

A COMPARATIVE STUDY OF DEEP LEARNING APPLICATIONS IN MEDICAL REPORT GENERATION USING CHEST X-RAYS

LIKITHA P¹, KRISH S SHAH¹, KSHITIJ AGARWAL¹, MAANASA GOWDA¹, SHARATH VISHWANATH²

¹Department of Computer Science and Engineering, PES University

²Department of Electrical and Electronics Engineering, PES University

E-MAIL: likithap@pesu.edu, [krishshah.krish17, kshitijagarwal2808, maanasa1203, sharathvishwanath46]@gmail.com

Abstract:

The automation of chest X-ray report generation has emerged as a transformative area in medical imaging, with the potential to alleviate radiologists' workload and enhance diagnostic accuracy and efficiency in healthcare. This paper presents a comprehensive survey of deep learning approaches aimed at automating the creation of clinical reports from chest X-rays, focusing on state-of-the-art methodologies, evaluation techniques, and real-world applicability.

To offer a holistic view the survey covers key model architectures, including Convolutional Neural Networks (CNNs), Transformers, and multimodal frameworks, providing an in-depth exploration of their capabilities and limitations in medical report generation. Notably, models like ViT-GPT2 and ResNet101+Tranformer achieve promising results, with ViT-GPT2 reporting a BLEU-4 score of 0.2020 and ResNet101+Tranformer reporting a BLEU-3 Score of 0.1765 post the proposed novel report standardization. The paper examines publicly available datasets such as IU-X Ray and MIMIC-CXR, which serve as benchmarks for training and evaluating these systems, alongside commonly used assessment metrics. The paper underscores challenges inherent to these datasets, such as biases, limited clinical diversity, while also discussing strategies to address these issues.

This survey highlights not only the progress made in this field but also the gaps and unresolved questions like clinical relevance and practical concerns that hinder widespread adoption. By synthesizing insights from existing studies, this work aims to guide researchers, clinicians, and developers in advancing the field of AI-driven chest X-ray report generation. It also identifies promising directions for future research and development, contributing to the broader goal of enhancing radiological practice and improving patient outcomes through trustworthy and effective AI solutions.

Keywords:

Chest X-ray, Medical report generation, Deep learning in radiology, Transformers in medical imaging, Medical natural language processing (NLP), Automated radiology reports, Vision-language models, Artificial intelligence in medical diagnostics

1 Introduction

Chest X-rays have emerged as critical diagnostic tool used to detect thoracic conditions such as pneumonia, tuberculosis, and COVID-19 due to their accessibility, low cost, and quick results [1, 2]. Their utility in early disease detection, particularly in emergency and resource-limited settings, was evident during the COVID-19 pandemic, where they played a key role in patient triage and monitoring [3]. Studies have shown their high diagnostic accuracy for diseases like tuberculosis and pneumonia, reinforcing their role in global health initiatives [4].

However, manual interpretation of chest X-rays is time-consuming and reliant on radiologist expertise as shown in Table 1. Rising imaging volumes [5] and radiologist shortages have led to delays [4, 6], diagnostic errors, and widespread burnout [7]. These challenges have driven the need for automated, accurate, and efficient reporting systems.

TABLE 1. Radiologists per 100,000 population across countries

Country	Radiologists per 100,000
Ghana	0.2
India	1.0
Singapore	7.6
United Kingdom	9.9
United States of America	12.5
Europe	13.0
Organisation for Economic Co-operation and Development (OECD) Average	12.8

AI, from early CAD [8] systems in the 1960s to modern deep learning, has shown promise in improving diagnostic workflows by delivering impressive performance in interpreting medical images and generating clinical reports [9, 10].

Real-world deployments like the qXR v2.1 software that has analyzed over 1.3 million chest X-rays across 33 UAE visa screening centers, achieving a 99.92% Negative Predictive Value with 88.2% of surveyed radiologists reported reduced turnaround times, and 82% saw improved diagnostic accuracy which validate the progress in this domain [11, 12].

Our key contributions are:

- A comparative technical study of state of the art CNN+Stacked LSTM, ResNet+Transformer, and ViT+GPT-2 architectures.
- Quantitative and qualitative analysis using BLEU, ROUGE, and METEOR metrics.
- Standardization techniques for consistent medical terminology in report generation.
- Assess practical applications of AI in radiology workflows.
- Highlight key challenges, including burnout and reporting delays
- Discussion of real-world deployment challenges and ethical considerations. [7, 11, 12].

2 Literature Survey

Exploration in the field of chest X-ray report generation requires a thorough investigation of the key concepts, including datasets, the evolution of algorithmic approaches, and the evaluation metrics that measure the performance and clinical relevance of these systems.

