Advances in Natural Language Processing for Radiology: State-of-the-Art Techniques, Applications, and Open Challenges

Kotha Chandrakala[1], Dr. Shahin Fatima[2]
Research scholar, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India[1]
Assistant Professor, Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India[2]

***Abstract*—Radiology reports encode critical clinical observations from medical imaging in an unstructured textual form that is central to modern clinical diagnosis and decision support. In this context, natural language processing (NLP) has emerged as a key clinical NLP technology for automatically extracting, classifying, and interpreting information from radiology reports. This study presents a structured review of more than sixty recent contributions on NLP for radiology, covering approaches that range from traditional rule-based pipelines to contemporary deep learning and transformer-based models. We examine how deep learning architectures, including BERT, GPT-4, multimodal transformers, and vision–language alignment networks, are applied to core tasks such as disease classification, tumor response assessment, cancer phenotype extraction, radiology report generation, cohort identification, quality assurance, and longitudinal patient follow-up. Particular attention is given to knowledge graph integration, multimodal cross-attention, and zero-shot learning strategies that adapt large language models to radiology-specific workflows. We also analyze key barriers to clinical adoption, including limited annotated data, domain generalization gaps across institutions, ethical and fairness concerns, and the need for transparent model explainability. Based on this synthesis, the review outlines future research directions for building interpretable, multimodal, and clinically robust NLP solutions that integrate technological, clinical, and operational perspectives to advance radiology report analysis and medical imaging–driven care.**

Keywords*—Natural language processing; radiology reports; deep learning; transformers, medical imaging; report generation; clinical NLP

I. Introduction

The integration of artificial intelligence (AI) into clinical workflows has reshaped healthcare delivery, particularly in medical imaging and diagnostic interpretation. Among the various AI techniques, natural language processing (NLP) has become a key approach for extracting clinically meaningful information from unstructured text, especially radiology reports. These reports are rich in contextual and diagnostic detail but are often ambiguous and stylistically heterogeneous, which limits their direct use in automated systems. NLP helps bridge the gap between free-text documentation and structured representations that can support robust clinical decision-making.

Radiology reports play a central role in clinical decision-making by translating complex imaging findings into concise narrative summaries. The number of imaging examinations and the corresponding volume of textual interpretations continue to increase, making it challenging for clinicians to assimilate all available information efficiently. NLP-based systems assist with tasks such as automated diagnosis [1], identification of additional or incidental findings [4], tumor response classification [9], and assessment of disease progression [14]. These tools can reduce cognitive workload, improve consistency, and support timely decision-making. Over time, radiology NLP has evolved from rule-based approaches to advanced deep learning and transformer-based models such as BERT and, more recently, GPT-4.

Recent developments in deep learning and multimodal architectures have further expanded the possibilities for NLP in radiology. Models that align image and text representations, integrate radiologists' eye-gaze patterns, or employ cross-modal transformers offer improved performance and interpretability compared with earlier pipelines [25], [56], [29]. Hybrid systems that combine convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and semantic evaluators have also shown strong results in radiology report generation and related downstream applications [22].

Despite these advances, several challenges hinder widespread clinical deployment. Model performance often deteriorates when applied to institutions other than the ones in which they were developed, due to differences in report style, structure, and terminology. The scarcity of large, high-quality annotated datasets limits the ability to train and validate models across diverse patient populations and imaging modalities. Concerns about data privacy, algorithmic bias, and lack of transparency remain, especially for deep learning–based systems used in high-stakes diagnostic settings. Domain-specific language models pre-trained on biomedical corpora (e.g., BioBERT, PubMedBERT) alleviate some of these issues, but the overall landscape remains fragmented and rapidly evolving [31].

This review concentrates on recent advances in NLP for radiology, encompassing a broad spectrum of techniques, applications, and clinical scenarios. It draws primarily on more

than 60 plus peer-reviewed studies published between 2020 and 2025 in venues such as IEEE Access, IEEE Transactions, and leading medical informatics journals. By mapping NLP methods to concrete radiology applications—including diagnostic automation, report generation, quality assurance, cohort identification, and clinical decision support—the review aims to provide a clear picture of current capabilities, limitations, and emerging directions [33].

The contributions of this work are threefold. First, it offers an updated synthesis of radiology-focused NLP methods, spanning rule-based systems, classical machine learning, deep learning, transformer models, and multimodal vision–language architectures. Second, it introduces a structured taxonomy that links technical approaches to clinical use cases such as disease classification, tumor response assessment, and longitudinal patient tracking. Third, it highlights cross-cutting challenges in data quality, domain generalization, explainability, and ethical deployment, and distills practical implications for developing clinically robust and trustworthy NLP solutions in radiology.

Unlike earlier reviews that primarily focus on individual application areas (e.g., breast cancer, de-identification, or report generation), this study provides a unified taxonomy that links techniques (rule-based, deep learning, transformers, foundation models) to end-to-end radiology workflows, including quality assurance, cohort identification, and longitudinal tracking. It also synthesizes very recent work (2020–2025) on multimodal vision–language models, synthetic data, and privacy-preserving training, which are not comprehensively covered in prior surveys.

The remainder of this study is organized as follows: Section II reviews NLP techniques applied to radiology reports, from foundational rule-based pipelines to modern deep learning and transformer-based models. Section III summarizes major application areas and representative use cases of radiology NLP. Section IV discusses technical approaches and methodological aspects, including data pre-processing, feature representation, model architectures, training strategies, and evaluation metrics. Section V provides a discussion of clinical translation, challenges, and limitations. Section VI outlines emerging trends and future research directions. Finally, Section VII concludes the study by synthesizing key insights and reflecting on the implications for future radiology-focused clinical NLP research and practice.

II. Background and Related Work: NLP Techniques in Radiology

A. Rule-Based and Statistical Methods

The initial NLP systems employed in the field of radiology made use of rule-based frameworks, which benefited from handcrafted lexical rules for extracting structured information from narrative reports. The clarity and ability to interpret these methods make them good candidates for clinical audits and quality control tasks [9], [13]. However, their problems with static ontologies and sensitivity to linguistic diversity limited scalability and generalizability. Techniques such as n-grams and TF-IDF were useful in document classification and term weighting, but only correlated according to the study [7].

