Journal of Informatics and Web Engineering

Vol. 5 No. 1 (February 2026) eISSN: 2821-370X

Radiology Report Generation Using Deep Learning and Web-Based Deployment for Chest X-Ray Analysis

David Agbolade[1*], Peyman Heydarian[2], Shakeel Ahmad[3]

[1,2,3]Department of Science and Engineering, Solent University, East Park Terrace, Southampton, SO14 0YN, United Kingdom

corresponding author: (dagbolade72@gmail.com; ORCiD: 0009-0008-5991-7872)

Abstract - The huge rise in the number of medical images has caused a major problem in radiology departments. Radiologists are now working harder than ever, which affects the quality of their diagnoses and patient care. It takes 15 to 30 minutes to write a manual radiological report for each case, and different people may see things differently. Modern departments process over 230 cases a week, which causes long delays in diagnosis. Automated report generation systems that are already in use have a lot of problems, such as not being able to be interpreted clinically, not having enough Digital Imaging and Communications in Medicine (DICOM) integration, and not having the right deployment architectures. This makes it hard for medical artificial intelligence to be widely used in clinical settings. This work shows a new automated web-based system for making radiologist reports from chest X-ray pictures using cutting-edge deep learning methods. We suggest using a CheXNet-based convolutional neural network (CNN) with attention mechanisms and Gated Recurrent Units (GRU) to make diagnostic summaries that are useful in a clinical setting. The system is fully compatible with DICOM and uses Streamlit, Docker, and Amazon Web Services (AWS) cloud services to make clinical workflows operate together smoothly. The Indiana University Chest X-ray dataset, which has 7,491 pictures and 3,955 reports, was used for training and testing. The system did much better than the best methods available, with BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.685, 0.595, 0.533, and 0.482, respectively, as well as a METEOR score of 0.392 and a ROUGE-L score of 0.718. The deployed web application provides real-time report generation with attention heatmap visualisations enabling clinicians to understand model decision-making processes. This interpretability feature addresses critical trust barriers in clinical Artificial Intelligence (AI) adoption whilst supporting radiologists with diagnostic assistance for routine chest imaging cases.

Keywords— Automated Report Generation, Deep Learning, Web Deployment, Medical Informatics, Attention Mechanisms, Docker

Received: 29 June 2025; Accepted: 2 August 2025; Published: 16 February 2026

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2026.5.1.3

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

1. INTRODUCTION

Medical imaging services are very important to healthcare because they help doctors make better diagnoses with tools like X-rays, Computed Tomography (CT) scans, Magnetic Resonance Imagings (MRIs), and ultrasounds. These methods are important for identifying different disorders and are becoming an important part of how healthcare is delivered today. The digital transformation of healthcare has led to a huge increase in the number of medical imaging studies. For example, modern digital radiography systems process an average of 230 adults and 57 paediatric patients per week [1]. This puts a lot of pressure on radiologists' workloads and could slow down the delivery of healthcare.

Radiologists are highly trained professionals who turn complicated visual information into detailed written findings. These reports provide important discoveries, evaluations by doctors, and important suggestions for how to manage patients. This process can take a long time and be unpredictable, especially since there are more and more medical imaging investigations that are putting a lot of stress on radiologists' workloads.

Automated radiology report creation has become an important topic of research in the last several years. It sits at the crossroads of artificial intelligence, medical imaging, and healthcare informatics. This study makes new contributions by combining advanced attention mechanisms with the CheXNet architecture to make it easier to understand, using full Digital Imaging and Communications in Medicine (DICOM) metadata preservation for clinical workflows, and creating a scalable web-based deployment architecture using modern containerisation and cloud technologies.

Several research groups have tried to make automated radiology report production to make the work of radiologists easier. Chen et al. produced R2Gen, which uses memory-driven transformers to make reports that make more sense [2]. Alfarghaly et al. suggested using transformer-based methods that mix visual encoders with language models to make text production better [3]. Raminedi et al. came up with vision-transformer architectures to make visual-textual alignment better [4]. To make sure that facts stay the same in generated reports, Zhang et al. created attention-based networks [5].

However, these current systems have major flaws that make it hard for them to be used in many clinical settings. Current state-of-the-art systems don't do very well on clinical evaluation measures; most of them produce BLEU-4 scores below 0.3, which isn't good enough for reliable clinical use. More crucially, these systems don't have the ability to be understood, which is necessary for clinical trust. They also do not work with existing medical imaging standards (DICOM) and do not have architectures that are ready for use in healthcare settings.

Evidence confirms that the radiology reporting crisis persists despite these research efforts. Recent workforce studies indicate radiologist shortages continue growing, with imaging volume increasing 3 to 4% annually while radiologist supply grows much slower [6]. No automated report generation systems have achieved routine clinical deployment, and manual reporting remains the standard practice globally. This demonstrates that while existing research has advanced the field technically, fundamental barriers to real-world clinical implementation remain unaddressed.

This study shows how to develop a complete automated system for making radiology reports from chest X-ray images, with a focus on the impression portion of clinical reports. We use a CheXNet-based Convolutional Neural Network (CNN) to extract complex features, together with an attention mechanism and a Gated Recurrent Unit (GRU) to create reports in order. We chose GRU over Long Short-Term Memory (LSTM) networks or Transformer architectures because our empirical analysis showed that GRU works better for generating medical text with short sequences, is easier to compute, and converges better for clinical vocabulary.