1. Domain:

Chest X-ray report generation is often framed as a specialized form of image captioning, wherein the goal is to generate descriptive text summarizing key clinical findings from radiographic images. Early efforts in this domain adopted convolutional neural networks (CNNs) for visual feature extraction and recurrent neural networks (RNNs) for sequence generation [13, 14]. These systems, while effective at generating brief descriptions, struggled with medical coherence and failed to capture region-specific abnormalities or inter-sentence dependencies. For instance, early CNN-RNN models generated sentence fragments that often lacked clinical depth or diagnostic accuracy [15].

To mitigate these shortcomings, hierarchical LSTM architectures were introduced to better model the report’s structural complexity by capturing both word- and sentence-level dependencies. These methods led to improved report quality by generating longer, semantically richer descriptions and capturing more nuanced clinical contexts [16]. Attention mechanisms further enhanced performance by enabling the models to selectively focus on salient image regions, aligning visual and textual features more effectively. The integration of self-attention and cross-attention modules marked a pivotal shift towards more clinically coherent report generation [7].

Building on these foundations, transformer-based architectures have emerged as the dominant paradigm. Vision-Language Models (VLMs) unify image understanding and language generation within a shared embedding space and have demonstrated robust performance in tasks like visual question answering and radiology report generation. Recent adaptations like the Medical-VLBERT [17] and MATNet [18] highlight the effectiveness of combining medical knowledge and visual cues through alternate learning strategies and multimodal transformers. However, these models often require extensive datasets and may overfit when trained on small-scale clinical corpora.

Hybrid models that combine retrieval and generation offer a promising direction for improving flexibility and accuracy. For example, the Hybrid Retrieval-Generation Reinforced Agent [15] dynamically chooses between retrieving template sentences and generating novel ones, using reinforcement learning to optimize decision-making and enhance clinical fluency. Similarly, AIMNET [19] employs adaptive mechanisms to balance visual and textual inputs, reducing bias and improving alignment with radiological findings.

Knowledge-grounded models also play an important role. Approaches such as “When Radiology Report Generation Meets Knowledge Graph” [20] incorporate domain-specific ontologies to guide report generation. While effective at improving interpretability, such models may struggle with adaptability due to their dependence on static medical knowledge graphs. In contrast, newer frameworks such as dynamic graph-enhanced contrastive learning [21] leverage evolving graph structures to align image-report pairs more flexibly, addressing domain drift and dataset shifts.

Another innovation involves the integration of strong encoders like ConvNeXt in conjunction with biomedical language models such as BioBERT. The CNX-B2 model [22] exemplifies this hybrid CNN-Transformer approach, showing superior performance across standard benchmarks like BLEU, METEOR, and CIDEr. Likewise, models such as Improving Chest X-Ray Report Generation by Leveraging Warm Starting [23]

demonstrate how initializing from pretrained vision and language checkpoints, including ViT and PubMedBERT, can enhance convergence and output quality.

Moreover, recent advancements like the Memory-Guided Transformer [24] address limitations in traditional transformers by incorporating spatio-semantic visual extractors. These networks are capable of fine-grained feature localization and better semantic reasoning, crucial for accurate clinical report generation. The use of deformable attention and semantic encoders allows such models to generalize well across different radiographic patterns and datasets.

Other contributions, such as the GRU-based encoder-decoder framework [25], although simpler, have been explored for their efficiency and adaptability. Models like S4M [26] further broaden the scope by generalizing report generation across multiple body parts, relying on cross-modal alignment and body-part specific features.

Together, these developments illustrate a dynamic and rapidly evolving research landscape. The trajectory from early CNN-RNN systems to sophisticated transformer-based models reflects a consistent drive toward generating clinically meaningful, coherent, and accurate medical reports. As architectures grow in complexity and incorporate retrieval, multimodal attention, domain knowledge, and memory-guided mechanisms, the field continues to push toward higher diagnostic utility and real-world applicability.

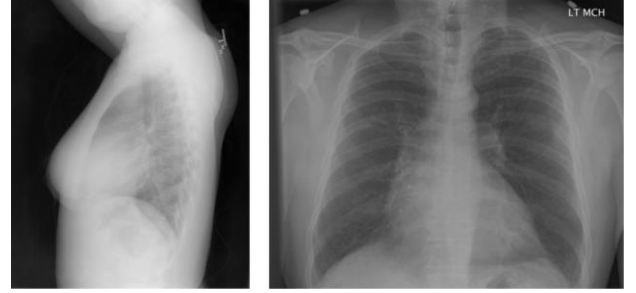
2. Dataset Used

The Indiana University - X ray dataset more popularly known as IU-Xray Dataset [27], a widely used publicly available dataset, contains 7,470 chest X-ray images along with corresponding radiology reports from 3,955 unique patients. The number of images for every unique patient varies from 1 to 4, most patients have 2 images, 1 corresponding for each chest X-ray view. These reports are comprehensive, including sections such as findings, impression, and caption, which guide the model in learning the relationship between visual abnormalities in X-rays and the appropriate textual descriptions.

