Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1703-1710
Journal homepage: http://iieta.org/journals/isi

Survey of Deep Learning Techniques for Radiology Report Generation

Greeshma Yedla*, Thirupathi Vadluri

School of Computer Science & Artificial Intelligence, SR University, Warangal 506371, India

Corresponding Author Email: yedla.greeshma12@gmail.com

https://doi.org/10.18280/isi.310526

Received: 14 October 2025
Revised: 15 January 2026
Accepted: 25 April 2026
Available online: 31 May 2026

Keywords:
radiology report generation, deep learning, vision-language models, transformers, knowledge graphs, medical imaging, natural language processing, clinical text generation

ABSTRACT

Radiology report generation has emerged as a highly significant research area owing to the increasing demand for automated and accurate clinical documentation in modern healthcare systems. Recent advances in deep learning have significantly improved the capability to generate clinically meaningful radiology reports by learning complex relationships between medical images and textual descriptions. This survey provides a comprehensive analysis of the evolution of deep learning techniques for automated radiology report generation, ranging from conventional encoder-decoder architectures to advanced attention-based, transformer-based, vision-language and knowledge graph-based models. The survey examines major developments, such as attention mechanisms, multimodal pretraining, and the integration of structured medical knowledge through knowledge graphs. Representative models, including Bootstrapping Language-Image Pretraining (BLIP), Global-Local Representation Learning for Image-Text Alignment (GLoRIA), and Biomedical Vision-Language Pretraining (BioViL), are critically analyzed to highlight their strengths and limitations. After analyzing various approaches to radiology report generation, the key challenges are identified, namely inadequate semantic representation of data, clinical hallucinations, and the use of manually created ontologies in the training process. To address these challenges, the survey identifies promising future research directions. Specifically, it involves integrating transformer-based neural networks with medical knowledge and Explainable Artificial Intelligence (XAI) techniques to improve the clinical accuracy, reliability, and interpretability of automated radiology report generation systems.

1. INTRODUCTION

With the fast-paced development of medical imaging techniques, the number of diagnostic imaging examinations performed has increased drastically. As a consequence, the need for precise and detailed radiology reports also escalated. Radiology reports play a crucial part in medical diagnostics and treatment planning. They serve as the primary source of information exchange between radiologists and other medical professionals. Their importance cannot be underestimated, as they are crucial for clinical practice [1, 2].

The preparation of radiology reports is a highly labor-intensive and time-consuming process that requires profound skills from the specialists responsible for it. Indeed, a radiologist has to thoroughly analyze medical images, find abnormalities in them and describe the findings in a correct and professional manner. In addition, the increasing workload in modern medical facilities makes preparing radiology reports rather difficult. Besides, in some areas of the world, there is a shortage of radiologists who can write informative and understandable radiology reports. This issue also adds to the difficulty of producing high-quality radiology reports and causes delays in the reporting process. The continuous growth in the number of imaging studies because of advances in diagnostic imaging techniques and increased availability of medical care leads to this problem [3, 4].

However, recent advances in artificial intelligence can help solve this issue to some extent. Deep learning has allowed building systems that can automatically produce clinically relevant radiology reports. Deep learning algorithms enable automatic generation of descriptive reports through analyzing the input images and drawing conclusions based on them. Using datasets with image–report pairs, deep learning algorithms have shown impressive results in automatically producing medically useful narratives similar to those created by humans [5, 6]. Nevertheless, the creation of fully autonomous radiology report generation systems still encounters challenges associated with multimodal learning, comprehensive encoding of medical knowledge, and a high degree of accuracy.

Various deep learning models have been developed recently, which attempt to generate radiology reports. Starting from basic encoder–decoder models and ending with advanced transformer and vision-language models (VLMs), deep learning algorithms have become increasingly complex. Finally, with the rise in the popularity of knowledge graphs, attempts at generating clinically informative radiology reports have been made. Such models incorporate a knowledge base in order to ensure clinical reasoning when generating narratives. This survey offers a comprehensive overview of recent developments in deep learning algorithms aimed at generating radiology reports. This paper analyzes and

compares major classes of radiology report generation systems and assesses their effectiveness and applicability in real-world medical settings. Moreover, by comparing current approaches to radiology report generation, this survey identifies some key challenges to their development and application. Lastly, the survey considers the practical issues related to integrating these systems into the healthcare industry.