B. Classical Machine Learning Approaches

With the help of machine learning (ML) models like support vector machines (SVMs), decision trees, and logistic regression, tagging of diseases associated with radiology reports resulted in semi-automated cohorts [6], [14]. A central role is played by feature engineering with bag-of-words, syntactic dependencies, and domain-specific lexicons. Machine learning (ML) models had greater adaptability than rule-based systems, but these models often relied on shallow representations of language and were not interpretable (see [8]).

C. Deep Learning and Hybrid Architectures

The introduction of deep neural networks, such as CNNs and RNNs, helped build richer contextual representations in radiologic natural language processing (NLP). CNN models worked well for spatial semantics performance, while LSTM architectures were effective for temporal dependency modelling of sequential clinical narratives [17], [22]. The combination of CNN-LSTM or attention-based RNN for various tasks, such as CT reporting in emergencies and response classification tasks [22], [26]. The clinical explainability challenge requires large annotated corpora despite the improved performance of the models [23].

D. Transformers and Foundation Models

Transformer architectural designs, such as BERT, BioBERT, and Clinical BERT, changed the paradigm of radiology NLP. Models pre-trained on large biomedical corpora can learn bidirectional contextual semantics better [21], [54]. Fine-tuned variants have been used successfully for tumor response classification [9], cancer report prioritization [10], and impression summarization [30]. In recent times, large language models LLMs like GPT-4 and ChatGPT have been used for radiology reporting, and they offer promise for zero-shot and few-shot, cross-institutional generalization, and adaptation [36], [60]. However, there are ongoing risks of hallucination, fairness, and clinical safety issues.

TABLE I. Summary of NLP Use Cases in Radiology

Application Area	Techniques Used	Representative Studies	Clinical Benefit
Disease Classification	Rule-based NLP, Transformers	[2], [6], [9], [10]	Faster triage, improved coding accuracy
Tumor Response & Phenotyping	BERT, LSTM, Attention Networks	[1], [3], [5], [14]	Personalized treatment planning
Report Generation	Multimodal Transformers, Reinforcement Learning	[20], [21], [25], [28], [37]	Streamlined documentation, reduced workload
Cohort Identification	Rule-based NLP, Knowledge Graphs	[4], [7], [14], [15]	Efficient trial recruitment, population health
Quality Assurance	Ontology Matching, NLP Audits	[12], [13], [16]	Standardized reports, error detection
Longitudinal Tracking	State-space Models, Cross-modal RNNs	[17], [34], [35], [40]	Time-based outcome prediction

The main NLP use cases in radiology identified in the reviewed literature—including disease classification, tumor response assessment, report generation, cohort identification, quality assurance, and longitudinal tracking—are summarized in Table I, and a comparative overview of rule-based, classical machine learning, deep learning, and transformer-based methods is provided in Table II.

TABLE II. COMPARATIVE SUMMARY

Technique	Strengths	Weaknesses	Representative Studies
Rule-based	Transparent, interpretable	Poor scalability, rigid syntax	[9], [13]
ML Models	Adaptable, efficient	Requires manual features	[6], [14]
Deep Learning	Captures complex context	Data-hungry, opaque	[22], [26], [23]
Transformers	Generalizable, state-of-the-art	Computationally intensive, ethical concerns	[21], [30], [54], [60]

Fig. 1 illustrates a generic NLP pipeline for radiology reports, from raw text input to downstream clinical tasks.

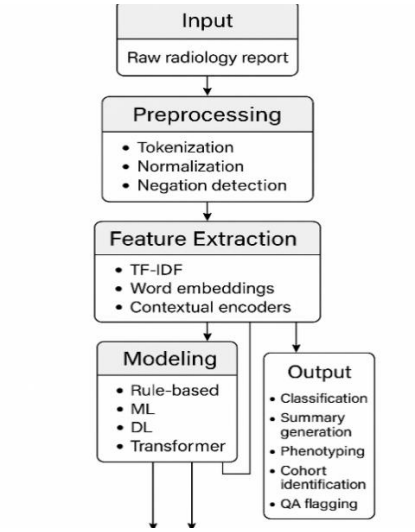

Fig. 1. NLP pipeline for radiology reports.

III. RESULTS AND APPLICATIONS OF NLP IN RADIOLOGY

Natural language processing (NLP) has emerged as a powerful tool to convert unstructured radiology narratives into structured ones. As radiology report volumes are exploding, the NLP-based solutions enable scalable, precise analysis for different diagnostic, operational, and research use cases. This section discusses important uses of NLP in radiology from various literature sources, as well as future perspectives.

A. Automated Disease Classification

One of the primary applications of NLP in radiology is the automatic classification of diseases from narrative reports. The techniques that have been used for the classification of pulmonary embolism, abdominal aortic aneurysms, and metastatic lesions are rule-based as well as machine-learning-based. Gaviria-Valencia et al. created a near-real-time NLP model to detect abdominal aortic aneurysms, which achieved a good recall and precision in clinical deployment scenarios [2]. Likewise, Laurent et al. employed a rule-based NLP pipeline within oncology workflows to classify tumor response outcomes in a timely manner [9].

BERT and other transformer-based models improved the classification accuracy of transformer models. Ayanouz et al. showed that the transformer-based NLP is better than traditional methods for prioritizing oncological radiology reports for review [10]. Moreover, Ong et al. introduced a neural NLP system that effectively identified disease responses from free-text radiology reports, with good agreement with a manual expert assessment [6].

B. Tumor Response and Cancer Phenotyping

NLP is used in oncology to help extract phenotypic information as well as track tumors. Fink et al. used deep learning for evaluating oncologic outcomes from structured radiology narratives, providing a scalable framework for retrospective outcome studies [1]. Maghsoudi and colleagues proposed a transformer-based NLP model for cancer-associated thrombosis. The model might facilitate real-time phenotype surveillance of hematologic malignancies [5].

According to the review, most of the breast cancer-related NLP applications aim to extract information related to tumour characteristics, receptor status, and treatment response using imaging studies [3]. In addition, predictive oncology is aided by the use of NLP pipelines to create digital twins of metastatic progression using longitudinal radiology report data [14] [18].