3. Evaluation Metrics

Automated report generation in healthcare requires metrics that assess both natural language quality and clinical accuracy. Common evaluation metrics include:

- **BLEU (Bilingual Evaluation Understudy):** BLEU, introduced by Papineni et al., measures the precision of n-gram overlaps between generated and reference texts. It is



COMPARISON:None.

INDICATION: Preop bariatric surgery.

FINDINGS: Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lungs. Inferior XXXX XXXX XXXX.

IMPRESSION: No acute pulmonary findings.

FIGURE 1. An example of a chest X-ray with its associated report from IU-XRAY

computed as:

$$\text{BLEU} = \text{BP} \times \exp \left(\frac{1}{N} \sum_{n=1}^N w_n \log p_n \right) \quad (1)$$

BLEU is effective for syntactic accuracy but has limitations in capturing semantic meaning and clinical relevance.

- **ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence):** Proposed by Lin, ROUGE-L evaluates the longest common subsequence (LCS) between generated and reference texts, emphasizing sequence-level matching. It computes:

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{|Y|} \quad (2)$$

ROUGE-L better captures sentence-level fluency and coherence compared to simple n-gram overlap.

3 Methodology

This study presents three distinct approaches to medical report generation using deep learning models. Each approach was designed to investigate the potential of different architectures and techniques in generating coherent and informative medical reports.

A. CNN + Stacked LSTM Network

The first study employed a Convolutional Neural Network (CNN) [28] combined with stacked Long Short-Term Memory (LSTM) [29] network to generate medical reports.

The CNN, DenseNet121 [30] pretrained on CheXpert [31], was used as a feature extractor, processing the input data into a compact representation that was then fed into the LSTM unit for sequential text generation. This architecture was inspired by previous work on natural language generation tasks, where the CNN was used to extract features from images and the LSTM provided temporal dependencies between words [32].

The input data for this study consisted of reports annotated with relevant information such as text descriptions. Each report was preprocessed by representing each word in the vocabulary using pretrained GloVe word embeddings (300d). The reports were split into training, validation, and testing sets following a 75:15:10 split.

A two-level Stacked LSTM network model was trained on the training set to learn the relationship between the visual features and the textual report. The model iteratively processed the word embeddings and the feature vector, using the knowledge gained at each step to predict the next word in the report sequence. During training, the loss function used was cross-entropy loss for sequence prediction, and the Adam optimizer with a learning rate of 10^{-4} was employed.

The performance of the model on the validation set was monitored to prevent overfitting, and training stopped after 10 epochs when the validation loss plateaued. The generated reports were evaluated using BLEU Score and ROUGE Score metrics.

In some cases, the model generated incomplete or inconsistent reports due to missing or ambiguous information in the X-rays. This was particularly evident when critical features were unclear in the images, impacting the overall coherence of the generated text.

The combination of DenseNet-121 and GloVe embeddings worked well in extracting both visual and textual features, leading to coherent and contextually accurate reports. The LSTM decoder helped effectively map the image features to the text. The model also showed higher BLEU and ROUGE scores for cases with two associated X-rays, indicating the importance of multiple views.

B. ResNet101 + Transformer

The second study investigated the use of a ResNet101 [33] pre-trained model as a feature extractor, followed by a Transformer [34] encoder for text generation. This architecture was

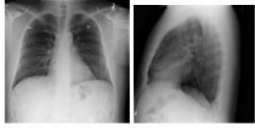
Image : CXR18_IM-0520	Ground Truth	Report generated by DenseNet121+Stacked LSTM
	Heart size within normal limits. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.	the lungs and pleural spaces show no acute abnormality . heart size and pulmonary vascularity within normal limits .

FIGURE 2. Ground truth and sample output generated by the DenseNet121 + Stacked LSTM model .

inspired by previous work on image-to-text tasks, where the ResNet101 was used to extract visual features from images and the Transformer provided efficient processing capabilities. The ResNet101 employed in this study consisted of 50 convolutional layers with batch normalization and ReLU activation functions. The output from the last layer was then fed into a Transformer encoder with 6 layers, each consisting of an attention mechanism and fully connected feed-forward networks. The number of self-attention heads used in the Transformer was set to 8. During our experimentation we realised the varying linguistics in the reports present, thus text standardization was carried out wherein the synonym vocabulary provided by domain experts was used to replace synonyms with standardized medical terms, ensuring consistency in tone, vocabulary, and structure across generated reports. Each report was tokenized using the NLTK library [35] and preprocessed by removing redundant information and normalizing the text to lowercase. The preprocessed reports were then split into training (80%) and testing sets (20%).