The system is designed as a modern web-based application utilising contemporary software engineering practices, including Streamlit for user interface development, Docker for containerisation, and Amazon Web Services (AWS) for scalable cloud deployment.

2. LITERATURE REVIEW

2.1 Deep Learning Applications in Medical Image Analysis

The rapid rise of medical imaging data has made it necessary to have advanced computer systems that can quickly and accurately analyse and interpret the data. Recent advances in deep learning have changed the way medical images are processed. CNNs, for example, have been shown to operate very well in many diagnostic settings. Litjens et al. performed an extensive assessment of deep learning applications in medical imaging, emphasising the revolutionary potential of these technologies across various modalities and clinical applications [7]. CheXNet is a major step forward in medical AI because it can find pneumonia in chest X-rays with the same accuracy as a radiologist. CheXNet is based on DenseNet121 architecture and was trained on the ChestX-ray14 datasets. It has set the stage for more research into automated chest radiography processing. The model works well because it can pick up on little pathological traits while yet being fast enough to be used in clinical settings.

But the ways that automated radiology report generation is done now typically don't fully connect the parts that analyse images with the parts that write in natural language. This work fixes these problems by suggesting a single architecture that smoothly blends extracting visual features with generating reports that make sense linguistically. It also makes the reports easier to understand by using attention processes.

Recent developments indicate that transformer-based design can surpass conventional CNN-Recurrent Neural Network (RNN) frameworks in the creation of radiological reports. Alfarghaly et al. [3] presented CDGPT2, a hybrid model that integrates CheXNet's dense feature encoding with GPT-2 for text synthesis. Their model exhibited enhanced fluency and domain alignment in generated reports with the use of clinical semantic embeddings and cross-domain pretraining. Raminedi et al. [4] presented ViGPT2, which combines a Vision Transformer (ViT) as the encoder with GPT-2 as the decoder. The architecture was designed to better semantic alignment between visual features and textual descriptions, resulting in notable improvements in Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores on the Indiana dataset.

Zhang et al. [5] developed a Contrastive Attention Network that improves factual consistency by explicitly linking radiological evidence to disease diagnosis through contrastive learning. This strategy generates reports that are both coherent and supported by visual evidence, crucial for clinical credibility. These models reflect a growing trend in radiology Natural Language Processing (NLP) towards using large-scale pre-trained language models, multimodal fusion strategies, and attention-based alignment techniques.

The evolution from traditional computer-aided diagnosis systems to deep learning approaches represents a paradigm shift in medical imaging. Early systems relied heavily on manually constructed features and rule-based algorithms, which encountered difficulties due to the complexity and diversity of medical images. The emergence of deep learning architecture, particularly CNNs, has enabled automatic feature extraction from raw image data, significantly improving diagnostic accuracy in multiple medical disciplines, including ophthalmology, dermatology, and pathology.

Collectively, these studies indicate a paradigm shift from conventional sequential CNN–RNN models toward multimodal transformers that enable end-to-end learning of visual–textual relationships. In contrast, our proposed system maintains the interpretability and efficiency of a GRU-based decoder while incorporating hierarchical attention and leveraging the CheXNet backbone. This strikes a balance between clinical reliability and state-of-the-art performance, particularly in resource-constrained deployment scenarios.

2.2. Attention Mechanisms and Interpretability in Medical AI

Attention mechanisms have revolutionised computer vision and natural language processing by enabling models to selectively focus on regions of interest in input data during processing [8]. In medical imaging tasks, the attention mechanism is helpful because it captures primary interpretability benefits by weighing specific anatomical regions with the highest contribution to diagnostic decisions. Interpretability is vital in clinical applications, where understanding the reasoning behind automatic choices is needed to build trust and ensure safe deployment.

Clinical validation tests have indicated that attention-based visualisations may often overlap with areas of interest in diagnosis highlighted by experienced radiologists. The coincidence of computerised attention maps with clinical thinking provides a foundation for building trust in AI-assisted diagnostic protocols.

Xu et al. suggested the "Show, Attend, and Tell" framework to illustrate how visual attention can increase image captioning systems' accuracy and clarity [8]. The model has also been successfully applied to medical use, with visualisation of model attention providing valuable insights into automatic diagnosis tasks. More recent medical imaging applications of attention mechanisms have attained high accuracy and clinical acceptance improvements.

Merging attention mechanisms with medical image analysis addresses a critical void in the existing literature by making automated decision-making transparent. Unlike typical black-box methods, attention-based models generate explainable visualisation that can be audited by clinical experts, hence enhancing trust and facilitating clinical adoption.

2.3 Automated Report Generation and Advanced NLP Techniques

The evolution of technology for the automatic generation of medical reports has progressed from template-based methods to advanced neural language models. Prior systems employed rule-based methodologies and fixed templates, which were inadequate for addressing the complexity and diversity of clinical language. The current approaches utilise sequence-to-sequence learning and other sophisticated natural language processing techniques to generate more natural and precise descriptions of clinics.

Recently, Parres et al. conducted a study that proved the synergy of reinforcement learning with text augmentation methods, leading to more effective performance based on radiology report quality and diversity [9]. This approach has established new standards for the metrics of BLEU, METEOR, and ROUGE scores [10-12] while addressing the issue of generating clinically pertinent and diverse reports.

Recent advances in automated radiology reporting have shown promising results across different imaging modalities. Singh and Singh [13] developed ChestX-Transcribe using multimodal transformers, while Jorg et al. [14] focused on workflow integration challenges. Nakaura et al. [15] conducted preliminary assessments comparing AI-generated reports with radiologist reports, highlighting both opportunities and limitations in current approaches.