C. Radiology Report Generation

Development of a report is one of the most active areas of research in the field of NLP, where a report-generating system creates a coherent report from the relevant information. Recent models use vision-language mechanisms, reinforcement learning, and attention methods. The visual-textual cross-modal interaction network designed by Zhang et al. allows contextual alignment of image features with generated text, improving the fluency and relevance of reports greatly. Likewise, Yi et al. proposed a hybrid reward-based learning model that improves the quality of language outputs while preserving diagnostic specificity [28] [39].

Various strategies have emerged for report synthesis, such as multi-phase supervision, hierarchical classification, and gaze-guided alignment. GPT-4-based transformer encoding models were used by Olivato et al. [21], which result in strong language understanding capabilities, improving adherence to medical conventions and report completeness.

D. Cohort Identification and Surveillance

Natural language processing helps find groups of patients easily for clinical trials and more. According to Sekler et al., NLP has immense potential in identifying secondary findings buried in text (PET/CT reports) and largely ignored in oncological cohort's manual review [4]. According to Donnelly et al. [7], Natural language processing (NLP) could enable automated audits of electronic medical records to assist with disease surveillance, if potentially scaled out across hospital systems [19].

Batch et al. further investigated NLP-driven cohort extraction through metastasis detection pipelines that combined

structured sequences of radiology reports, which aid in building cancer digital twins [14].

In addition, Paul et al. showed that cohort curation was ethically compliant when de-identification models were built on radiology reports while preserving data utility [15].

E. Quality Assurance and Clinical Auditing

It is important to have quality and uniformity in radiological documentation for diagnosis and accreditation. Fei et al. developed a framework that uses NLP to monitor the quality of reports of pulmonary nodules. It highlighted variations in the vocabulary used and the completeness of reports. Kale et al. used knowledge graph techniques to detect inconsistencies in radiologist dictation to standardize reports and facilitate retrieval [12].

These NLP tools can be used for internal auditing within the institution. They also allow benchmarking against national or institutional reporting. According to Yu et al., NLP and machine learning models must undergo external validations before they may be used or integrated into a clinical quality system [16].

F. Longitudinal Analysis and Outcome Prediction

In addition to static analysis, NLP facilitates the long-term monitoring of disease progression and treatment response. Chuang et al. introduced a machine learning framework that organized CT liver tumor reports over time, making treatment response tracking easier and interpretable [17]. Zhang et al. used state-space modeling to create radiology reports sequentially to ensure temporal, spatial, and temporal dependencies [35].

Gao et al. applied multi-modal representation learning to predict the response to neoadjuvant therapy in breast cancer from its radiology text. Here, insights from NLP are aligned with therapeutic planning [40]. Likewise, Song et al. proposed a temporal residual network that includes time-sensitive report features to enhance prediction accuracy in radiologic sequences [34].

The datasets examined in this review draw on both large, publicly available chest X-ray collections and institution-specific oncology cohorts. Widely used resources such as MIMIC-CXR, CheXpert, and OpenI offer substantial volumes of radiology images and accompanying reports for model development, while in-house hospital and oncology datasets typically provide CT, MRI, and PET/CT reports with more detailed phenotypic and tumor response information. Table III outlines the main public and private datasets included in the surveyed studies, along with their imaging modalities and annotation characteristics.

TABLE III. PUBLIC/PRIVATE DATASETS FOR RADIOLOGY NLP

Dataset Name	Type	Modality	Annotation Type
MIMIC-CXR	Public	X-ray	Findings
CheXpert	Public	X-ray	Labels
OpenI	Public	X-ray	Reports
Private Hospital A	Private	CT, MRI	Phenotypes
Private Oncology Group B	Private	PET, CT	Tumor Response

IV. RESEARCH METHODOLOGY AND TECHNICAL APPROACHES

In radiology, natural language processing (NLP) draws on a wide range of technical strategies, including both traditional machine learning and modern deep learning. Each family of methods varies in its ability to represent clinical language, cope with variation in reporting styles, and remain robust across different healthcare institutions. In practice, radiology NLP workflows can be viewed as passing through five main stages: data preprocessing, linguistic analysis, feature construction, model training, and performance evaluation.

A. Data Preprocessing and Normalization

This study highlights the basics of NLP, specifically the first step in carrying out NLP on Medical text. Radiology reports often include abbreviations, negations, terms of anatomy, and ungrammatical forms. The text can be segmented properly using tokenization, sentence segmentation, and part-of-speech tagging. Techniques such as negation detection (NegEx and its variants) are specifically important in the medical domain to determine present or absent findings [13], [59].

Mapping synonyms to medical ontologies, such as SNOMED-CT or RadLex, also belongs to Normalization. According to some studies, Information extraction (IE) is used to convert free text phrases to structured representations for downstream machine learning systems.

B. Feature Engineering and Representation

Standard approaches use handcrafted features like term frequency-inverse document frequency (TF-IDF), bag-of-words, and n-gram statistics [7]. Early implementations were effective, but weaker in semantics, and failed to capture the connection between entities. The move towards distributed representations – word embeddings (Word2Vec, GloVe) and contextual embeddings (ELMo, BERT) – enhances the expressivity of the input features [21], [54].

In radiology contexts, embeddings adapted to domains like clinical BERT or Blue BERT trained on biomedical corpus have shown much improvement in performance on classification and generation tasks [17], [24]. The availability of models that project image and text features into shared cross-modal embeddings has made it easier to compare and align human and automated annotations. [25], [26], [56].

Table IV groups the models used in the reviewed studies into major categories and counts how many works fall into each category.

TABLE IV. CATEGORIZATION OF NLP MODELS USED IN STUDIES

Model Category	Example Algorithms	No. of Studies
Rule-based	Regex, Ontologies	6
Classical ML	SVM, RF	4
Deep Learning	CNN, RNN	10
Transformers	BERT, GPT	20
Multimodal	ViT-BERT, CLIP	8

C. Modeling Architectures

The architecture selection often depends on the task. Conventional methods for classification include support vector machines (SVM), random forests (RF), and logistic regression [6], [14]. Use of deep learning models on unstructured inputs has seen an increase in popularity over traditional methods.