The model was trained using the Adam optimizer with a learning rate of 0.001, and the loss function used was cross-entropy loss between the predicted output and the true labels. The performance of the model was evaluated using BLEU, ROUGE-L, and METEOR metrics.

Effectively representing large radiology images with patch-level features while retaining contextual information proved challenging. Additionally, balancing relational memory updates with the complexity of integrating memory-driven normalization layers in the decoder added further difficulty. Another challenge was mitigating variations in sentence structure, vocabulary, and tone to ensure that the generated reports remained coherent and consistent.

Using ResNet101 for patch feature extraction resulted in robust and discriminative visual embeddings, improving the model's ability to generate accurate reports. The incorporation of relational memory allowed for dynamic context-aware updates, leading to more coherent multi-sentence reports. To

address challenges with diverse vocabulary and inconsistent phrasing, a report-standardization step was introduced, leading to significant improvements in BLEU and ROUGE scores. This step enhanced alignment with reference texts and increased overlap in critical n-grams and phrases, streamlining outputs and improving clinical relevance. Quantitative evaluation demonstrated significant improvements in BLEU and ROUGE scores, with clinical assessments confirming strong alignment with expert-annotated data.

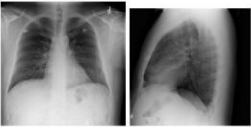
Image : CXR18_IM-0520	Ground Truth	Report generated by ResNet101 + Transformer
	Heart size within normal limits. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.	the lungs and pleural spaces show no acute abnormality . heart size and pulmonary vascularity within normal limits .

FIGURE 3. Ground truth and sample output generated by the ResNet101 + Transformer.

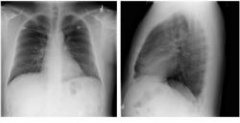
Image : CXR18_IM-0520	Ground Truth (standardized report)	Report generated by ResNet101 + Transformer (standardized report)
	heart size within normal limits . no focal pulmonary pulmonary alveoli consolidation no definite pleural effusion seen . no typical findings lung edema . no pneumothorax .	heart size within normal limits stable mediastinal and hilar contours . no focal pulmonary pulmonary alveoli consolidation no definite pleural effusion seen . no typical findings lung edema . no pneumothorax .

FIGURE 4. Ground truth and sample output generated by the ResNet101 + Transformer with Standardized Reports.

C. ViT + GPT-2

The third study explored the use of a Vision Transformer (ViT) [34] to generate medical report summaries. This architecture was inspired by previous work on vision-to-text tasks, where the ViT was used to extract visual features from images and the GPT-2 provided efficient processing capabilities. The ViT employed in this study consisted of 12 self-attention layers with a hidden state size of 128. The output from the last layer was then fed into a GPT-2 decoder with 24 layers, each consisting of an attention mechanism and fully connected feed-forward networks.

The input data for this study consisted of medical reports annotated with relevant information such as patient demographics, diagnosis, and treatment plans. Each report was tokenized

using the NLTK library [35] and preprocessed by removing redundant information and normalizing the text to lowercase. The preprocessed reports were then split into training (80%) and testing sets (20%).

The model was trained using the Adam optimizer with a learning rate of 0.001, and the loss function used was cross-entropy loss between the predicted output and the true labels. The performance of the model was evaluated using BLEU, ROUGE-L, and METEOR metrics. A key highlight of the training pipeline was the integration of synonym standardization during preprocessing. This step played a pivotal role in improving the consistency of input data, leading to notable gains in evaluation metrics. By unifying semantically similar expressions, the model achieved better alignment with reference reports and improved coherence in the generated outputs demonstrated the value of targeted preprocessing enhancements in radiology report generation tasks.

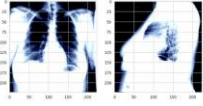
Image : CXR865_IM-2385-1001	Ground Truth	Report generated by ViT + GPT2
	Cardiac silhouette is normal in size. Normal mediastinal contour and pulmonary vasculature. The lungs are without focal airspace consolidation, large pleural effusion, or pneumothoraces.	cardiomediastinal silhouette is within normal limits. lungs are clear without areas of focal consolidation no pneumothorax or large pleural effusion

FIGURE 5. Ground truth and sample output generated by the ViT + GPT-2.

4 Results

We evaluate three model architectures across key metrics to understand their strengths and limitations in radiology report generation. The results demonstrate how architectural design and preprocessing choices influence learning dynamics and report quality.