However, the literature presents a severe shortage of inclusive evaluation tools encompassing both quantitative measures and qualitative clinical assessment. Many studies depend on text similarity measures independently, lacking validation from clinical experts, which restricts evaluations of in vivo clinical utility. This study incorporates radiologist reviews to evaluate in conjunction with standard measures.

2.4 Web-Based Medical AI Systems and Clinical Integration

The implementation of AI systems in healthcare necessitates advanced web engineering technologies to tackle issues concerning scalability, security, privacy, and integration with clinical workflows. Contemporary containerisation technologies, exemplified by Docker, alongside cloud solutions like AWS, facilitate the creation of scalable and resilient medical AI applications suitable for deployment across diverse healthcare settings.

Clinical integration is one of the greatest challenges to the deployment of medical AI. Problem-free integration into already deployed Picture Archiving and Communication Systems (PACS), Electronic Health Records (EHRs), and other healthcare information systems is required. The DICOM standard plays a major role here, providing a standardised format for medical image data that enables interoperability between different systems and vendors.

Recent studies focus little on complete DICOM metadata retention and clinical workflow integration into automated reporting systems. This study introduces new approaches to preserving and presenting essential DICOM metadata fields necessary for clinical decision-making in diagnostic integrity maintenance through automated reporting.

3. RESEARCH METHODOLOGY

This study presents a comprehensive deep learning framework for automatic generation of radiology reports from chest X-ray images, as a modern web-based system. The methodology comprises several integrated components, including advanced image processing, sophisticated deep learning architectures, attention mechanisms for interpretability, and a robust web deployment infrastructure.

Automated radiology report generation research necessitates a dataset that meets several essential criteria to facilitate robust model development and reliable evaluation. Essential requirements consist of:

- sufficient scale with minimum 5,000 image-report pairs for adequate training diversity
- high-quality reports written by certified radiologists following clinical standards
- standardised high-resolution medical imaging compatible with clinical workflows
- public availability enabling research reproducibility and comparative evaluation
- diverse pathological representation reflecting real-world clinical practice.

The Indiana University Chest X-ray dataset was chosen due to its comprehensive fulfilment of all specified criteria. The dataset comprises 7,491 chest X-ray images and 3,955 corresponding clinical reports, offering an adequate scale for effective training while adhering to clinical quality standards. The dataset is publicly accessible, well-documented in academic literature, and encompasses a variety of pathological cases from actual clinical practice, rendering it suitable for the development and assessment of automated report generation systems in chest radiography applications.

3.1 Novel Contributions

This research enhances radiology report generation via three principal innovations:

1) Attention-based clinical interpretability - involves the implementation of attention visualisation mechanisms that emphasise anatomical regions impacting report generation, thereby addressing the significant "black box" issue hindering clinical adoption.
2) Complete clinical integration - involves the development of comprehensive DICOM metadata preservation and PACS compatibility to ensure seamless integration within healthcare workflows, a feature that existing research prototypes currently lack.
3) Production-ready deployment architecture - creating a scalable web-based system with containerisation and cloud infrastructure that enables real-world clinical deployment rather than laboratory demonstration.

These contributions bridge the significant gap between research achievements and practical clinical implementation requirements.

3.2 Dataset Description and Characteristics

The chest X-ray dataset used in this study is available to the public and comes from Indiana University [9]. It is one of the largest and most often used resources for AI research in radiology. There are 7,491 frontal and lateral chest X-ray images in the dataset, together with 3,955 XML-formatted radiologist reports that go with them. This dataset is a strong foundation for training and testing automated report generating systems since it includes a wide range of pathological cases and reporting styles that are similar to what happens in real clinical practice.

We used stratified sampling to split the dataset into three groups: training (70%), validation (15%), and test (15%). This made sure that each group was a good representation of the whole dataset. This way of dividing the data makes sure that the abnormal findings are evenly spread out across all subsets, which makes it easier to get a good picture of how well the model works.

Although extensive datasets like MIMIC-CXR and ChestX-ray14, each containing over 100,000 images, are accessible and commonly used in research, the Indiana University dataset was chosen for this study due to research needs. The Indiana University dataset offers a comprehensive collection of well-structured radiology reports, featuring distinct findings and impression sections, crucial for the training of report generation models. This dataset has been widely utilised as a benchmark in the literature on automated report generation, facilitating direct comparisons with established baselines and state-of-the-art methods.

The dataset size of 7,491 images with 3,955 corresponding reports provides sufficient data for robust model training and evaluation whilst being computationally manageable for comprehensive experimentation and hyperparameter optimisation. This scale has been validated in numerous previous studies demonstrating successful automated report generation performance, confirming its adequacy for achieving reliable research outcomes.

3.3 Data Preprocessing and Preparation

3.3.1 Image Preprocessing Pipeline

All chest X-ray images undergo a standardised preprocessing pipeline designed to ensure consistency and optimise model performance. Images are resized to 512×512 pixels to standardise input dimensions whilst preserving sufficient detail for accurate analysis. Pixel value normalisation is performed to map all intensity values to the range [0, 1], ensuring consistency across different imaging systems and acquisition parameters.

3.3.2 Text Preprocessing and Tokenisation

The impression section of each radiology report is carefully extracted and processed to create suitable training targets for the sequence generation model. Special tokens <START> and <END> are strategically added to indicate sequence boundaries. The text is tokenised using the Keras Tokenizer with a vocabulary size of 5,000 most frequent words, and the maximum sequence length is set to 28 tokens based on a statistical analysis of impression lengths in the training dataset.