Recurrent neural networks (RNNs) and their gated variants [e.g., recurrent neural networks, long short-term memory (LSTM), gated recurrent unit (GRU)] are commonly used to model the sequence of radiology text. CNN-LSTM hybrids have been utilized for the interpretation of emergency CT reports [22], while attention-based RNNs enhance focus on clinically salient terms [26]. The most researched models in recent literature are transformers like BERT, RoBERTa, and GPT-based models due to their contextual understanding and transfer learning capabilities [10], [21], [30].

Multimodal models incorporate both visual and textual data. Architecture involving eye-gaze guided networks [25], hierarchical attention transformers [21], and semantic fusion [56] facilitates cross-modal alignment of radiology images and reports.

Fig. 2 depicts the main technical components and data flow within radiology NLP architectures.

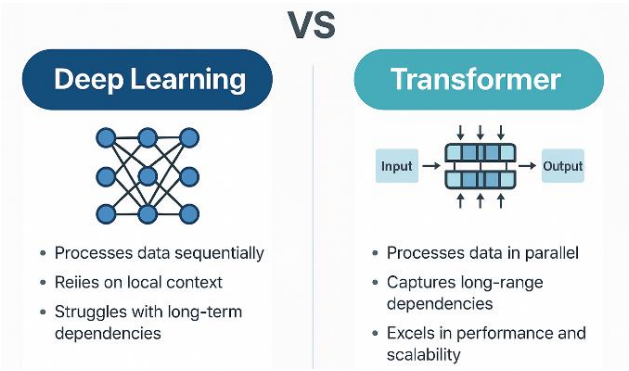

Fig. 2. Radiology NLP architectures: Technical components and data flow.

D. Model Training and Optimization

Training models of NLP in radiology faces challenges of class imbalance, noisy annotations, and limited labeled datasets. To mitigate these, researchers adopt techniques such as.

- Data augmentation refers to paraphrasing, back-translation, and synonym replacement.
- Transfer Learning: Fine-tuning pretrained models like BioGPT and GPT-4 on clinical corpora [36], [21].
- Semi-supervised Learning uses pseudo-labels or weak supervision on big, unlabeled datasets.
- Reinforcement learning was used for report generation (to enhance the quality of generated reports) with task-specific reward functions [28], [37].

Optimization methods may range from simple gradient descent to complex pieces such as AdamW, and regularization (dropout, weight decay) may be applied to prevent overfitting.

E. Evaluation Metrics and Benchmarking

Assessment of radiology NLP models requires both quantitative and qualitative evaluation. For tasks involving classification and extraction, standard metrics are precision, recall, F1-score, and AUC-ROC [2], [9], [11]. In the context of report generation, the BLEU, ROUGE, and METEOR scores are used to measure the linguistic similarity of reports with ground-truth text [20], [59].

There may be clinical relevance that these metrics miss. So, expert-in-the-loop validation, qualitative scoring by radiologists, and utility-based evaluation (e.g., diagnostic usefulness) are gaining more and more interest [16], [23], [24].

Table V outlines the key stages in typical radiology NLP pipelines, along with representative techniques and example studies.

TABLE V. KEY TECHNICAL STAGES IN RADIOLOGY NLP PIPELINES

Stage	Key Techniques	Challenges Addressed	Representative Works
Preprocessing	Tokenization, NegEx, Lexical mapping	Noise reduction, ambiguity handling	[13], [59], [62]
Feature Representation	TF-IDF, Word2Vec, BERT, Cross-modal embeddings	Semantic encoding, domain adaptation	[7], [21], [25], [56]
Modeling	SVM, CNN-LSTM, BERT, GPT-4, Hierarchical Transformers	Accuracy, context-awareness, scalability	[6], [22], [30], [54], [60]
Training	Fine-tuning, Reinforcement Learning, Semi-supervised	Data scarcity, generalization	[27], [28], [36], [61]
Evaluation	F1, AUC, BLEU, Human scoring	Performance, clinical relevance	[9], [16], [23], [59]

V. DISCUSSION: CLINICAL TRANSLATION, CHALLENGES, AND LIMITATIONS

In this section, we examine how radiology-focused NLP systems are being incorporated into clinical practice and highlight the main barriers reported in the literature. The discussion centers on data quality and variability, domain adaptation and robustness, interpretability and clinician trust, ethical and regulatory issues, and the practical challenges of embedding these tools into everyday radiology workflows.

Radiology and natural language processing (NLP) have advanced considerably, but scaling their use in routine clinical practice still faces major obstacles. The restrictions include technical, infrastructural, ethical, and regulatory. To move research-grade models into healthcare, we need to know these challenges.

A. Data Quality and Heterogeneity

Variations in radiology reports exist between institutions, modalities, and radiologists. It is difficult to develop generalized NLP models because the reporting styles, terminologies, and report structures vary. Further, many public

datasets are small or insufficiently diverse to enable transfer across institutions [16], [24], [59]. There is a lack of annotated datasets for fine-grained tasks, which include tracking tumor response and rare disease classification [5], [6].

Also, clinical documentation contains a lot of abbreviations, misspellings, and vague phrases. These problems cause a decline in performance of the rule-based and traditional machine learning models, which require robust pre-processing and normalization pipelines [13], [62].

To illustrate how datasets and methods align in practice, Table VI highlights representative combinations of application areas, datasets, and NLP techniques.

TABLE VI. KEY APPLICATIONS, DATASETS, AND METHODS

Application Area	Dataset Used	NLP Techniques
Disease Classification	MIMIC	Rule-based, ML
Tumour Monitoring	Private Oncology B	LSTM, BERT
Report Generation	CheXpert	GPT-4, RLHF
Cohort ID	OpenI	Keyword Search, Graph NLP
QA & Audits	Institutional Reports	BERT + Ontology

B. Domain Adaptation and Model Generalization

Pretrained language models (e.g., BERT, GPT) show good performance on benchmark datasets, but often underperform when deployed in new hospital systems, due to domain shift [24], [60],[63]. Phase post-training methods can be applied to fine-tune Clinical BERT, since their flexible design is suitable for diverse tasks.

Although powerful, transformer-based models are data-hungry and overfit to institutional-specific data [36]. Also, it is still a challenge for globally-operable NLP systems to generalize across languages, cultures and regions.