A. CNN + Stacked LSTM

The CNN + Stacked LSTM model generated reports that captured basic observations but often lacked clinical detail. BLEU and ROUGE scores reflected moderate alignment with references, revealing limitations in handling long-range dependencies and diverse terminology.

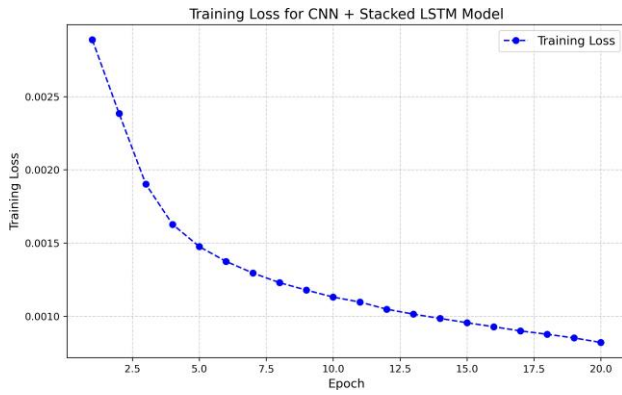


FIGURE 6. Training loss for CNN + Stacked LSTM model.

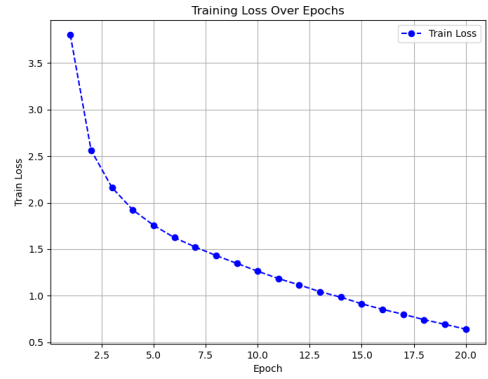


FIGURE 8. Training loss for ResNet101 + Transformer.

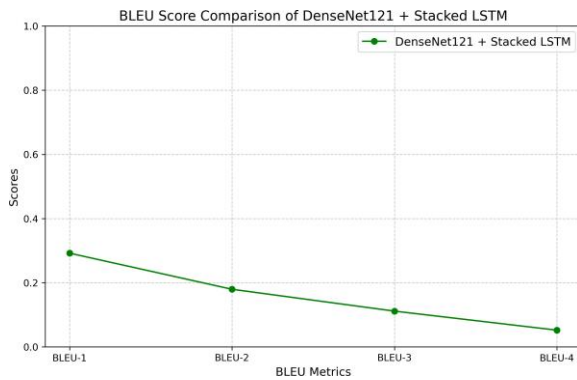


FIGURE 7. Validation BLEU scores for CNN + Stacked LSTM model.

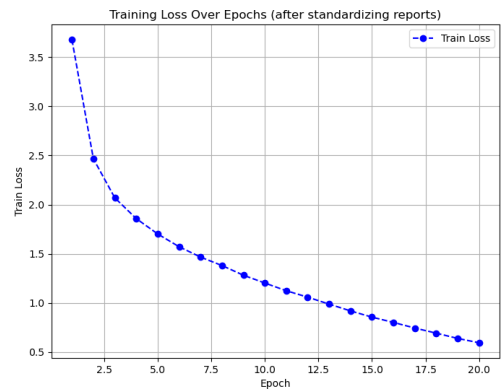


FIGURE 9. Training loss after report standardization.

B. ResNet101 + Transformer

This model outperformed the CNN + LSTM approach, generating more coherent and structured multi-sentence reports. Report standardization improved vocabulary consistency and yielded higher BLEU, ROUGE, and METEOR scores.

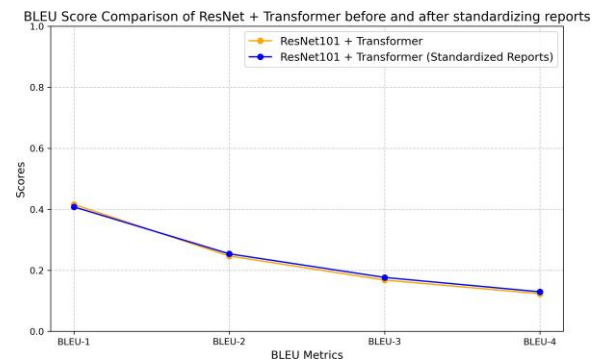


FIGURE 10. BLEU score before and after standardization.

C. ViT + GPT-2

The ViT + GPT-2 model excelled at capturing global context and generating detailed, structured narratives. A synonym-standardization step further improved scores, and this model achieved the highest performance across all metrics.