1.1 Contributions of this survey

This survey makes several key contributions to research on deep learning algorithms in radiology report generation:

1. It provides a detailed and systematic classification of radiology report generation techniques. The survey categorizes these techniques into several broad categories, namely, encoder–decoder architectures, attention-based models, transformers and vision language models, and finally, knowledge graph based models.
2. Secondly, the review conducts a comparative analysis of current approaches to radiology report generation and identifies the benefits and drawbacks of each approach.
3. It identifies several research challenges in the field, such as lack of clinical grounding, hallucination of findings, inability to correctly describe rare diseases, dataset bias, and lack of appropriate evaluation metrics.
4. It discusses some practical challenges in implementing automatic radiology report generation systems in a hospital setting, specifically discussing PACS and HIS integration requirements.
5. Lastly, the presented survey outlines promising avenues of research in this area based on the challenges identified by the authors. These avenues include hybrid architecture combining transformers, knowledge graphs, multimodal representation learning, and explainable AI.

1.2 Research methodology

To provide an exhaustive overview of the research topic in question, a systematic study of existing works is performed. Relevant papers were obtained through a thorough search conducted in the main academic resources, among which there are IEEE Xplore, SpringerLink, ScienceDirect, PubMed, and arXiv. Keywords used during the searching procedure include radiology report generation, medical image captioning, deep learning in radiology, VLMs in healthcare, and automated medical report generation to make sure that the widest range of studies is covered.

As a means to select high-quality papers from the large set obtained through the searching procedure, a set of inclusion and exclusion criteria was used. Among the factors included in these criteria were deep learning application in radiology report generation, the development of innovative approaches, or the publication of a comprehensive literature review within the area of interest. All peer-reviewed scientific articles and reputable preprints, both journal publications and conference proceedings, were taken into account.

At the same time, several exclusion criteria were defined. They consisted of irrelevant topics, insufficient contribution, purely theoretical papers, works concerning only general-purpose image captioning without any specific medical applications, or works lacking complete information on methodology. Studies featuring no results of experiment or no analysis of achieved results were also excluded from the further analysis.

The next step involves a systematic classification of studies based on their methodological basis. The following classes are currently distinguished in this survey: encoder–decoder models, attention-based models, transformer-based architecture, VLMs, and knowledge graph-based approaches. Such an organized classification of works allows conducting a comparative analysis of existing approaches and evaluating the potential of further innovations in radiology report generation. In addition, a structured approach provides insights into the emerging research areas and reveals new directions for further investigation.

2. LITERATURE REVIEW

In recent years, the field of automatic radiology report generation has significantly progressed due to rapid advancements in deep learning and machine vision technologies. Initially, research on this issue was concerned with establishing the possibility of generating textual descriptions of medical images in a way similar to how it can be done using sequence-to-sequence frameworks. Over time, new architectures were proposed to facilitate the generation of highly accurate and linguistically adequate medical descriptions of images. Currently, radiology report generation methods have evolved to advanced transformer-based, vision-language, and knowledge graph frameworks.

Development of new techniques in the course of research indicates the rising importance of enhancing both linguistic and clinical adequacy of the text generated. Initially, convolutional and recurrent neural network models were used to solve this problem. Then, attention-based architectures were applied for the purpose of better alignment of regions in medical images and corresponding textual descriptions. Recent research shows the successful implementation of transformers and VLMs. Simultaneously, the use of knowledge graphs is gaining momentum.

To provide an organized view of different types of radiology report generation models, it is necessary to classify existing works according to the common features of the architectures used. Therefore, this survey introduces a systematic categorization of studies based on the type of model used, including encoder–decoder models, attention-based models, transformer-based architectures, VLMs, and knowledge graph-based approaches.