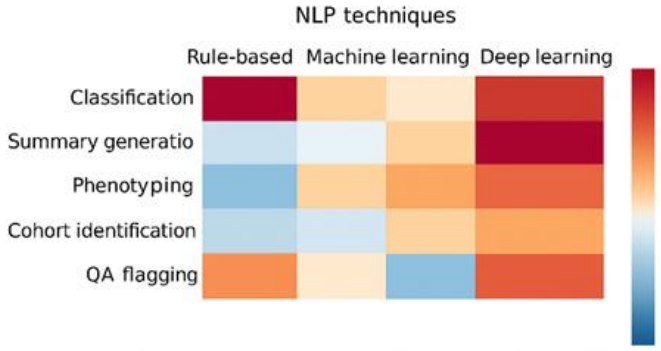

Fig. 3. Application-wise NLP techniques heatmap.

As illustrated in Fig. 3, the heatmap contrasts how frequently different NLP techniques are applied across major radiology tasks such as report generation, information extraction, and decision support. This visualization helps highlight which approaches are currently dominant in specific applications and where there is still scope for methodological exploration [52].

C. Explainability and Clinical Trust

The clinical use of NLP tools based on deep learning may face challenges not easy to overcome. Clinicians often demand explanations for automated decisions, especially in high-stakes settings like oncology or emergency [23]. Although rule-based methods are transparent, they do not scale. The adoption of transformer models in clinical workflows has been limited, despite the emergence of attention visualization, saliency mapping, and explainable AI (XAI) frameworks [27], [28].

Guidelines From U.S. And Other Countries. Essentially, the "provenance" and "autonomy" standards set by the FDA or EU MDR also require traceability and human oversight, which are not yet fully established in most deep learning NLP pipelines.

D. Ethical, Legal, and Privacy Concerns

NLP systems that work on radiology reports must abide by strict privacy regulations such as HIPAA in the U.S. or GDPR in Europe. Even though automated de-identification tools have been improved, there remains a risk for re-identification, especially when text is paired with imaging metadata. If certain demographics or disease types are underrepresented in the training data, resulting models may encode bias and deliver uneven performance across patient groups [59].

Improperly constrained large language models (LLMs) can hallucinate clinically inaccurate content, which raises serious safety concerns [30, 36]. Models are fine-tuned on specific data.

E. Integration into Clinical Workflows

Integrating NLP models into radiology information systems (RIS) or electronic health records (EHR) poses various and unique challenges. Legacy systems compatibility, real-time processing, and usability are some of the issues. According to Laurent et al. [9] and Sekler et al. [4], tight integration is important for adoption success. The models should also be efficient in delivering output while not affecting model workload and reporting workflows.

Automation only works well when there is a match in user interface, design, and time of decision [7], [23].

Taken together, these issues show that technological improvements by themselves are not enough. For radiology NLP tools to be used reliably in everyday care, they need rigorous external evaluation, clear governance frameworks, and joint design with radiologists, other clinicians, and hospital IT teams. The next section extends this discussion by describing emerging trends and research directions that seek to overcome these challenges.

VI. EMERGING TRENDS AND FUTURE DIRECTIONS

As Natural Language Processing is rapidly evolving, various changing trends are modifying its role in Radiology. These advances in multimodal foundation models, explainable AI, and the generation of synthetic data will tackle current shortcomings and spur off-the-shelf clinical use. This section mentions the future directions from the recent literature.

A. Multimodal and Vision-Language Models

One major change is the creation of vision-language models that combine imaging features into their textual data sets. With the help of these architectures, the interpretation of diagnostic images and their descriptions can occur simultaneously, thereby aiding in consistent reports and appropriate characterization of disease. Recent works like Med3DVLM [43] [45], Eye-Gaze Guided Alignment Networks [25], and crossmodal causal representation learning [29] illustrate this trend [44].

Techniques for pretraining vision and language models, specifically contrastive learning and attention mechanisms, are intended to achieve multi-modal generalization (e.g., X-ray, CT, MRI) and clinical adaptability (e.g., [25], [42], [56]).

Recent foundation-model efforts in chest imaging and cross-modal alignment further illustrate this trajectory [44], [45], [49], [50].

Table VII summarizes common evaluation metrics used in the reviewed studies and typical performance ranges reported for each.

TABLE VII. EVALUATION METRICS AND REPORTED PERFORMANCE

Metric	Used In	Typical Score Range
Accuracy	Classification	0.85-0.95
Precision	Detection	0.80-0.93
Recall	Detection	0.75-0.92
F1 Score	Overall	0.78-0.94
BLEU	Text Gen	20-35
ROUGE	Summarization	40-55

B. Generative Language Models and Radiology Report Synthesis

The emergence of GPT-4, BioGPT, and similar large generative models permits the development of fluent, clinically relevant report generation. Models fine-tuned on specific domain data can generate summaries, impressions, and full reports from both structured and unstructured inputs [21], [30], [36]. According to Zhang et al. [20] and Yi et al. [28], generating text through guided reinforcement learning and using a hybrid reward function can help maintain the factuality of generated text.

Additionally, ChatRadio-Valuer [36] and DDaTR [34] examine longitudinal context modeling to provide a consistent narrative across follow-up images, which is essential in oncology and chronic disease.

Several new radiology report generators refine longitudinal coherence and task-specific optimization across diverse datasets [34], [39], [51]–[53].

C. Explainable and Transparent NLP Models

The need for interpretability remains a clinical trust and regulatory concern. The new methods embed explanatory features within the model output. For instance, the observation-state-aware report generator [38] and the reward-based attentive mechanism [28] mark clinician-relevant tokens or phrases, providing insight into the decisions. Other works use knowledge-enhanced graph representations [12], [40] to track inferences and enable auditability [41].

As these methods become more developed, they can be incorporated into PACS viewers or RIS dashboards to provide radiologists with real-time, interpretable aids.

D. Self-Supervised and Few-Shot Learning Paradigms

Data annotation in medical NLP is resource-intensive. Self-supervised learning (SSL), implemented through masked language modeling or contrastive pretraining on unlabeled radiology corpora, and is increasingly used in this domain [24], [26]. The implementation of S3-Net [26] and zero-shot grounding technologies [46] [47] demonstrates how SSL allows for flexible reuse of models across tasks with minimal supervision.