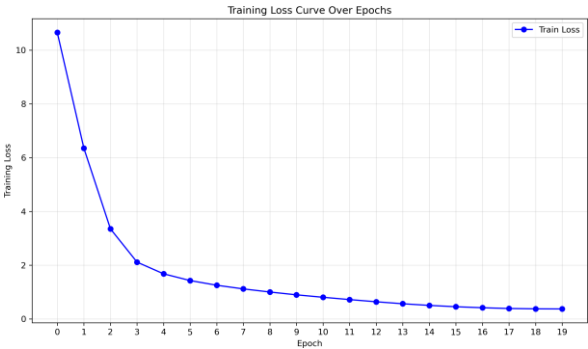


FIGURE 11. Training loss for ViT + GPT-2.

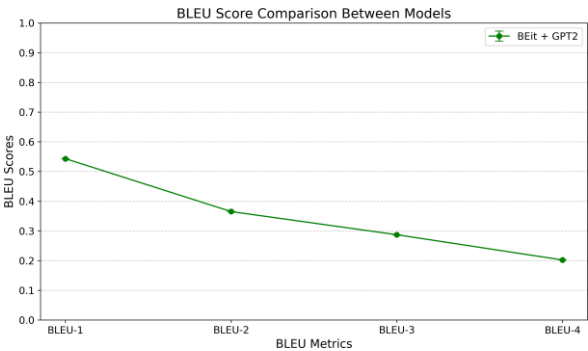


FIGURE 12. Validation BLEU scores for ViT + GPT-2.

D. Comparative Analysis

Figure 13 shows BLEU score comparisons across all models. ViT + GPT-2 led overall, followed by the standardized ResNet101 + Transformer. CNN + LSTM trailed, highlighting the advantage of transformer-based vision-language models.

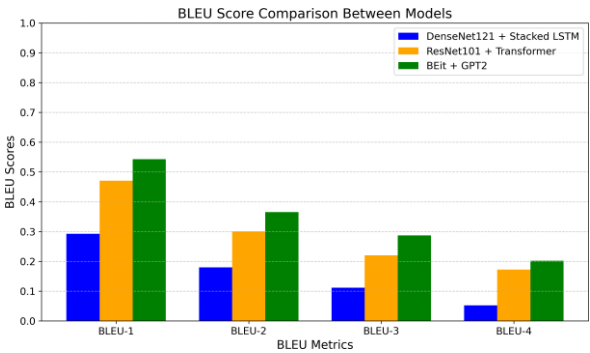


FIGURE 13. BLEU score comparison across ViT + GPT2, CNN + Stacked LSTM and ResNet101 + Transformer.

The sample outputs (Figure 14) illustrate qualitative differences, with ViT + GPT-2 producing the most clinically accurate and coherent descriptions.


Image : CXR865_IM-2385-1001	Ground Truth	Report generated by DenseNet121+Stacked LSTM	Report generated by ResNet101 + Transformer with Relational Memory	Report generated by ViT + GPT2
	Cardiac silhouette is normal in size. Normal mediastinal contour and pulmonary vasculature. The lungs are clear without focal consolidation, large pleural effusion, or pneumothorax.	the heart normal size. the mediastinum unremarkable. the lungs are clear.	the cardiomedastinal silhouette and vasculature are within normal limits for size and contour. the lungs are normally inflated and clear. osseous structures are within normal limits for patient age.	cardiomedastinal silhouette is within normal limits. lungs are clear without areas of focal consolidation no pneumothorax or large pleural effusion.

FIGURE 14. Qualitative comparison of generated reports across CNN + Stacked LSTM, ResNet101 + Transformer and ViT + GPT2 illustrating differences in clinical accuracy and coherence

This figure underscores the progressive improvement in report generation with increasingly sophisticated architectures, demonstrating how advanced vision-language models like ViT + GPT-2 enable more clinically relevant and precise outputs, aligning better with expert standards.

TABLE 2. Performance comparison of BLEU scores for different models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
DenseNet121 + Stacked LSTM	0.2920	0.1795	0.1114	0.0521
ResNet101 + Transformer	0.4157	0.2468	0.1680	0.1231
ResNet101 + Transformer with Standardized Reports	0.4077	0.2539	0.1765	0.1294
ViT + GPT-2	0.5430	0.3650	0.2870	0.2020

5 Related Work

The use of pretrained language models has significantly advanced automatic chest X-ray (CXR) report generation, especially those adapted to biomedical and radiological domains. These models typically serve as encoders or decoders in multi-modal systems, improving both the linguistic and clinical quality of generated reports.