2.1 Encoder–decoder models

The earliest approach to generating radiology reports based on deep learning is the encoder-decoder architecture, which makes use of convolutional neural networks (CNN) to encode image information and recurrent neural networks (RNN), especially long short-term memory (LSTM) networks to generate sentences. Early attempts to generate medical reports by this architecture were performed by Pang et al. [1] and Wang et al. [2]. Hierarchical LSTM architectures were introduced later, in which the goal was to generate the report either on the sentence-level or word-level. Even though these early approaches showed some promise in mapping image features to medical reports, encoder-decoder models still have difficulties with handling long-range dependencies and

complex clinical interactions.

2.2 Attention-based models

Attention-based models have emerged to resolve some of the drawbacks that encoder-decoder models suffer from. In contrast to encoder-decoder architecture, attention-based models can focus selectively on parts of the input image during the generation process. This concept was introduced to computer vision by Liu et al. [3] for image captioning tasks, and subsequently, similar models were applied for medical images by Monshi et al. [4] and Mamdouh et al. [5]. Hierarchical attention-based models were also introduced to improve the performance of report generation. However, even though attention-based models have significantly advanced visual grounding, they still lacked integration with structured clinical knowledge.

2.3 Transformer-based models

In addition to solving some of the shortcomings of attention-based models, transformers have proven to be effective models for long text generation. Transformer-based architectures eliminate the need for recurrence, and they can efficiently capture dependencies within sequences. Works by Sloan et al. [6] and Huang et al. [7] show that transformer-based architectures can successfully generate complex long radiology reports. Moreover, transformer architectures have proven to work well when used in combination with vision transformer (ViT) for improved feature extraction from images. Unfortunately, even though transformers outperform other architectures in many cases, they are still prone to hallucinations, meaning that they sometimes produce findings that do not exist in an image. Furthermore, high resource requirements make transformer models unsuitable for some environments in healthcare settings.

2.4 Vision-language models

Multimodal VLMs are models which jointly train visual and language modalities to learn the intermodal relationship. Most VLMs are trained using contrastive learning and masked language modeling methods on large-scale image-caption datasets. Some of the most popular VLMs used for medical report generation include GLoRIA [8], BLIP [9], and BioViL [10], all of which achieved high-quality performance in image-region-to-sentence alignment task. Nevertheless, the use of VLMs in radiology report generation remains challenging, since it requires adequate clinical domain adaptation. Furthermore, pre-training on image-caption datasets might cause undesired biases in some cases. For example, GLoRIA showed great success in performing image-region to sentence alignment. However, it failed to recognize some rare conditions due to lack of clinical grounding. For a more complete overview of VLMs and multimodal large language models, please see recent surveys such as Hartsock and Rasool [10], and Yi et al. [11].

2.5 Knowledge graph-based models

Knowledge graph-based models make use of clinical domain knowledge encoded in a knowledge graph to generate medical reports. Knowledge graph-based approaches represent clinical relationships between medical entities such as diseases, symptoms, and anatomy using graph neural networks (GNN). The work by Mou [12] used clinical metadata to generate context-aware radiology reports. Singh [13] and Singh and Singh [14] generated radiology reports using deep learning techniques. More recent studies focus on combining knowledge graph-based approaches with transformers to take advantage of both. As knowledge graph approaches improve interpretability of models and their clinical accuracy, they still face some challenges due to their resource requirements and need for domain knowledge [15, 16]. The graph-based techniques to aid semantic reasoning and generate clinical reports. Liu et al. [17] introduced the concept of Dual Graph Convolutional Network that would enhance semantic relations between visual and text data, whereas Li [18] stressed the need for structured clinical knowledge to be used in the deep learning model to increase accuracy and semantic coherence of generated radiology reports. Hybrid approaches combining knowledge graphs with deep learning have also been researched recently [19-21]. A more detailed comparison of existing radiology report generation models is shown in Tables 1 and 2.