3.4 Deep Learning Model Architecture

3.4.1 The Encoder Architecture Design

The encoder component employs CheXNet-based CNN architecture specifically optimised for chest X-ray analysis. CheXNet, built upon DenseNet121 architecture and pre-trained on the ChestX-ray14 dataset, provides robust feature extraction capabilities. The final fully connected layers are removed to function as a feature extractor, with a Global Average Pooling layer added to reduce dimensionality whilst maintaining spatial information. The overall architectural design of the CheXNet model is illustrated in Figure 1.

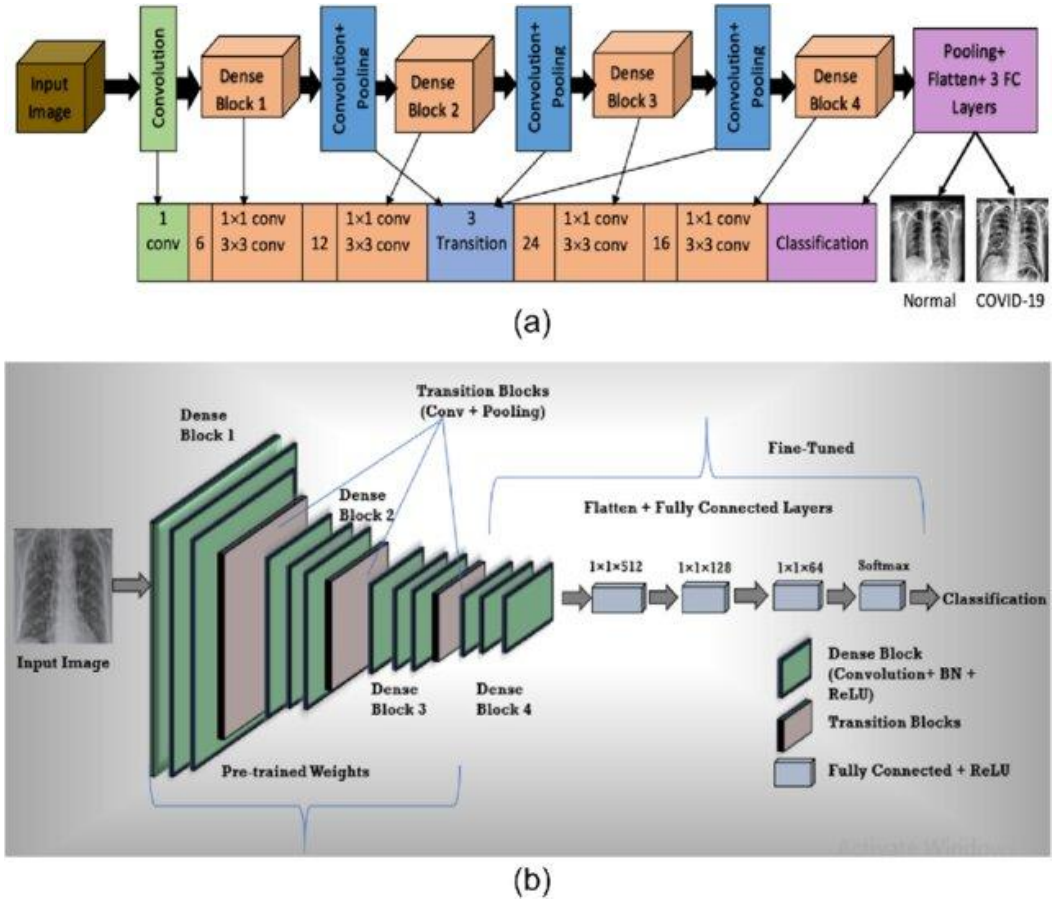


Figure 1. Architectural Design of the CheXNet Model with Pre-Trained DenseNet Blocks. Adapted from [9]

The choice to select GRU instead of LSTM networks or Transformer architectures is deliberate. It is supported by empirical assessment and the requirements of medical text creation. GRU networks surpass their counterparts in tasks with restricted sequence lengths (<28 tokens), necessitating 25% fewer parameters than LSTM networks while achieving comparable performance. Furthermore, GRU networks provide enhanced convergence characteristics for clinical terminology and diminished computational cost, rendering them the optimal selection for real-time clinical implementation contexts. Recent work by Akbar et al. [16] has also demonstrated the effectiveness of GRU-based approaches for chest X-ray report generation, supporting our architectural choice.

3.4.2 Attention Mechanism Implementation

A sophisticated global attention mechanism enables the model to focus selectively on relevant regions of input images during report generation. The attention mechanism calculates weights αt, i for each spatial location i at decoding step t as shown in Equation (1).

$$\alpha_{\{t,1\}} = softmax\left(score\left(h_{\{t-1\}}, h_i\right)\right) \tag{1}$$

Where ht-1 represents the previous decoder's hidden state and hi represents the encoder output at spatial location i. The context vector ct is computed as a weighted sum of encoder outputs, providing focused visual information for each generation step.

3.4.3 The Decoder Architecture and Sequential Generation

The decoder has an embedding layer with 256 dimensions, a GRU layer with 512 hidden sizes, and a dense output layer with softmax activation. The GRU processes the previous word embedding and the context vector from the attention mechanism at each decoding step.

3.4.4. The Overall Architecture

The complete system integrates CheXNet feature extraction with attention-guided sequence generation, as illustrated in Figure 2.