5.1 General-Purpose vs. Domain-Specific Language Models

Initial work in this area leveraged general models such as BERT [36] and GPT-2 [37], which, despite their versatility, struggle with medical language due to vocabulary mismatches and lack of domain exposure [38].

Domain-adapted variants like BioBERT [38] and ClinicalBERT [39] addressed this by continuing BERT's training on biomedical and clinical texts. However, since they retain BERT's original vocabulary, they are limited in handling specialized radiological terms [40].

5.2 PubMedBERT

PubMedBERT [41] was trained from scratch on over 18 billion biomedical tokens from PubMed and PMC, using a domain-specific tokenizer. This enables more accurate representation of medical terms and improved performance across biomedical NLP tasks. Its alignment with clinical language makes it a strong candidate for initializing decoders in CXR report generation pipelines [42].

5.3 RadBERT

RadBERT [43], in contrast, is trained solely on 2.7 million radiology reports across modalities like X-ray, CT, and MRI. Its vocabulary and tokenizer, derived entirely from radiology text, better capture the abbreviated and telegraphic style of such reports. It outperforms general and biomedical models on radiology-specific tasks like semantic similarity and natural language inference, making it well-suited for radiology report generation.

5.4 Implications for Report Generation

PubMedBERT and RadBERT highlight the value of domain-specific pretraining. While PubMedBERT supports broader biomedical understanding, RadBERT offers specialized adaptation to radiological language. Incorporating these models—either as warm-start encoders or within hybrid architectures—can improve the *accuracy*, *coherence*, and *clinical validity* of automated CXR reports [42, 43]. Future studies may benefit from directly comparing these models in full-generation pipelines for radiology.

6 Conclusion

This study examined the development and evaluation of automated radiology report generation systems, leveraging deep learning advancements across vision and language domains. Starting from foundational architectures like DenseNet121 + Stacked LSTM, we iteratively explored more sophisticated models such as ResNet101 + Transformer and Vision-Language Transformers (VLMs), observing substantial improvements in evaluation metrics like BLEU, ROUGE, and CIDEr. The ViT + GPT-2 model consistently outperformed earlier approaches, highlighting the power of tightly coupled vision-language modeling. However, despite these gains, there remains notable room for improvement in contextual fluency, clinical accuracy, and interpretability of generated reports.

Beyond model design, this work emphasizes the importance of data quality, clinical relevance, and trust in AI-driven systems. Academic research in this area not only drives innovation but also plays a crucial role in enhancing diagnostic workflows, supporting radiologists, and promoting clinical education. As the field evolves, future directions should focus on fine-tuning VLMs for richer contextual understanding, integrating dynamic knowledge graphs for up-to-date medical insight, and leveraging ensemble learning for region-specific radiology tasks. There is also great potential in addressing dataset limitations through self-supervised and federated learning, and extending model capabilities across modalities like CT and MRI. Additionally, embedding explainability and uncertainty quantification will be key to building clinician trust and facilitating real-world deployment.

Ultimately, automated report generation stands at the intersection of AI innovation and clinical impact. With careful attention to model robustness, ethical considerations, and practical integration, these systems can become valuable tools for enhancing diagnostic accuracy, improving healthcare delivery, and supporting medical professionals in delivering better patient outcomes.

References

- [1] M. S. Ahmed, A. Rahman, F. AlGhamdi, S. AlDakheel, H. Hakami, A. AlJumah, Z. AlIbrahim, M. Youldash, M. A. Alam Khan, and M. I. Basheer Ahmed, "Joint diagnosis of pneumonia, covid-19, and tuberculosis from chest x-ray images: A deep learning approach," *Diagnostics*, vol. 13, no. 15, p. 2562, 2023.
- [2] M. S. Ahmed, A. Rahman, F. AlGhamdi, S. AlDakheel, H. Hakami, A. AlJumah, Z. AlIbrahim, M. Youldash,

