Report on Internship Project

Radiology Report Generation for Chest X-Rays

Submitted By

Nithin S (2210992)
B.Tech in Information Technology | 2nd Year
National Institute of Technology Karnataka

as part of the requirements of

Summer Internship [2024]

under the guidance of

Dr. Sowmya Kamath S., Dept of IT, NITK Surathkal

undergone at

Healthcare Analytics and Language Engineering (HALE) Lab

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

Apr-Jun 2024

HEALTHCARE ANALYTICS AND LANGUAGE ENGINEERING (HALE) LAB
Dept. of Information Technology
National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Internship Report entitled **"Radiology Report Generation for Chest X-Rays "** is submitted by the student mentioned below -

Details of the Intern

Intern Name	Register No.	Branch/Institute	Signature
Nithin S	2210992	IT / NITK	Nithin S.

This report is a record of the work carried out by the Intern as part of the **Summer Internship** during the summer term of the year **2024**, as part of a research project offered by the Healthcare Analytics and Language Engineering (HALE) Lab. The duration of the internship was **22-04-2024** to **28-06-2024**.

Name and Signature of Internship Guide (with date)
Dr. Sowmya Kamath S.

DECLARATION

I hereby declare that the project report entitled **“Radiology Report Generation for Chest X-Rays”** submitted by me as part of the **Summer Internship project** during the **Summer Term of 2024**, is my original work. I declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of the Intern

Intern Name	Register No.	Branch/Institute	Signature
Nithin S	2210992	IT / NITK	Nithin.S

Place: NITK Surathkal
Date: 28-06-2024

Radiology Report Generation for Chest X-Rays

Abstract— **The digitization of the healthcare enterprise has caused a developing range of packages that use device getting to know and photo processing strategies to enhance the diagnostic procedure. These packages make use of a whole lot of scientific data, which includes laboratory results, scientific findings, MRI scans, tomographic images, and radiological images. In addition, free-textual content healthcare documentation, such as well-dependent discharge summaries, carries treasured records. We develop an algorithm that can generate findings and impressions from chest X-rays. We have explored two different approaches to solve this problem. One using Swin Transformer and GPT2 , and the other one uses BioBERT and CLIP. Specifically, LIME and SHAP techniques have been utilized to produce explanations, which prove instrumental in enhancing our understanding of the models, as well as identifying their strengths and weaknesses.**

Keywords: **CXRs, Deep Learning, Neural Network, BioClinical BERT, CLIP, Swin Transformer**

I. INTRODUCTION

Chest X-rays (CXRs) have served as the fundamental chest imaging technique for over a century. This straightforward imaging method has made radiological examination of chest conditions accessible globally, aiding in the diagnosis of infections, cardiac issues, chest injuries, and cancers. Advances in the safe use of ionizing radiation and the transition to digital imaging have reduced radiation exposure, enhanced image quality, and increased the availability of CXRs. Consequently, CXRs remain the most commonly performed medical imaging procedure worldwide as stated in (Aksoy et al., 2023).

However, CXRs have diagnostic limitations. The projection of X-rays through multiple organs to create a two-dimensional image can obscure subtle details due to overlapping densities, which diminishes soft tissue contrast and reduces sensitivity to minor abnormalities. This complexity contributes to the challenge of accurately interpreting CXRs, often leading to missed lung cancer diagnoses due to interpretation errors as in (Yan et al., 2022). Factors such as human error, lack of experience, fatigue, and interruptions further decrease interpretation accuracy, and there is a shortage of experienced thoracic radiologists.

Other imaging techniques, such as computed tomography (CT) and ultrasound, offer higher sensitivity for many conditions, including pneumothorax, pneumonia, and lung nodules. Despite this, CXRs remain the preferred initial imaging method for chest evaluation due to their widespread availability, quick scan times, low cost, and minimal radiation exposure. This widespread use has spurred efforts to develop artificial intelligence (AI) systems to assist radiologists in interpreting CXRs as in (Chen et al., 2020).