Table 1. Summary of previous research work on radiology report generation

Author(s)	Focus Area	Key Contributions	Limitations
Monshi et al. [4]	Deep learning survey	Introduced CNN–RNN and attention models; highlighted datasets and metrics	Did not cover transformers or VLMs
Pang et al. [1]	Medical report generation	Comprehensive review of encoder–decoder pipelines and evaluation metrics	Limited coverage of knowledge graphs and VLMs
Liu et al. [3]	Transformers & multimodal models	Focused on LLMs, dataset bias, and generalization	Limited focus on clinical explainability
Sloan et al. [6]	Recent advances	Summarized encoder–decoder to transformer evolution and training strategies	Preprint; limited clinical validation
Messina et al. [22]	Explainability	Emphasized interpretability and trust in AI systems	Limited performance evaluation
Hartsock and Rasool [10]	Vision-language models (VLMs)	Reviewed VLMs for report generation and VQA tasks	No novel model proposed
Yi et al. [11]	Multimodal LLMs	Benchmarked VLMs in radiology	Limited knowledge integration analysis
Liu et al. [15]	GCN + Transformer	Systematic review of hybrid models	No implementation insights
Chen et al. [16]	Vision-language applications	Survey of medical vision-language systems	Generalized beyond radiology

Apart from the model-specific analysis described above, several systematic reviews [15, 16] shed light on multimodal

learning as applied in healthcare for vision-language applications. The tables have been adjusted to distinguish clearly between the studies that focus on surveys and those employing models.

A comparative analysis of existing approaches to building radiology report generation models demonstrates certain peculiarities and drawbacks associated with each method and technology used. Thus, encoder-decoder approaches employing CNN and RNN models perform efficiently, yet lack capabilities of capturing long dependencies and understanding clinical semantics. Attention-based models represent an improvement since they allow better matching of corresponding image patches with texts. Still, they do not incorporate the application-specific knowledge required for generating reports. Transformer models are even more efficient because they use the concept of self-attention. Thus, better context understanding and more coherent report generation can be achieved. At the same time, this type of models is susceptible to hallucinations when they generate incorrect or non-existent findings, resulting in their limited use in medicine. VLMs, namely BLIP and GLoRIA, improve upon these shortcomings by employing multimodal pre-training to align visual and linguistic representations better. However, they are incapable of handling rare or disease-specific diseases since their reports suffer from the lack of clinical grounding. Knowledge graph-based methods address this problem by introducing medical semantics and domain-specific knowledge into the model, but they are complicated to design.

In conclusion, it appears that hybrid models combining transformer and knowledge graph approaches are particularly promising. Besides, knowledge graph-based models are better at clinical grounding than transformers and vision-language approaches, which excel at fluency and scalability.

Figure 1 depicts the evolution of deep learning models used for generating radiology reports.

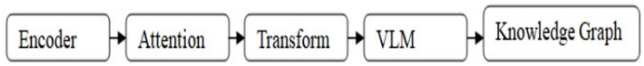

Figure 1. Taxonomy of deep learning models for radiology report generation

Apart from comparing model architectures based on their performance in terms of fluency, consistency, and other aspects, another aspect to consider is the trade-off of different models. While encoder-decoder and attention-based models provide computational simplicity, they lack capabilities of modeling complex clinical dependencies. On the other hand, transformers and VLMs increase the model's capabilities of modeling semantic context and improving fluency of generated reports, yet they have the tendency to hallucinate and ignore clinical knowledge. Finally, knowledge graph-based models solve this problem by adding structured medical knowledge and thus making generated reports more consistent and semantically correct; yet, they become more complex and depend on curated datasets. All this demonstrates that all these model architectures have trade-offs, which means that the choice of a particular model depends on certain requirements and constraints. The conclusion that was derived through this analysis is that there is no universally optimal architecture and each one comes with certain advantages and disadvantages. Another interesting finding was that the inclusion of structured medical knowledge and medical representations helped improve the reliability of models [18]. The comparative analysis is summarized in Table 3.