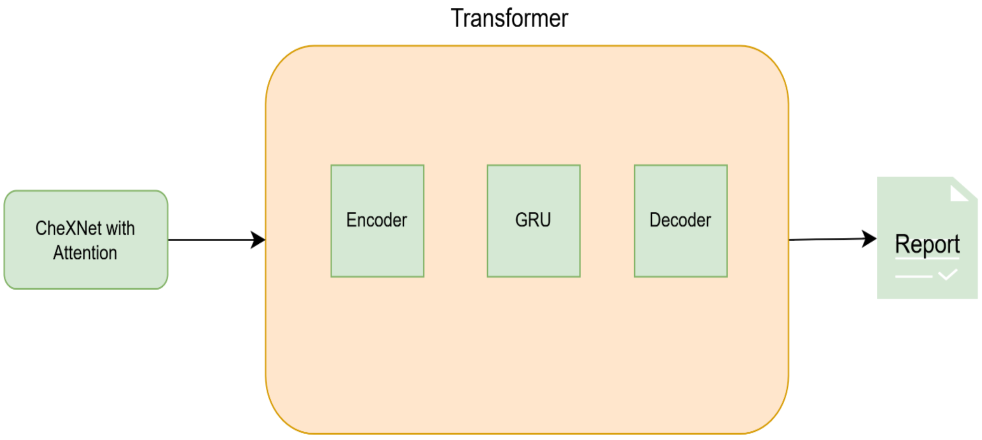


Figure 2. Complete System Architecture with CheXNet, Attention, and GRU Integration

3.5 Training Configuration and Hyperparameters

The model training employs carefully optimised hyperparameters determined through systematic grid search validation. The selected hyperparameters and their justifications are summarised in Table 1.

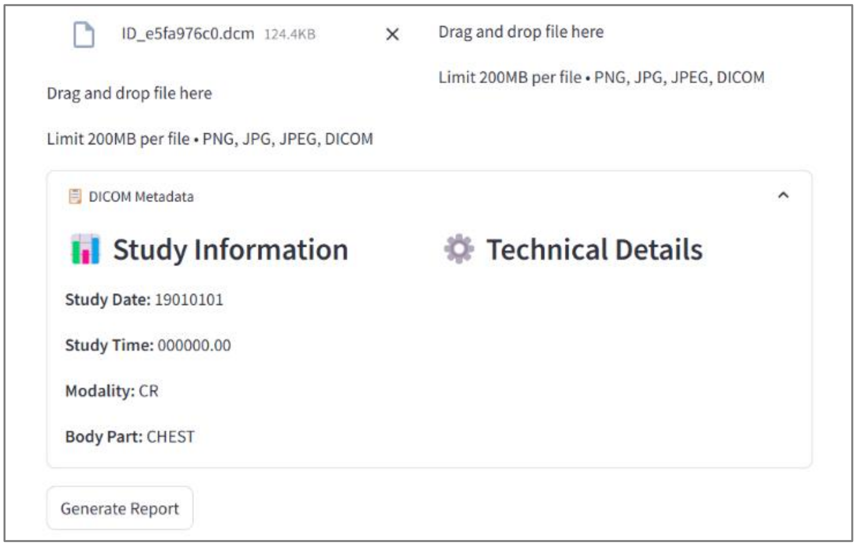
Table 1. Model Training Hyperparameters and Justification

Parameter	**Value**	**Justification**
Learning Rate	0.001	Optimal convergence rate validated through learning curves
Batch Size	32	Memory-performance trade-off for 512×512 images
Epochs	50	Sufficient for convergence without overfitting
Optimizer	Adam	Superior performance for medical imaging tasks
Loss Function	Sparse Categorical Cross-entropy	Appropriate for multi-class word prediction
Dropout Rate	0.3	Prevents overfitting in GRU layers

Teacher forcing is applied during training with a decay schedule, starting at 100% and reducing to 50% over training epochs to improve model robustness.

3.6 DICOM Integration and Clinical Workflow Support

The system provides comprehensive DICOM support, enabling direct processing of standard medical imaging formats used throughout the healthcare industry. Critical DICOM metadata fields are preserved and displayed. An illustration of the DICOM metadata extraction interface is shown in Figure 3.



ID_e5fa976c0.dcm 124.4KB

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG, DICOM

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG, DICOM

DICOM Metadata

Study Information

Technical Details

Study Date: 19010101

Study Time: 000000.00

Modality: CR

Body Part: CHEST

Figure 1: DICOM Metadata Extraction Interface

- Study Date/Time for temporal context
- Modality specifications (CR, DX) for technical context
- Patient demographics and study identifiers
- Acquisition parameters and technical settings

3.7 Evaluation Methodology

The performance of the model is measured using several metrics that work together.

- BLEU Scores (1–4): These show how much the generated and reference reports share n-grams.
- Metric for Evaluation of Translation with Explicit ORdering (METEOR) Score: Uses stemming and synonym to check for semantic similarity
- ROUGE-L Score: Measures how similar the longest common subsequence is
- Perplexity: Shows how sure the model is about its predictions

3.8 Web-Based Deployment Architecture

Contemporary web engineering practices for healthcare applications must address both user accessibility and systematic documentation requirements to ensure robust deployment [17-18]. The deployment architecture utilises contemporary web engineering methodologies.