Training foundation models on medical imaging data opens up a vast number of possible tasks. Artemova et al. showed that it is possible to do few-shot and prompt-based learning in the context of radiology [42]. For example, Lu et al. [48] demonstrated the feasibility of prompt-based learning for medical visual question answering using the BLIP language–vision model. This can make large foundational models accessible to resource-poor hospitals facing rare disease scenarios [45].

E. Integration with Real-Time Clinical Decision Support

Future NLP systems will be integrated and used in the clinical workflow in real-time. Some types of clinical decision support services are point-of-care decision support, triage alerts, and automated billing code assignment. For example, automatically identifying the priority of reports and flagging incidental findings can impact patient safety and throughput.

Technologies such as iterative feedback loops with radiologists [30] and systems built into emergency settings [22] are already demonstrating potential improvements in efficiency without increasing cognitive burden.

F. Synthetic Data Generation and Privacy-Preserving Models

The creation of synthetic data with the help of text generators and image augmentation can counter the issues associated with privacy and the shortage of data. Models such as LHR-RFL [28] and GAN-based simulators facilitate controlled variation of training samples that mitigate overfitting and enhance robustness.

In addition, federated learning frameworks and privacy-aware architectures (e.g., differential privacy and secure multi-party computation) are being investigated for cross-institutional model training without centralizing patient data [15], [32].

Fig. 4 outlines the evolving NLP landscape in radiology, mapping emerging trends such as large language models, multimodal integration, and real-time clinical decision support. It highlights how these advances are expected to reshape reporting workflows, reduce clinician burden, and enable more personalized, data-driven imaging care in the near future.

Privacy-preserving training and federated learning are being actively explored to address data sharing constraints [15], [32], [55], [57], [58].

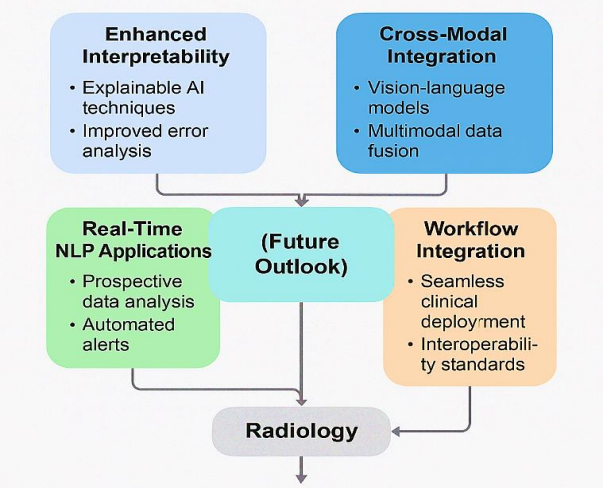

Fig. 4. Future outlook: NLP landscape in radiology.

VII. Conclusion

Over the past decade, natural language processing has become a central component of research at the interface of radiology, medical imaging, and clinical informatics. The studies reviewed in this study show that NLP can meaningfully support clinical workflows by automating parts of disease classification, tumor response assessment, radiology report generation, cohort identification, quality assurance, and longitudinal tracking of patient outcomes. When combined with electronic medical records and modern machine learning, radiology NLP systems have the potential to reduce manual effort, improve consistency, and surface clinically relevant information that might otherwise remain buried in free-text reports.

At the same time, multiple barriers continue to limit routine clinical deployment. Performance often drops when models are applied to data from new institutions or imaging protocols, reflecting differences in report style, vocabulary, and documentation practices. The scarcity of large, carefully annotated datasets constrains both model development and robust external validation. Concerns about data privacy, fairness, and the opacity of complex deep learning and transformer-based models add further friction, particularly in high-stakes diagnostic settings. Addressing these issues requires not only better algorithms but also clearer standards, transparent evaluation, and close collaboration between radiologists, data scientists, and hospital IT teams.

Looking ahead, several technical and methodological trends are likely to shape the next phase of radiology NLP. Multimodal vision–language models, self-supervised and weakly supervised learning, and robustness-enhancing strategies such as domain adaptation and adversarial training offer avenues for building systems that generalize more reliably across populations and institutions. Integration with knowledge graphs, structured clinical ontologies, and explainable AI techniques can help make model outputs more interpretable and clinically trustworthy. If these advances are coupled with rigorous benchmarking, prospective validation, and careful consideration of ethical and regulatory constraints, NLP tools can move from research prototypes to dependable components of everyday radiology practice.

Ultimately, real progress will depend on sustained cooperation across disciplines, updated guidelines for data curation and evaluation, and continuous monitoring of deployed systems in real-world settings. With a more integrated and clinically grounded approach, radiology-focused NLP can evolve from promising experimental systems into widely adopted tools that support radiologists and clinicians in delivering timely, accurate, and equitable care.

References

[1] Fink, M., et al. (2022). Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports. *Radiology: Artificial Intelligence*, 4(2). https://doi.org/10.1148/ryai.210148

[2] Gaviria-Valencia, S., et al. (2022). Near real-time natural language processing for the extraction of abdominal aortic aneurysm diagnoses from radiology reports: Algorithm development and validation study. *JMIR Medical Informatics*, 10(7), e35618. https://doi.org/10.2196/35618

[3] Saha, A., et al. (2023). A scoping review of natural language processing of radiology reports in breast cancer. *Frontiers in Oncology*, 13, 1112454. https://doi.org/10.3389/fonc.2023.1112454

[4] Sekler, J., et al. (2022). Identifying secondary findings in PET/CT reports in oncological cases: A quantifying study using automated natural language processing. In *2022 IEEE International Conference on Big Data* (pp. 5174–5179). IEEE. https://doi.org/10.1109/BigData55660.2022.10021128

[5] Maghsoudi, A., et al. (2023). A transformer natural language processing algorithm for cancer associated thrombosis phenotype. *Blood*, 142(Suppl. 1), 2476–2477. https://doi.org/10.1182/blood-2023-172236

[6] Ong, H. X., et al. (2023). Automated classification of disease response in radiology reports using natural language processing. *JCO Global Oncology*, 9, e2300128. https://doi.org/10.1200/GO.23.00128