Table 2. Model-based approaches for radiology report generation

Author	Model Type	Representative Model	Key Findings	Limitations
Wang et al. [2]	Encoder–Decoder	CNN + LSTM	Established baseline for report generation using sequence modeling	Poor long-term dependency handling, lacks semantic understanding
Mamdouh et al. [5]	Attention-Based	Show-Attend-Tell	Improved spatial alignment between image regions and text	Weak clinical interpretability, lacks domain knowledge
Sloan et al. [6]	Transformer	Vision Transformer	Generates fluent and context-aware reports	Prone to hallucination, high computational cost
Huang et al. [7]	Transformer	Metransformer	Unified transformer architecture for report generation	Limited external validation
Li et al. [8]	Vision–Language	GLoRIA	Strong image-text alignment using global-local features	Requires clinical grounding for reliability
Zhang et al. [9]	Vision–Language	BLIP	Effective multimodal pretraining for text generation	Not specifically optimized for medical domain
Mou [12]	Knowledge Graph	CliCon	Incorporates patient context into report generation	Requires complete structured metadata
Singh [13]	Knowledge Graph	GCN-KG	Integrates structured medical knowledge for better semantic accuracy	High complexity and slow inference
Liu et al. [15]	Hybrid (KG + VLM)	KG-enhanced VLM	Combines knowledge graphs with vision-language models (VLMs) for improved accuracy	High model complexity
Liu et al. [17]	Hybrid (Graph + Transformer)	Dual GCN	Enhances semantic relationships using dual graph structures	Limited scalability
Tian et al. [21]	Hybrid (Transformer + Knowledge)	Clinical Graph Transformer	Improves domain-specific report generation	Limited to specific clinical domain

Table 3. Comparative analysis of radiology report generation models

Model Type	Strength	Limitation	Clinical Suitability
CNN + LSTM	Simple and fast	Poor long-term dependency	Low
Attention Models	Better image-text alignment	No clinical reasoning	Medium
Transformer	Fluent and contextual output	Hallucination problem	Medium
Vision-Language Models (VLMs)	Strong multimodal learning	Weak for rare diseases	High
Knowledge Graph Models	High semantic accuracy	Complex and data dependent	Very High

3. OBSERVATIONS AND GAPS IDENTIFIED

Based on this analysis of existing model architectures, the following observations can be made:

1. Although CNN + LSTM-based architectures can be seen as the baseline of all others, they lack capabilities to learn complex dependencies in longer reports as well as context in the form of medical semantic representations.
2. Attention models are able to improve visual–textual alignment, but they still operate independently from clinical ontologies, which makes their results hard to understand in practical application cases.
3. Models based on transformer architecture provide increased fluency and contextuality; however, they are still susceptible to hallucination, which is extremely dangerous for models' use in clinical practice.
4. VLMs like BLIP, GLoRIA, and BioViL exhibit excellent generalization capability on publicly available datasets; however, they struggle with unusual pathological cases.
5. Modeling using knowledge graphs increases semantic accuracy and clinical relevance of generated reports, yet they require manually curated ontologies and have scalability issues.

3.1 Research gap

There are several research gaps in radiology report generation, despite recent advancements in this field. First of all, there is usually no effective incorporation of medical knowledge, which results in hallucinations and reduces clinical reliability of the models. Second, most of the approaches rely on big corpora, but they fail to achieve optimal performance in cases with rare pathologies or when the case becomes complicated. Finally, current evaluation methods of models are mostly focused on linguistic similarity, and thus they lack clinical assessment. Apart from that, little attention is paid to deployment on healthcare infrastructure, such as PACS or HIS. These research gaps clearly demonstrate the need for hybrid and knowledge-enhancing approaches for the development of linguistically fluent and clinically accurate report generation models suitable for practical application. To fill these gaps, this survey provides a review of existing models and discusses hybrid approaches combining transformer architecture with knowledge graphs.

3.2 Proposed research direction

To address the above-mentioned research gaps, some directions of research can be considered. It is important to develop techniques for incorporating medical domain knowledge into deep learning models for minimization of hallucinations and improvement of the model's reliability in clinical practice. Also, VLMs need to be fine-tuned with clinically curated datasets for improvement of their accuracy in unusual pathological cases. The creation of explainable architecture is also necessary for generation of reports, as the model must align generated output to clinically relevant entities and image regions. Incorporation of multimodal and context information like patients' medical history or previous examinations can be helpful for accuracy improvement as well.

One possible approach to solving the problem can be developing of hybrid deep learning models combining transformer architecture with knowledge graph-based reasoning and interpretability. Such models can help achieve increased semantic grounding, reduction of hallucinations, and generation of clinically relevant and accurate reports without losing scalability to multiple imaging modalities.

