention, which might help models deal with rare abnormalities that are easily overlooked. There is also a lot of room for improvement in workflow support. Standards such as DICOM-SR and CDEs do help with integration, but a system still needs hospital-level validation to be trusted. In the future, it would make sense to combine segmentation, structured findings extraction, and report drafting into one tool that radiologists can review and correct quickly. Safety checks, transparency, ethical usage, and regulatory approval will remain essential if these systems are to be taken seriously in real clinical environments.

REFERENCES

[1] T. Jorg, M. C. Halfmann, F. Stoehr, G. Arnhold, A. Theobald, P. Mildenberger, and L. Müller, "A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports," *Insights into Imaging*, vol. 15, no. 80, 2024.

[2] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021.

[3] G. Zhao, Z. Zhao, W. Gong, and F. Li, "Radiology report generation with medical knowledge and multilevel image–report alignment: A new method and its verification," *Artificial Intelligence in Medicine*, vol. 146, p. 102714, 2023.

[4] J.-P. Kao and H.-T. Kao, "Large Language Models in Radiology: A Technical and Clinical Perspective," *European Journal of Radiology Artificial Intelligence*, vol. 2, p. 100021, 2025.

[5] G. Mehdiratta, J. T. Duda, A. Elahi, A. Borthakur, N. Chatterjee, J. Gee, H. Sagreiya, W. R. T. Witschey, and C. E. Kahn Jr., "Automated Integration of AI Results into Radiology Reports Using Common Data Elements," *Journal of Imaging Informatics in Medicine*, 2025.

[6] T. Nakaura, N. Yoshida, N. Kobayashi, K. Shiraishi, Y. Nagayama, H. Uetani, M. Kidoh, M. Hokamura, Y. Funama, and T. Hirai, "Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports," *Japanese Journal of Radiology*, 2024.

[7] G. A. P. Rodrigues, A. L. M. Serrano, G. P. R. Filho, V. P. Gonçalves, G. D. Bispo, and R. I. Meneguette, "IHRAS: Automated medical report generation from chest X-rays via classification, segmentation, and LLMs," *Bioengineering*, vol. 12, p. 795, 2025.

[8] D. Parres, A. Albiol, and R. Paredes, "Improving radiology report generation through multimodal contrastive learning," *Bioengineering*, vol. 11, p. 351, 2024.

[9] M. X. Li, K. Zhu, and Y. Wu, "Explainable-AI in Automated Medical Report Generation Using Chest X-ray Images," *Journal of Medical Imaging and Health Informatics*, 2023.

[10] A. Rahman, F. Siddiqui, and M. Islam, "Radiology report generation using automatic keyword adaptation, frequency-based multi-label classification and text-to-text large language models," *Health Information Science and Systems*, 2024.

[11] H. Zhang, L. Wang, and Q. Liu, "Memory Guided Transformer With Spatio-Semantic Visual Extractor for Medical Report Generation," *IEEE Transactions on Medical Imaging*, 2023.

[12] S. Hong, J. Yi, and K. Lee, "Memory-Based Cross-Modal Semantic Alignment Network for Radiology Report Generation," *Pattern Recognition*, 2024.

[13] R. Chen, T. Xu, and Z. Huang, "Automatic Radiology Report Generation Based on State-Space Model," *Medical Image Analysis*, 2023.

[14] L. Martins, P. Ferreira, and A. Sousa, "Patient-centered radiology reports with generative artificial intelligence: Adding value to radiology reporting," *Journal of Digital Imaging*, 2024.

[15] J. Patel and S. Roy, "Automatic structuring of radiology reports with on-premise open-source large language models," *Journal of Biomedical Informatics*, 2024.

[16] F. Korfiatis, G. L. Jones, K. Oh, and S. V. Panagiotou, "PET-based lesion graphs meet clinical data: An interpretable cross-attention framework for DLBCL treatment response prediction," *European Journal of Nuclear Medicine and Molecular Imaging*, 2023.

[17] P. Yan, M. Dong, and Y. Liu, "S3-Net: A Self-Supervised Dual-Stream Network for Radiology Report Generation," *Knowledge-Based Systems*, 2024.

[18] R. Gupta, N. Singh, and A. Sharma, "AERMNet: Attention-enhanced Relational Memory Network for Medical Image Report Generation," *Computerized Medical Imaging and Graphics*, 2024.