[7] Donnelly, L. F., et al. (2022). Use of natural language processing in evaluation of radiology reports: An update on applications and technology advances. *Seminars in Ultrasound, CT and MR*, 43(6), 631–637. https://doi.org/10.1053/j.sult.2022.08.007

[8] Gonçalves, L. B., et al. (2022). Natural language processing and graph theory: Making sense of imaging records in a novel representation frame. *JMIR Medical Informatics*, 10(9), e38802. https://doi.org/10.2196/38802

[9] Laurent, G., et al. (2023). Automatic classification of tumour response from radiology reports with rule-based natural language processing integrated into the clinical oncology workflow. *JCO Clinical Cancer Informatics*, 7, e2300104. https://doi.org/10.1200/CCI.23.00104

[10] Ayanouz, S., et al. (2023). Unleashing the power of NLP and transformers: A game-changer in medical research and clinical practice and a revolution of medical text analysis: Case study: Cancer report classification by priority. In *2023 6th International Conference on Networking, Intelligent Systems & Security* (pp. 1–6). IEEE. https://doi.org/10.1109/NISS57604.2023.10073526

[11] Bikdeli, B., et al. (2023). Developing validated tools to identify pulmonary embolism in electronic databases: Rationale and design of the PE-EHR+ study. *Thrombosis and Haemostasis*, 123(6), 493–503. https://doi.org/10.1055/a-2022-0887

[12] Kale, K., et al. (2022). Knowledge graph construction and its application in automatic radiology report generation from radiologist's dictation. *arXiv preprint*. https://arxiv.org/abs/2212.10022

[13] Fei, X., et al. (2022). Quality management of pulmonary nodule radiology reports based on natural language processing. *Bioengineering*, 9(8), 339. https://doi.org/10.3390/bioengineering9080339

[14] Batch, K., et al. (2022). Developing a cancer digital twin: Supervised metastases detection from consecutive structured radiology reports. *Frontiers in Artificial Intelligence*, 5, 1016882. https://doi.org/10.3389/frai.2022.1016882

[15] Paul, T., et al. (2022). Utility of features in a natural-language-processing-based clinical de-identification model using radiology reports for advanced NSCLC patients. *Applied Sciences*, 12(13), 6439. https://doi.org/10.3390/app12136439

[16] Yu, A. C., et al. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, 4(5), e220061. https://doi.org/10.1148/ryai.220061

[17] Chuang, Y. H., et al. (2022). Effective natural language processing and interpretable machine learning for structuring CT liver-tumour reports. *IEEE Access*, 10, 77934–77946. https://doi.org/10.1109/ACCESS.2022.3190426

[18] Elmoufidi, A. (2022). Deep multiple instance learning for automatic breast cancer assessment using digital mammography. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–10. https://doi.org/10.1109/TIM.2022.3142016

[19] Mehta, V., & Dharia, R. (2023). Automated approach to selecting neurological medical imaging orders using natural language processing. In *2023 IEEE International Conference on Smart Healthcare and AI* (pp. 1–5). IEEE. https://doi.org/10.1109/SHCAI58232.2023.10123614

[20] Zhang, W., et al. (2024). Visual-textual cross-modal interaction network for radiology report generation. *IEEE Signal Processing Letters*, 31, 984–988. https://doi.org/10.1109/LSP.2024.3379005

[21] Olivato, M., et al. (2024). Language models for hierarchical classification of radiology reports with attention mechanisms, BERT, and GPT-4. *IEEE Access*, 12, 69710–69727. https://doi.org/10.1109/ACCESS.2024.3402066

[22] Tomassini, S., et al. (2025). Multi-branch CNN-LSTM fusion network-driven system with BERT semantic evaluator for radiology reporting in emergency head CTs. *IEEE Journal of Translational Engineering in Health and Medicine*, 13, 61–74. https://doi.org/10.1109/JTEHM.2025.3535676

[23] 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

[24] Wu, Z., et al. (2025). Exploring the trade-offs: Unified large language models vs. local fine-tuned models for highly-specific radiology NLI task. *IEEE Transactions on Big Data*, 11(3), 1027–1041. https://doi.org/10.1109/TBDATA.2025.3536928

[25] Peng, P., et al. (2024). Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(12), 7406–7419. https://doi.org/10.1109/JBHI.2024.3422168

[26] Pan, R., et al. (2024). S3-Net: A self-supervised dual-stream network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(3), 1448–1459. https://doi.org/10.1109/JBHI.2023.3345932

[27] Tao, Y., et al. (2024). Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(7), 4145–4156. https://doi.org/10.1109/JBHI.2024.3393018

[28] Yi, X., et al. (2025). LHR-RFL: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE Transactions on Medical Imaging*, 44(3), 1494–1504. https://doi.org/10.1109/TMI.2024.3507073

[29] Chen, W., et al. (2025). Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 34, 2970–2985. https://doi.org/10.1109/TIP.2025.3568746

[30] Ma, C., et al. (2024). An iterative optimizing framework for radiology report summarization with ChatGPT. *IEEE Transactions on Artificial Intelligence*, 5(8), 4163–4175. https://doi.org/10.1109/TAI.2024.3364586

[31] Ahmad, E. (2020). *Cardiovascular diseases detection using machine learning algorithms. International Journal for Research in Applied Science and Engineering Technology, 8*(12). https://doi.org/10.22214/ijraset.2020.6376

[32] Ahmad, F., & Kumar, P. (2024). *An update of new/potential cardiovascular markers: A narrative review. Molecular Biology Reports.* https://doi.org/10.1007/s11033-023-08978-1

[33] Ali, Z., Naseer, N., & Nazeer, H. (2022). *Cardiovascular disease detection using deep learning approaches. International Journal of Advanced Computer Science and Applications.*

[34] Song, S., Tang, H., Yang, H., & Li, X. (2025). DDaTR: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3591364

[35] Zhang, D., Tan, Y., Qin, J., & Xiang, X. (2025). Automatic radiology report generation based on state-space model. *IEEE Journal of Biomedical and Health Informatics.* https://doi.org/10.1109/JBHI.2025.3579888

[36] Zhong, T., et al. (2025). ChatRadio-Valuer: A chat large language model for generalizable radiology impression generation on multi-institution and multi-system data. *IEEE Transactions on Biomedical Engineering.* https://doi.org/10.1109/TBME.2025.3597325