3.3 Benefits and clinical impact of automated radiology report generation

Necessity for automating radiology report generation process is driven by various reasons: increasing number of imaging studies required; the shortage of specialists in many countries; fast report delivery; uniform terminology; and many other. Thus, such tools will have the potential of helping with these problems. Moreover, they might help reducing inter-rater variability in radiologists, which will result in higher diagnostic accuracy.

3.4 Evaluation metrics for radiology report generation

Evaluation of radiology report generation models is usually performed using natural language processing metrics like BLEU, ROUGE, and CIDEr. BLEU stands for bi-lingual evaluation understudy metric and evaluates the n-grams overlap, so it is widely used for evaluating text generation models; however, this metric does not consider any clinical correctness. ROUGE is also based on n-grams overlap, and it measures recall between the generated text and the reference text. However, it is a surface-level metric for assessing generated text. CIDEr stands for Consensus-based Image Description Evaluation and is used for image caption generation task; therefore, it is able to evaluate descriptive quality of generated text.

Although these metrics are good for evaluation purposes, they are unable to evaluate clinical correctness of the models' output since they do not evaluate accuracy of medical facts or their relevancy for the report generation task. Therefore, clinical metrics involving expert review of generated reports are necessary for real-world evaluation of radiology report generation models. Domain-specific evaluation frameworks that incorporate linguistic quality together with clinical correctness should be developed as an important future research direction. The overview of existing metrics and their

benefits and limitations is provided in Table 4.

Such a comparison shows the trade-off between language skills and clinical accuracy in various architectures. Usually, Transformer-based models and VLMs demonstrate better results in terms of metrics like BLEU, ROUGE, and CIDEr, than encoder-decoder models. Nonetheless, better metrics do not necessarily mean that the output will be clinically accurate.

Table 4. Comparison of evaluation metrics

Metric	Purpose	Advantage	Limitation
BLEU	Measures n-gram overlap between generated and reference text	Simple and widely used	Does not capture clinical correctness or meaning
ROUGE	Measures recall-based overlap of text	Good for evaluating coverage	Focuses on surface similarity, not semantics
CIDEr	Evaluates similarity using consensus among multiple references	Better for descriptive quality	Not designed for medical domain
Clinical Accuracy	Measures correctness of medical findings	Reflects real clinical usefulness	Requires expert evaluation, time-consuming

3.5 Applications of radiology report generation systems

The creation of radiology report generation systems involves numerous advantages for both clinicians and patients. For clinicians, the system provides a tool to produce preliminary reports, find important points, and save time on redundant reporting to focus on analyzing cases. Moreover, these algorithms can serve as secondary reviewors and provide consistency in diagnostics, thus increasing diagnostic accuracy. Besides, the system proves to be useful for education purposes as it helps to learn from standard reports. These automatic generators can prove to be effective in performing studies and clinical trials since it provides quick access to structured information.

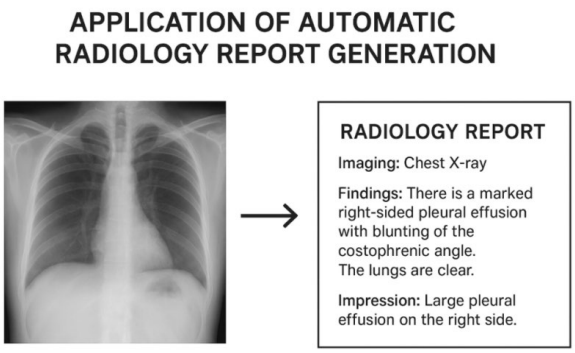

Figure 2. Application of an automatic radiology report generation system

In practice, the workflow may look the following way. The image is uploaded into PACS first, then the algorithm performs analysis, and finds possible pathological features and generates a report. Further, a clinician reviews the generated report and approves the findings, and the final version is saved in HIS. Thus, these tools decrease the time needed for report preparation, increase accuracy, and facilitate decision-making, especially in high-stress environments, like an ER department. The same approach applies to CT/MRI reports since the system supports the production of structured reports for complex images. As one can see, a model may generate "no abnormalities found" report even though there are some pathologies in the image, and human validation is still crucial. Figure 2 demonstrates application of an automatic radiology report generation system.