- M. A. Alam Khan, and M. I. Basheer Ahmed, "Joint diagnosis of pneumonia, covid-19, and tuberculosis from chest x-ray images: A deep learning approach," *Diagnostics*, vol. 13, no. 15, p. 2562, 2023.
- [3] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," in *Machine Learning for Healthcare Conference*, pp. 249–269, PMLR, 2019.
- [4] N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell, "A systematic review of fatigue in radiology: is it a problem?," *American Journal of Roentgenology*, vol. 210, no. 4, pp. 799–806, 2018.
- [5] M. Winder, A. J. Owczarek, J. Chudek, J. Pilch-Kowalczyk, and J. Baron, "Are we overdoing it? changes in diagnostic imaging workload during the years 2010–2020 including the impact of the sars-cov-2 pandemic," in *Healthcare*, vol. 9, p. 1557, MDPI, 2021.
- [6] T. Malikowski, M. Mahmood, T. Smyrk, L. Raffals, and V. Nehra, "Tuberculosis of the gastrointestinal tract and associated viscera," *Journal of clinical tuberculosis and other mycobacterial diseases*, vol. 12, pp. 1–8, 2018.
- [7] C. X. Y. Goh and F. C. H. Ho, "The growing problem of radiologist shortages: perspectives from singapore," *Korean Journal of Radiology*, vol. 24, no. 12, p. 1176, 2023.
- [8] P. H. Meyers, C. M. Nice Jr, H. C. Becker, W. J. Nettleton Jr, J. W. Sweeney, and G. R. Meckstroth, "Automated computer analysis of radiographic images," *Radiology*, vol. 83, no. 6, pp. 1029–1034, 1964.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [10] Y. Huang, H. Xue, B. Liu, and Y. Lu, "Unifying multi-modal transformer for bi-directional image and text generation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1138–1147, 2021.
- [11] A. A. Mohamed AlJasmi, H. Ghonim, M. E. Fahmy, A. M. Nair, S. Kumar, D. Robert, A. A. Mohamed, H. Abdou, A. Srivastava, and B. Reddy, "Post-deployment performance of a deep learning algorithm for normal and abnormal chest x-ray classification: A study at visa screening centers in the united arab emirates," *Available at SSRN 4905858*.
- [12] C. J. Liew, P. Krishnaswamy, L.-T. Cheng, C. H. Tan, A. C. Poh, and T. C. Lim, "Artificial intelligence and radiology in singapore: championing a new age of augmented imaging for unsurpassed patient care," *Ann Acad Med Singapore*, vol. 48, no. 1, pp. 16–24, 2019.
- [13] K. O'Shea, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.
- [14] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [15] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, vol. 31, 2018.
- [16] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.
- [17] G. Liu, Y. Liao, F. Wang, B. Zhang, L. Zhang, X. Liang, X. Wan, S. Li, Z. Li, S. Zhang, *et al.*, "Medical-vlbart: Medical visual language bert for covid-19 ct report generation with alternate learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3786–3797, 2021.
- [18] C. Shang, S. Cui, T. Li, X. Wang, Y. Li, and J. Jiang, "Matnet: Exploiting multi-modal features for radiology report generation," *IEEE Signal Processing Letters*, vol. 29, pp. 2692–2696, 2022.
- [19] J. Shi, S. Wang, R. Wang, and S. Ma, "Aimnet: Adaptive image-tag merging network for automatic medical report generation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7737–7741, IEEE, 2022.
- [20] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 12910–12917, 2020.
- [21] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3334–3343, 2023.

- [22] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. Truong, C. D. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, *et al.*, “Benchmarking saliency methods for chest x-ray interpretation,” *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022.
- [23] A. Nicolson, J. Dowling, and B. Koopman, “Improving chest x-ray report generation by leveraging warm starting,” *Artificial intelligence in medicine*, vol. 144, p. 102633, 2023.
- [24] P. Divya, Y. Sravani, C. Vishnu, C. K. Mohan, and Y. W. Chen, “Memory guided transformer with spatio-semantic visual extractor for medical report generation,” *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [25] W. Akbar, M. I. U. Haq, A. Soomro, S. M. Daudpota, A. S. Imran, and M. Ullah, “Automated report generation: A gru based method for chest x-rays,” in *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–6, IEEE, 2023.
- [26] Q. Chen, Y. Xie, B. Wu, M.-S. To, J. Ang, and Q. Wu, “S4m: Generating radiology reports by a single model for multiple body parts,” *arXiv preprint arXiv:2305.16685*, 2023.
- [27] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. A. Rodriguez, D. Milentijevic, E. Apostolova, S. Antani, G. R. Thoma, and C. J. McDonald, “Preparing a collection of radiology examinations for distribution and retrieval,” *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.
- [28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- [29] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- [31] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Illcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpan-skaya, *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.
- [32] B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” *arXiv preprint arXiv:1711.08195*, 2017.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [34] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [35] E. Loper and S. Bird, “Nltk: The natural language toolkit,” *arXiv preprint cs/0205028*, 2002.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- [37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [38] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [39] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.
- [40] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, and B. C. Wallace, “Does bert pretrained on clinical notes reveal sensitive data?,” *arXiv preprint arXiv:2104.07762*, 2021.
- [41] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-specific language model pretraining for biomedical natural language processing,” *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

- [42] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest x-ray report generation by leveraging warm starting," *Artificial intelligence in medicine*, vol. 144, p. 102633, 2023.
- [43] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e210258, 2022.