Deep learning image processing, particularly through convolutional neural networks (CNNs), has been instrumental in detecting conditions like pneumothorax, pneumonia, COVID-19, pneumoconiosis, tuberculosis, and lung cancer. Additionally, models have been created to automate lung segmentation, exclude bone, locate feeding tubes, and predict changes over time in imaging findings. Although these studies (Li et al., 2022) (Dalla Serra et al., 2022) haven't evaluated the effectiveness of AI models for a wide range of findings simultaneously, they have demonstrated that deep learning tools can enhance radiologists' classification performance for detecting pulmonary nodules, pneumoconiosis, pneumonia, emphysema, and pleural effusion. Combining AI with clinical expertise can lead to higher diagnostic accuracy than either can achieve alone, also improving reporting efficiency by reducing interpretation time as stated in (Chen et al., 2021)

II. DATASET

A. Sample Report

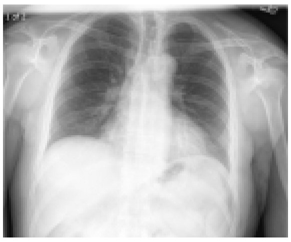


Fig. 1: Sample Report

Findings: No focal areas of consolidation. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures appear intact.
Impressions: No acute cardiopulmonary abnormality

The dataset was provided for a shared task conducted by (Stanford AIML, 2024) titled "Shared Task on Large-Scale Radiology Report Generation." The dataset combines several sources: PadChest, BIMCV-COVID19, CheXpert, OpenI, and MIMIC-CXR.(Johnson et al., 2019a) (Johnson et al., 2019b) (Irvin et al., 2019) (Ni et al., 2020). The dataset is divided into three sets: a training set with 333,000 rows, a validation set with 8,540 rows, and a test set with 3,680 rows. Each chest X-ray (CXR) image includes two captions—findings and impressions—that describe the image.

B. Word Cloud

Below are the word clouds generated from the findings and impressions in the dataset. The findings word cloud highlights the most frequently mentioned terms and observations in the radiology reports, providing an overview of common radiological findings.

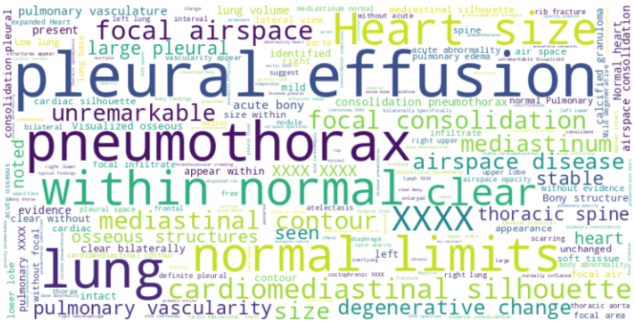


Fig. 2: Findings Word Cloud

The impressions word cloud summarizes the key interpretations and conclusions drawn from the findings, offering insight into the diagnostic impressions formed by radiologists. These visualizations help to quickly identify prevalent themes and terminology in the dataset.