3.6 Integration with healthcare information systems

In a clinical setting, radiology report generation systems should be compatible with already existing healthcare systems like PACS and HIS. Figure 3 shows how the process of generating radiology reports is carried out from end to end through image capture, processing of model, report generation, radiologist validation, and integration into the health systems.

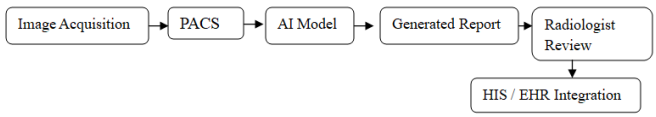

Figure 3. System architecture of radiology report generation integrated with healthcare information systems

In a typical workflow, medical images are first acquired and stored in PACS. Next, they are retrieved and processed through deep learning models for report generation. The resulting report undergoes review and validation by radiologists and then stored in HIS. In terms of integration between components in the clinical reporting system, the proposed end-to-end workflow allows smooth data flow among components. Consequently, the clinical workflow not only includes automated radiology report generation but also guarantees clinical oversight over the reports generated. Moreover, efficient data storage, security, system interoperability, and other system-level considerations such as scalability, efficiency, and reliability are vital during implementation for large-scale hospitals. In this regard, the presented integration facilitates deployment of radiology report generation models into the clinical setting.

3.7 Future directions and challenges

Specific areas that require attention in future research include addressing a number of challenges related to the application of deep learning to generate radiology reports. First, it is crucial to find methods for dealing with the problem of hallucinations in vision-language and transformer-based models that involve generation of either non-existent or wrong findings. To tackle this issue, researchers need to come up with suitable grounding mechanisms. Also, designing evaluation metrics specific to radiology tasks rather than just using NLP measures like BLEU or ROUGE becomes a priority as well. Moreover, improving model performance on rare diseases and uncommon medical findings represents another important area to consider since existing models tend to have problems in such situations. From the point of view of system design, researchers need to pay more attention to real-time integration

of report generation systems with PACS and HIS, as well as to the incorporation of multimodal data (e.g., patient history, lab results). Such approaches would facilitate better report generation. Finally, cross-modal reasoning capabilities are also crucial in order to improve clinical applicability of future solutions. This is the message of recent studies [21, 22]. In addition, recent research has demonstrated that efficient feature extraction techniques and lightweight deep learning architectures can significantly improve computational efficiency while maintaining high predictive performance [19, 23].

4. CONCLUSION

This survey provides a comprehensive review on the development and current status of research on the application of deep learning in generating automated radiological reports. From simple encoder-decoder models to highly advanced transformer-based architectures and multimodal approaches, there have been significant developments in the field, focusing on achieving fluency, semantics, and accuracy in the process of creating radiology reports. Nonetheless, important problems like hallucinations, insufficient clinical grounding, and non-interpretable systems prevent the safe application of such approaches in real-world applications within the health sector.

We believe that the use of knowledge graphs and multimodal fusion can serve as effective avenues for addressing some of the problems mentioned above. In light of these issues, we propose a novel research approach involving a combination of transformer-based vision-language modeling, structured medical knowledge, and techniques in explainable AI.

REFERENCES

[1] Pang, T., Li, P.G., Zhao, L.J. (2023). A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine, 22: 48. https://doi.org/10.1186/s12938-023-01113-y

[2] Wang, X.Y., Figueredo, G., Li, R.Z., Zhang, W.E., Chen, W.T., Chen, X. (2024). A survey of deep learning-based radiology report generation using multimodal data. arXiv preprint arXiv:2405.12833. https://doi.org/10.48550/arXiv.2405.12833

[3] Liu, X.Y., Xin, J.C., Shen, Q., Huang, Z.H., Wang, Z.Q. (2025). Automatic medical report generation based on deep learning: A state-of-the-art survey. Computerized Medical Imaging and Graphics, 120: 102486. https://doi.org/10.1016/j.compmedimag.2024.102486

[4] Monshi, M.M.A., Poon, J., Chung, V. (2020). Deep learning in generating radiology reports: A survey. Artificial Intelligence in Medicine, 106: 101878. https://doi.org/10.1016/j.artmed.2020.101878