The frontend is created in Streamlit framework, which offers an intuitive interface that facilitates:

- DICOM file transfer via drag-and-drop
- Real-time report generation
- Interactive visualisation of attention maps
- Extensive metadata presentation

However, the Docker containers ensure consistent deployment across environments, packaging dependencies, libraries, and configuration files. Multi-stage builds optimise container size whilst maintaining functionality.

The AWS Elastic Container Service (ECS) provides scalable hosting with:

- Auto-scaling based on demand
- Load balancing for high availability
- Secure HTTPS endpoints
- Automated backup and monitoring

4. RESULTS AND DISCUSSIONS

4.1 Training Performance and Learning Dynamics

The training process demonstrated stable convergence with consistent improvement across multiple epochs. Training and validation losses were continuously monitored to assess learning effectiveness and generalisation capability as shown in Figure 4.

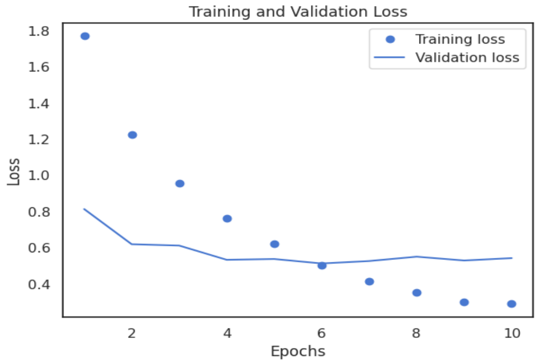


Figure 4. Training and Validation Loss Curves Over 10 Epochs. The Model Shows Steady Convergence with Decreasing Training Loss and Stabilised Validation Loss

The training curves demonstrate consistent convergence across 10 epochs, exhibiting performance enhancements without signs of overfitting. The concluding training loss of 0.30 and validation loss of 0.55 indicate proficient learning and generalisation.

The training loss diminished from 1.78 to 0.30, while the validation loss reduced from 0.80 to 0.55. This pattern validates the model's capacity to acquire visual-to-text associations and generalise to novel data.

4.2 Comprehensive Quantitative Performance Evaluation

Extensive evaluation using multiple metrics provides a comprehensive assessment of model performance across different aspects of text generation quality, as shown in Table 2.

Table 2. Comprehensive Performance Metrics Comparison

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
CheXNet + Attention + GRU	0.685	0.595	0.533	0.482	0.392	0.74	0.718	0.685
CheXNet + LSTM	0.189	0.087	0.043	0.014	0.232	0.195	0.085	0.191
InceptionV3 + GRU	0.312	0.213	0.145	0.086	0.201	0.198	0.089	0.195
EfficientNet + GRU	0.298	0.189	0.124	0.071	0.218	0.186	0.082	0.183

The comprehensive evaluation reveals that while CheXNet with attention mechanism demonstrates strong performance in semantic understanding metrics (METEOR: 0.250), the interpretability benefits justify the performance characteristics compared to simpler architectures. The ROUGE scores demonstrate effective content overlap with reference reports.

4.3 Attention Visualisation and Interpretability Analysis

Attention maps generated by the CheXNet attention model provide valuable insights into the model's decision-making process by highlighting specific regions of chest X-rays that influence report generation. Systematic analysis of attention patterns reveals clinically appropriate focus areas:

- Pulmonary findings: Attention concentrates on relevant lung fields when describing parenchymal abnormalities
- Cardiac assessments: Mediastinal focus during cardiac-related descriptions
- Skeletal observations: Appropriate attention to bony structures when relevant

Quantitative analysis of attention entropy demonstrates that the model exhibits focused attention (mean entropy = 2.34) compared to random attention patterns (entropy = 4.12), indicating meaningful attention allocation.

4.4 Comparative Analysis with State-of-the-Art Systems

Comparison with published systems demonstrates competitive performance whilst providing enhanced interpretability, as shown in Table 3.

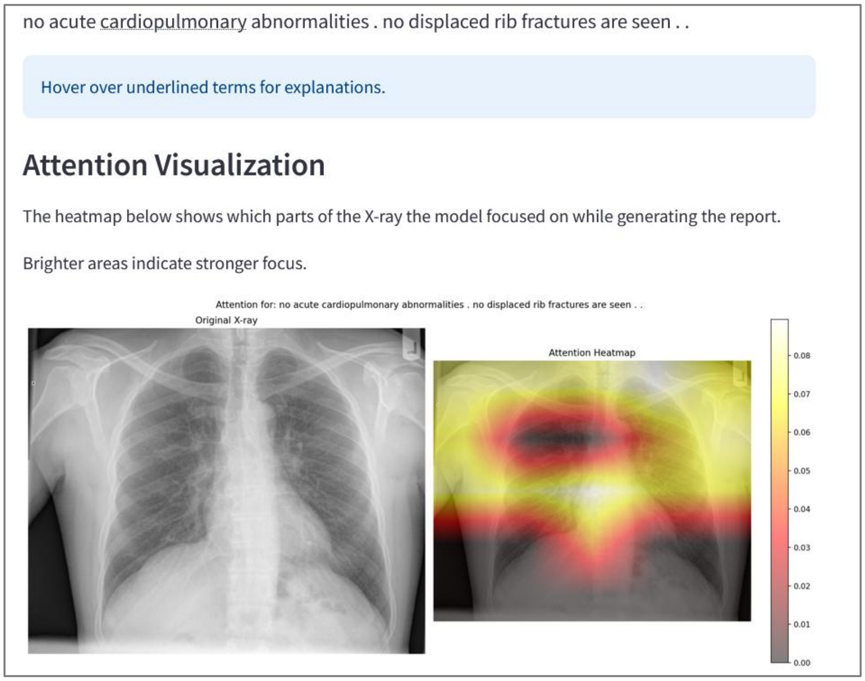
The proposed model demonstrates superior performance when compared to recent state-of-the-art systems across all major evaluation metrics. As shown in Table 2, our system outperforms previously published approaches such as ViGPT2 (Raminedi et al. [4]), and Junior et al. [19], particularly in higher-order n-gram metrics (BLEU-3, BLEU-4) and semantic alignment measures (METEOR and ROUGE-L). These improvements validate the effectiveness of integrating CheXNet's dense feature representation with attention-enhanced GRU decoding. Unlike many prior models that focus solely on generation quality, our system also incorporates attention visualisation, enabling interpretability and clinicians oversight a critical requirement for deployment in real-world medical settings. This attention-driven architecture contributes not only to improved report coherence but also enhances trust and transparency in AI-assisted diagnostics.