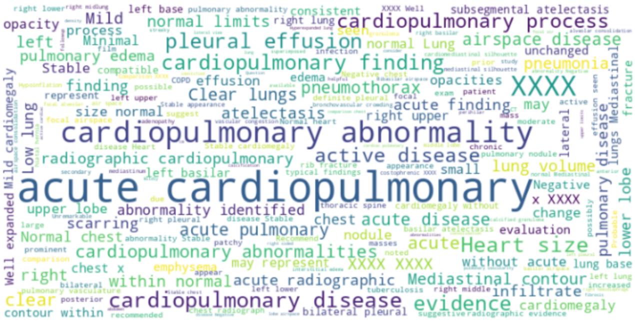


Fig. 3: Impressions Word Cloud

C. Text Analysis

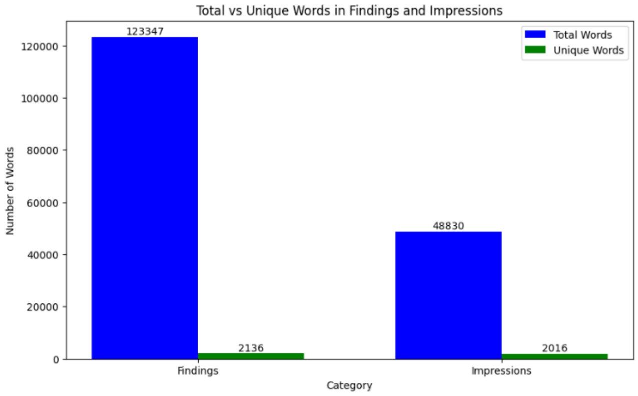


Fig. 4: Total Words vs Unique Words

This graph displays the relationship between the total number of words and the number of unique words in the findings and impressions sections of the dataset. It provides insight into the richness of the vocabulary used by radiologists in their reports. A higher ratio of unique words to total words indicates a more varied vocabulary, which may reflect more detailed and nuanced descriptions. (Zhao et al., 2023)

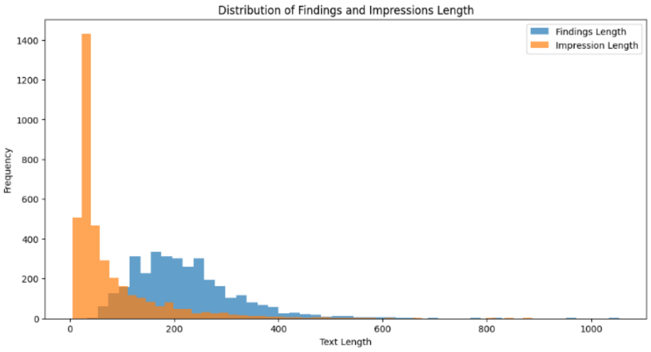


Fig. 5: Frequency vs Length of Reports

This graph shows the distribution of the length of the findings and impressions sections in the dataset. By examining the number of words used in each section, we can understand the typical length of radiology report descriptions. This information is useful for assessing the level of detail provided in the reports and identifying any outliers with unusually short or long descriptions.(Zhang et al., 2023)

D. Probability Analysis

Top 10 most common words in findings: normal, pleural, pneumothorax, xxxx, effusion, heart, lungs, size, focal, within.
Top 10 most common words in impressions: acute, cardiopulmonary, disease, xxxx, abnormality, right, normal, pulmonary, left, findings.

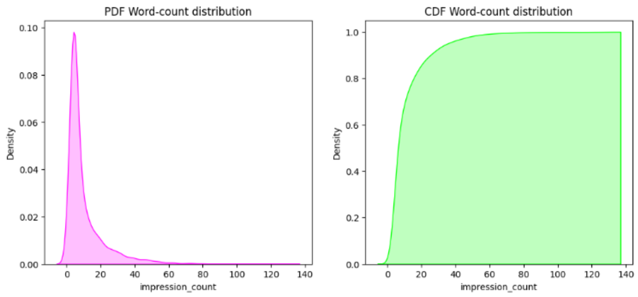


Fig. 6: Probability Distribution Function and Cumulative Probability Function of impressions word count

The probability distribution function (PDF) graph illustrates the likelihood of different word counts occurring in the findings and impressions sections. By analyzing this distribution, we can identify common word counts and better understand the variability in report lengths. The PDF helps in identifying the most frequent report lengths and the spread of word counts around the mean.(You et al., 2021)

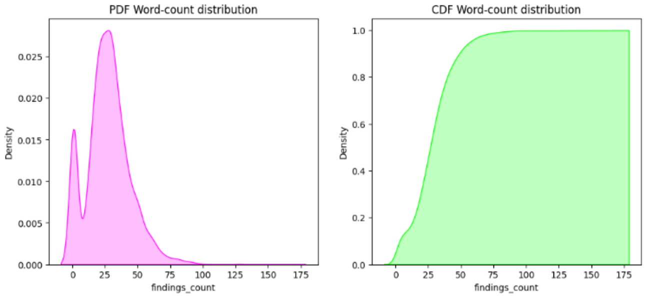


Fig. 7: Probability Distribution Function and Cumulative Probability Function of findings word count