[5] Mamdouh, D., Attia, M., Osama, M., Mohamed, N., and Lotfy, A., Arafa, T., Rashed, E.A., Khoriba, G. (2025). Advancements in radiology report generation: A comprehensive analysis. Bioengineering, 12(7): 693. https://doi.org/10.3390/bioengineering12070693

[6] Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M. (2025). Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering, 18: 368-387. https://doi.org/10.1109/RBME.2024.3408456

[7] Huang, S.C., Shen, L.Y., Lungren, M.P., Yeung, S. (2021). GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 3922-3931. https://doi.org/10.1109/ICCV48922.2021.00391

[8] Li, J.N., Li, D.X., Xiong, C.M., Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086. https://doi.org/10.48550/arXiv.2201.12086

[9] Zhang, Y.X., Wang, X.S., Xu, Z.Y., Yu, Q.H., Yuille, A., Xu, D.G. (2020). When radiology report generation meets knowledge graph. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07): 12910-12917. https://doi.org/10.1609/aaai.v34i07.6989

[10] Hartsock, I., Rasool, G. (2024). Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence, 7: 1430984. https://doi.org/10.3389/frai.2024.1430984

[11] Yi, Z.R., Xiao, T., Albert, M.V. (2025). A survey on multimodal large language models in radiology for report generation and visual question answering. Information, 16(2): 136. https://doi.org/10.3390/info16020136

[12] Mou, Y.L. (2025). Knowledge graph-enhanced vision-to-language multimodal models for radiology report generation. In the Semantic Web: ESWC 2024 Satellite Events, pp. 115-124. https://doi.org/10.1007/978-3-031-78955-7_12

[13] Singh, S. (2024). Designing a robust radiology report generation system. arXiv preprint arXiv:2411.01153. https://doi.org/10.48550/arXiv.2411.01153

[14] Singh, P., Singh, S. (2025). ChestX-Transcribe: A multimodal transformer for automated radiology report generation from chest x-rays. Frontiers in Digital Health, 7: 1535168. https://doi.org/10.3389/fdgth.2025.1535168

[15] Liu, C., Tian, Y.H., Song, Y. (2023). A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199. https://doi.org/10.48550/arXiv.2311.14199

[16] Chen, Q., Zhao, R.S., Wang, S.N., Phan, V M.H., et al. (2024). A survey of medical vision-and-language applications and their techniques. arXiv preprint arXiv:2411.12195. https://doi.org/10.48550/arXiv.2411.12195

[17] Liu, Y.L., Zhang, G.Y., Ma, C., Gu, Z.Y. (2024). Dual graph convolutional networks for chest X-ray report generation. Procedia Computer Science, 247: 1416-1425. https://doi.org/10.1016/j.procs.2024.10.170

[18] Li, M.J. (2023). Exploring clinical knowledge to enhance deep learning models for medical report generation. ProQuest Dissertations & Theses. University of Technology Sydney (Australia). http://hdl.handle.net/10453/171249.

[19] Satla, S., Manchala, S. (2021). Dialect identification in Telugu language speech utterance using modified features with deep neural network. Traitement du Signal, 38(6): 1793-1799. https://doi.org/10.18280/ts.380623

[20] Senior, H., Slabaugh, G., Yuan, S.X., Rossi, L. (2025). Graph neural networks in vision-language image understanding: A survey. The Visual Computer, 41: 491-

516. https://doi.org/10.1007/s00371-024-03343-0
[21] Tian, Y.H., Su, C., Duan, J.W., Song, Y. (2025). Computed tomography visual question answering with cross-modal feature graphing. arXiv preprint arXiv:2507.04333. https://doi.org/10.48550/arXiv.2507.04333
[22] Messina, P., Pino, P., Parra, D., and Soto, A., et al. (2020). A survey on deep learning and explainability for automatic report generation from medical images. arXiv preprint arXiv:2010.10563. https://doi.org/10.48550/arXiv.2010.10563
[23] Vardhini, P.A.H., Prasad, S.S., Vishnu Sai, M.H.S., Santoshi, C., Konduru, D. (2024). Pioneering minimalist speech analysis through optimized spectral features machine learning models. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 1-6. https://doi.org/10.1109/ESCI59607.2024.10497288