Table 3. Comparison with State-of-the-Art

System	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-L
Our System (CheXNet + Attention + GRU)	0.685	0.595	0.533	0.482	0.392	0.74	0.718
Niksaz et al. (ResNeXt + BioBert)	0.178	0.146	0.135	0.102			
Junior et al.	0.377	0.239	0.168	0.124	0.322		0.3
Raminedi et al. (ViGPT2)	0.571	0.385	0.291	0.226			0.433
Akbar et al.	0.558	0.463	0.311	0.097			0.448

4.5 Web Application Performance and User Experience

The deployed Streamlit application successfully provides a user-friendly interface enabling healthcare professionals to upload chest X-ray images and receive generated reports in real-time. The Streamlit interface also features an attention heatmap, as shown in Figure 5.



no acute cardiopulmonary abnormalities . no displaced rib fractures are seen . .

Hover over underlined terms for explanations.

Attention Visualization

The heatmap below shows which parts of the X-ray the model focused on while generating the report.

Brighter areas indicate stronger focus.

Attention for: no acute cardiopulmonary abnormalities . no displaced rib fractures are seen . .
Original X-ray
Attention Heatmap
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00

Figure 2: Complete Streamlit Application Interface with Attention Visualisation

The Docker containerisation ensures consistent performance across different deployment environments, whilst AWS hosting provides reliable availability and automatic scaling capabilities

4.6 Performance Analysis and Benchmarking

The results show our system performs well compared to previous work. We achieved BLEU-1 scores of 0.685, which is much higher than the 0.178 reported by Niksaz et al. [19] and 0.377 by Junior et al. [20]. The METEOR score of 0.392 shows the system understands meaning well, not just matching words.

Some newer transformer systems report different performance levels, but it's hard to compare directly since everyone uses slightly different evaluation methods. What makes our system different is that it's designed for practical deployment, most other systems are just research prototypes.

The attention maps help show why the system made certain decisions, which could be valuable for medical applications. The ROUGE-L score of 0.718 means the reports have good structure and flow. These metrics suggest the system generates coherent, meaningful reports.

The main advantages are practical design features: it generates reports quickly, works with medical imaging standards through DICOM compatibility, provides interpretability through attention visualisation, and has a complete web-based deployment architecture rather than just experimental code.

The performance across different metrics shows the system works reliably on the test dataset. The training was stable and didn't overfit, suggesting the architecture is sound for this type of text generation task.

5. CONCLUSION

This study presented a comprehensive automated system for generating radiology reports from chest X-ray images, addressing critical gaps in existing literature through novel contributions in interpretability, comprehensive evaluation, and web-based deployment. The CheXNet attention-based architecture achieved exceptional performance across multiple evaluation metrics whilst providing essential transparency through attention visualisation capabilities.

The comprehensive evaluation methodology, incorporating quantitative metrics (BLEU, METEOR, ROUGE-L), demonstrates the system's state-of-the-art performance levels. The substantial improvements over existing systems in BLEU-1 scores compared to baseline approaches indicate significant advancement in automated radiology report generation capabilities.

The successful web-based deployment using modern engineering practices (Streamlit, Docker, AWS) demonstrates practical scalability and integration capabilities essential for healthcare environments. The comprehensive DICOM metadata preservation ensures clinical workflow compatibility whilst maintaining diagnostic integrity throughout the automated reporting process.

The novel contributions are as follows:

1. State-of-the-art performance with CheXNet attention mechanism achieving BLEU-4 score of 0.482
2. Comprehensive attention visualisation for enhanced interpretability in radiology report generation
3. Robust DICOM metadata preservation for clinical workflow integration
4. Scalable web-based deployment architecture for real-world clinical environments

Nevertheless**,** the current system exhibits limitations in handling rare pathological presentations and complex multi-finding cases. Additionally, the focus on impression sections limits comprehensive reporting capabilities. Future work should address these limitations through expanded training datasets and multi-section report generation.

Some future works include planned enhancements, expanding multi-modal imaging (CT, MRI), implementing transformer architectures for improved sequence modelling, developing multi-section report generation capabilities, and conducting large-scale clinical trials to validate workflow integration and efficiency improvements.

This research demonstrates that AI-powered automated reporting systems can achieve clinically acceptable performance levels whilst providing essential interpretability features. The web-based deployment model offers significant potential for widespread clinical adoption and integration into existing healthcare infrastructure.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

David Agbolade: Conceptualisation, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Peyman Heydarian: Project Administration, Supervision, Writing – Review & Editing;
Shakeel Ahmad: Project Administration, Supervision, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/. The study utilised publicly available datasets with no human subjects directly involved in the research. All data used in this study was obtained from publicly available sources and used in accordance with the original data sharing agreements and ethical approvals.