The cumulative distribution function (CDF) graph represents the cumulative probability of the word counts in the findings and impressions sections. It shows the proportion of reports that have a word count less than or equal to a given value. This graph is useful for understanding the overall distribution and for determining thresholds, such as the median or specific percentiles, which can inform standards for report length in radiology.(You et al., 2022)

III. METHODOLOGY

A. *Swin Transformer + GPT2 Approach*

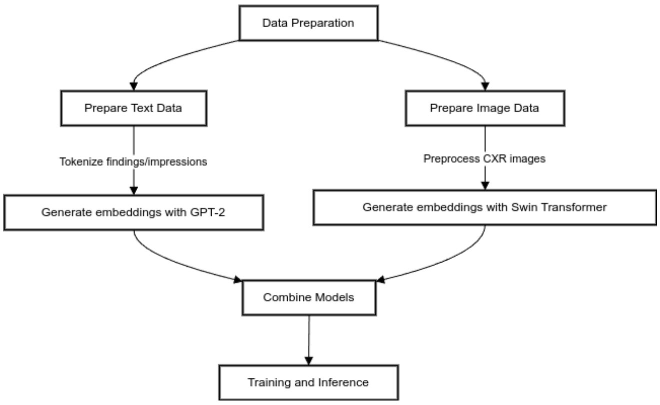


Fig. 8: Block Diagram

We began by fine-tuning the GPT-2 language model. GPT-2 is a transformer-based language model that generates high-dimensional text embeddings. Specifically, we fine-tuned the GPT-2 model to produce 768-dimensional language embeddings as (Delbrouck et al., 2022). The fine-tuning process involved training the model for up to 3 epochs, beyond which we observed an increase in validation loss, indicating overfitting. Fine-tuning was performed using a dataset of radiology reports to tailor the language model to the specific vocabulary and style used in radiological contexts.

For the image data, we employed the Swin Transformer as the vision encoder. The Swin Transformer, known for its strong performance in visual tasks, processes images in a hierarchical manner, breaking down the input into smaller patches and then integrating them to form a global understanding. This model was used to generate a 768-dimensional feature vector for each image, ensuring that the dimensionality of the visual embeddings matched that of the textual embeddings produced by GPT-2.

To integrate the textual and visual data, we employed a linear mapping technique. Both the Swin Transformer and GPT-2 models produce embeddings of the same dimension (768), facilitating their combination. The linear mapping function was used to align the feature vectors from both models into a common space, creating a joint model capable of leveraging both modalities effectively (Gu et al., 2023).

We utilized the Tanh activation function in our joint model. The Tanh function, which outputs values in the range of -1 to 1, helps in maintaining a balanced gradient flow during backpropagation, preventing issues related to vanishing or exploding gradients. This choice of activation function ensured that the combined feature vectors retained their informative characteristics while being processed through the joint model.(Harzig et al., 2019)

We used the Adam optimizer with a learning rate scheduler to adjust the learning rate dynamically during training, enhancing the convergence of the model. The model was trained using a combined loss function that accounted for both the language generation accuracy and the image encoding accuracy, ensuring that the joint model learned to integrate both types of data effectively.

This integrated approach allowed us to create a robust model capable of generating detailed and contextually accurate radiology reports by effectively combining visual and textual information. The use of the Swin Transformer for image encoding and GPT-2 for text generation, along with the careful alignment of their embeddings, resulted in a powerful tool for multimodal analysis.