[37] Chen, Z., et al. (2025). Enhancing radiology report generation via multi-phased supervision. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3580659

[38] Yang, H., Tang, H., Song, S., & Li, X. (2025). OS-RRG: Observation state-aware radiology report generation with balanced diagnosis and attention intervention. *IEEE Transactions on Neural Networks and Learning Systems.* https://doi.org/10.1109/TNNLS.2025.3589103

[39] Deng, X., et al. (2025). MvKeTR: Chest CT report generation with multi-view perception and knowledge enhancement. *IEEE Journal of Biomedical and Health Informatics.* https://doi.org/10.1109/JBHI.2025.3581289

[40] Gao, Y., et al. (2025). Multi-modal longitudinal representation learning for predicting neoadjuvant therapy response in breast cancer treatment. *IEEE Journal of Biomedical and Health Informatics.* https://doi.org/10.1109/JBHI.2025.3540574

[41] Jin, H., Che, H., He, S., & Chen, H. (2025). A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3585765

[42] Lian, C., Zhou, H.-Y., Liang, D., Qin, J., & Wang, L. (2025). Efficient medical vision-language alignment through adapting masked vision models. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3575853

[43] Xin, Y., Ates, G. C., Gong, K., & Shao, W. (2025). Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics.* https://doi.org/10.1109/JBHI.2025.3604595

[44] Wang, F., & Yu, L. (2025). Scaling chest X-ray foundation models from mixed supervisions for dense prediction. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3589928

[45] Gao, Y., et al. (2025). Hierarchical contrastive learning for precise whole-body anatomical localization in PET/CT imaging. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3599197

[46] Zou, K., et al. (2025). Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* https://doi.org/10.1109/TPAMI.2025.3596878

[47] Vilouras, K., Sanchez, P., O'Neil, A. Q., & Tsaftaris, S. A. (2024). Zero-shot medical phrase grounding with off-the-shelf diffusion models. *IEEE Journal of Biomedical and Health Informatics.* https://doi.org/10.1109/JBHI.2024.3494246

[48] Lu, Z., Zeng, Q., Lu, M., Chen, G., & Xia, Y. (2025). Bridging the semantic gap in medical visual question answering with prompt learning. *IEEE Transactions on Medical Imaging.* https://doi.org/10.1109/TMI.2025.3580561

[49] Khan, M. S., et al. (2025). Evaluating the effect of drug loading on the acoustic response of nanobubbles in stable and inertial cavitation

regimes. *IEEE Open Journal of Ultrasonics, Ferroelectrics, and Frequency Control*. https://doi.org/10.1109/OJUFFC.2025.3618629

[50] Zannoni, E. M., Yang, C., Cai, L., Wilson, M. D., Chen, C.-T., & Meng, L.-J. (2025). The Alpha-SPECT-Mini: A small-animal SPECT system based on hyperspectral compound-eye gamma cameras. *IEEE Transactions on Radiation and Plasma Medical Sciences.* https://doi.org/10.1109/TRPMS.2025.3560558

[51] McCune, E. P., et al. (2025). Displacement-guided focused ultrasound of the median nerve modulates somatosensory evoked potentials in humans. *IEEE Transactions on Biomedical Engineering.* https://doi.org/10.1109/TBME.2025.3602291

[52] Wu, H., et al. (2025). A miniaturized multidirectional stacking ultrasound transducer for endo-bronchoscopy lung nodule ablation. *IEEE Transactions on Biomedical Engineering.* https://doi.org/10.1109/TBME.2025.3622642

[53] Zhang, W., Cai, B., Hu, J., Qin, Q., & Xie, K. (2024). Visual-textual cross-modal interaction network for radiology report generation. *IEEE Signal Processing Letters, 31,* 984–988. https://doi.org/10.1109/LSP.2024.3379005

[54] Olivato, M., Putelli, L., Arici, N., Gerevini, A. E., Lavelli, A., & Serina, I. (2024). Language models for hierarchical classification of radiology reports with attention mechanisms, BERT, and GPT-4. *IEEE Access, 12,* 69710–69727. https://doi.org/10.1109/ACCESS.2024.3402066

[55] Tomassini, S., Duranti, D., Zeggada, A., Quattrocchi, C. C., Melgani, F., & Giorgini, P. (2025). Multi-branch CNN-LSTM fusion network-driven system with BERT semantic evaluator for radiology reporting in emergency head CTs. *IEEE Journal of Translational Engineering in Health and Medicine, 13,* 61–74. https://doi.org/10.1109/JTEHM.2025.3535676

[56] Peng, P., et al. (2024). Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics, 28*(12), 7406–7419. https://doi.org/10.1109/JBHI.2024.3422168

[57] Liu, A., Guo, Y., Yong, J.-H., & Xu, F. (2024). Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Medical Imaging, 43*(7), 2657–2669. https://doi.org/10.1109/TMI.2024.3372638

[58] Yi, X., Fu, Y., Yu, J., Liu, R., Zhang, H., & Hua, R. (2025). LHR-RFL: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE Transactions on Medical Imaging, 44*(3), 1494–1504. https://doi.org/10.1109/TMI.2024.3507073

[59] 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

[60] Ma, C., et al. (2024). An iterative optimizing framework for radiology report summarization with ChatGPT. *IEEE Transactions on Artificial Intelligence, 5*(8), 4163–4175. https://doi.org/10.1109/TAI.2024.3364586

[61] Iqbal, S., Qureshi, A. N., Khan, F., Aurangzeb, K., & Akbar, M. A. (2024). From data to diagnosis: Enhancing radiology reporting with clinical features encoding and cross-modal coherence. *IEEE Access, 12,* 127341–127356. https://doi.org/10.1109/ACCESS.2024.3449929

[62] Liao, Y., Xiang, H., Liu, H., & Spasić, I. (2024). Using information extraction to normalize the training data for automatic radiology report generation. *IEEE Access, 12,* 185103–185116. https://doi.org/10.1109/ACCESS.2024.3504378

[63] Kulsoom, U., Glavin, F. G., & Bendechache, M. (2025). Natural language processing and machine learning for analysis of radiology reports: A systematic review. *IEEE Access, 13,* 112215–112254. https://doi.org/10.1109/ACCESS.2025.3582728