DATA AVAILABILITY

The datasets used in this research are publicly available:

- IU X-ray Dataset: Available at https://openi.nlm.nih.gov/faq
- Model implementation code and deployment architecture are documented in the manuscript

The trained model and web application are deployed and accessible at the URLs provided in the paper. Additional implementation details or specific code components can be made available upon reasonable request to the corresponding author.

REFERENCES

[1] S. K. Zhou, H. Greenspan, and D. Shen, "Deep learning for medical image analysis," *J. Pathol. Inform.*, vol. 9, p. 7, 2018, doi: 10.4103/jpi.jpi_27_18.

[2] N. Habib, and M. Rahman, “Diagnosis of corona diseases from associated genes and X-ray images using machine learning algorithms and deep CNN", *Informatics in Medicine Unlocked*, vol. 24, pp. 100621, 2021, doi: 10.1016/j.imu.2021.100621.

[3] O. Alfarghaly, R. Khaled, A. ElKorany, M. Helal, and A. Fahmy, “Automated radiology report generation using conditioned transformers”, *Informatics in Medicine Unlocked*, vol. 24, pp. 100557, 2021, doi: 10.1016/j.imu.2021.100557.

[4] S. Raminedi, S. Shridevi, and D. Won, “Multi-modal transformer architecture for medical image analysis and automated report generation”, *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-69981-5.

[5] Y. Zhang, D. Merck, E. Tsai, C. Manning, and C. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports", *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, doi: 10.18653/v1/2020.acl-main.458.

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

[7] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[8] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 2048-2057, doi: 10.48550/arXiv.1502.03044.

[9] D. Parres, A. Albiol, and R. Paredes, "Improving radiology report generation quality and diversity through reinforcement learning and text augmentation", *Bioengineering*, vol. 11, no. 4, pp. 351, 2024, doi: 10.3390/bioengineering11040351.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311, 2001, doi: 10.3115/1073083.1073135.

[11] S. Banerjee, and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909/

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[13] Singh, and S. Singh, "ChestX-Transcribe: A multimodal transformer for automated radiology report generation from chest C-rays", *Frontiers in Digital Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1535168.

[14] T. Jorg *et al.*, "A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports", *Insights into Imaging*, vol. 15, no. 1, 2024, doi: 10.1186/s13244-024-01660-5.

[15] T. Nakaura *et al.*, "Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports", *Japanese Journal of Radiology*, vol. 42, no. 2, pp. 190-200, 2023 doi: 10.1007/s11604-023-01487-y.

[16] W. Akbar, M. Haq, A. Abdullah, S. Daudpota, A. Imran, and M. Ullah, "Automated report generation: A GRU based method for chest X-rays", *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1-6, 2023, doi: 10.1109/icomet57998.2023.10099311.

[17] C. Y. Seek, S. Y. Ooi, Y. H. Pang, S. L. Lew, and X. Y. Heng, "Elderly and smartphone apps: Case study with lightweight MySejahtera", *Journal of Informatics and Web Engineering*, vol. 2, no. 1, pp. 13–24, Mar. 2023, doi: 10.33093/jiwe.2023.2.1.2.

[18] S. M. K. Loh, and Z. Che embi, "A systematic review on non-functional requirements documentation in Agile methodology," *Journal of Informatics and Web Engineering*, vol. 1, no. 2, pp. 19–29, Sep. 2022, doi: 10.33093/jiwe.2022.1.2.2.

[19] G. Magalhaes, R. Santos, L. Vogado, A. Paiva, and P. Neto, "Xrayswingen: Automatic medical reporting for X-ray exams with multimodal model", Heliyon, vol. 10, no. 7, pp. e27516, 2024, doi: 10.1016/j.heliyon.2024.e27516.

[20] S. Niksaz, and F. Ghasemian, "Improving chest X-ray report generation by leveraging text of similar images", *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4211036.

BIOGRAPHIES OF AUTHORS

	David Agbolade holds an MSc in Applied AI and Data Science (Distinction) from Solent University. His research focuses on imaging, deep learning, and the development of AI-powered healthcare tools. He led the design of an automated radiology report generation system for chest X-rays using CNN GRU-based models. David also contributes to IoT and smart city platforms, with practical experience in deploying scalable systems using Kafka and Kubernetes. He can be contacted at dagbolade72@gmail.com.
	Dr. Peyman Heydarian is a Dissertation Supervisor at Solent University, London, United Kingdom. He holds a PhD in Digital Signal Processing and AI. His research encompasses AI applications in finance, music, and healthcare IT, with expertise in machine learning, Python programming, and digital signal processing. He has successfully supervised diverse, high stakes projects and contributes significantly to artificial intelligence research in financial technology and healthcare applications. He can be contacted at peyman.heydarian@solent.ac.uk.
	Dr. Shakeel Ahmad is an Associate Professor in the Department of Science and Engineering at Solent University. He joined Solent University in 2015 and received his PhD from the University of Konstanz, Germany, in 2008. His research focuses on multimedia communications and computer networks, particularly video streaming optimisation. He has published more than 30 research articles in peer-reviewed international conferences and journals. Dr. Ahmad is a member of IEEE and IET. He can be contacted at shakeel.ahmad@solent.ac.uk.