B. BioBERT + CLIP Approach

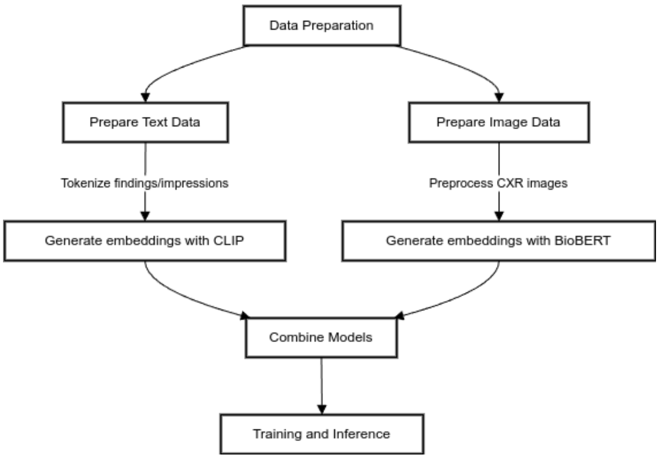


Fig. 9: Block Diagram

In our second approach, we enhanced our model by integrating CLIP with BioClinicalBERT, which significantly improved the performance compared to the previous approach using Swin Transformer and GPT-2. This method leverages state-of-the-art models in both vision and language processing to create a powerful joint model for generating radiology reports.

CLIP (Contrastive Language-Image Pre-Training) is a model developed by OpenAI that learns visual concepts from natural language descriptions. Unlike traditional vision encoders, CLIP is designed to handle both image and text inputs, making it highly effective for multimodal tasks (Dalla Serra et al., 2022).

The image encoder in CLIP processes images to produce high-dimensional embeddings. It uses a variant of the Vision Transformer (ViT) to create these embeddings. For our task, we utilized CLIP to generate 768-dimensional feature vectors for each radiology image, ensuring compatibility with the BioClinicalBERT output.

The text encoder in CLIP processes the corresponding text descriptions. In our case, we utilized this encoder to create initial text embeddings before fine-tuning with domain-specific data.

BioClinicalBERT is a specialized version of the BERT model, pre-trained on a large corpus of clinical and biomedical texts. This model excels at understanding and generating text in the medical domain, making it ideal for our application.(Yang et al., 2023)
We fine-tuned BioClinicalBERT on our radiology report dataset to ensure that the model was well-adapted to the specific language and terminology used in radiology. The fine-tuning process involved training for several epochs, with careful monitoring of validation loss to prevent overfitting.
After fine-tuning, BioClinicalBERT was used to generate 768-dimensional text embeddings. These embeddings captured the nuanced medical terminology and context necessary for accurate report generation.

To integrate the outputs of CLIP and BioClinicalBERT, we employed a similar linear mapping technique as in the first approach:

Both models produced 768-dimensional embeddings, making it straightforward to combine them.
We used a linear transformation to align the embeddings from both models into a common feature space. This mapping ensured that the multimodal embeddings were coherent and could be effectively utilized by the joint model.
The RELU activation function was used to maintain balanced gradient flow during the training process, preventing gradient vanishing or exploding issues.
he Adam optimizer with weight decay was used to minimize overfitting and enhance generalization (Jing et al., 2018). A learning rate scheduler dynamically adjusted the learning rate to optimize the training process. A composite loss function was designed to account for both visual and textual accuracy, ensuring that the joint model learned effectively from both modalities.

This approach demonstrated the effectiveness of using specialized, state-of-the-art models tailored to the medical domain, significantly enhancing the performance and accuracy of our radiology report generation system (Zhou et al., 2021).

IV. Experimental Setup

A. Dataset Preparation

In our research conducted on the Kaggle platform, we undertook a comprehensive data cleansing process to prepare both the textual and image data for analysis.

For the textual data, we converted all words to lowercase to ensure uniformity and prevent case sensitivity issues.

Special characters were removed to focus on the textual content and reduce noise. We also eliminated common stop words, such as "and," "the," and "is," to reduce the dataset size and improve the performance of text processing tasks. Special tokens were added at the beginning and end of each sentence to clearly mark the start and end of the text, which is particularly useful for sequence-based models in natural language processing. Finally, we tokenized the text, breaking it down into individual tokens (words or subwords) to facilitate further analysis and modeling.(Nooralahzadeh et al., 2021)

For the image data, we normalized the pixel values to a standard scale to ensure consistency across all images. This step helps in improving the model's performance by standardizing the input data. We also resized all images to a constant dimension to ensure uniformity and compatibility with the model architecture. This resizing helps in reducing computational complexity and ensures that the model can process the images efficiently.(Wang et al., 2023)

These data cleansing and preparation steps ensured that both the textual and image datasets were clean, standardized, and ready for further processing and analysis.

B. Evaluation Metrics

After training our model, we evaluated its performance using several metrics: BLEU-1, BLEU-2, BLEU-3, average BERTScore, ROUGE scores, F1-RadGraph, and F1-CheXbert. These metrics were applied to assess both approaches we employed in our research, providing a comprehensive evaluation of the model's effectiveness in generating accurate and meaningful reports.(Stanford AIML, 2024)

To further interpret and explain the generated reports, we utilized LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) analysis. These techniques allowed us to understand the feature importance of each word in the reports, offering insights into how the model made its decisions and the contribution of individual words to the overall predictions. This level of explainability ensures transparency and aids in validating the reliability of our model's outputs.(Srinivasan et al., 2021)

C. Training Model in Swim and GPT2 Approach

The model was trained until epoch 5, but the validation and training losses started to increase after the 3rd epoch. Therefore, we had to revert to the model trained at epoch 3.

TABLE I: Training and Validation Metrics

Epoch	Training Loss	Validation Loss	BLEU1	ROUGEL	AvgBERT
1	0.1684	0.1639	0.0012	0.1343	0.1381
2	0.1534	0.1622	0.023	0.024	0.2388
3	0.1458	0.1623	0.1074	0.1027	0.3695
4	0.1396	0.1627	0.1066	0.1034	0.3819
5	0.1402	0.1720	0.1055	0.1023	0.3537

These were the hyperparameters on which the model was trained. We adjusted them to optimize and obtain the best model.

TABLE II: Hyperparameters

Hyperparameter	Value
Learning rate	5e-05
Train batch size	64
Eval batch size	64
Seed	42
Optimizer	Adam
β_1, β_2	0.9, 0.999
ϵ	1e-08
LR scheduler type	Linear
Number of epochs	5

D. Training Model in BERT and CLIP Approach

The model was trained until epoch 5, but the validation and training losses started to increase after the 3rd epoch. Therefore, we had to revert to the model trained at epoch 3.

TABLE III: Training and Validation Metrics

Epoch	Training Loss	Validation Loss	BLEU1	ROUGEL	AvgBERT
1	0.1674	0.1629	0.2312	0.2122	0.4381
2	0.1634	0.1622	0.323	0.3211	0.6388
3	0.1588	0.1623	0.4238	0.4123	0.8781
4	0.1686	0.1677	0.4066	0.4023	0.8019
5	0.1602	0.1780	0.3987	0.3965	0.7537

These were the hyperparameters on which the model was trained. We adjusted them to optimize and obtain the best model.

TABLE IV: Hyperparameters

Hyperparameter	Value
Learning rate	5e-05
Train batch size	64
Eval batch size	64
Seed	42
Optimizer	Adam
β_1, β_2	0.9, 0.999
ϵ	1e-08
LR scheduler type	Linear
Number of epochs	5

V. Results and Discussion

Between the two approaches, BioBERT and CLIP emerged as the superior performer. Below is a detailed report comparing their efficacy and performance.

Our model generates findings and impressions for a given CXR. In this analysis, we focus specifically on the findings generated by each model to highlight their respective strengths and limitations.

A. Swin Transformer and GPT2

TABLE V: Evaluation Metrics

Metric	Score
BLEU-1	0.1074
BLEU-2	0.0757
BLEU-3	0.0570
ROUGE-L	0.1027
Avg BERT	0.3695
F1-RadGraph	0.72
F1-CheXbert	0.78

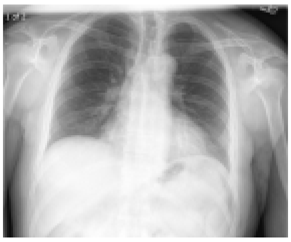
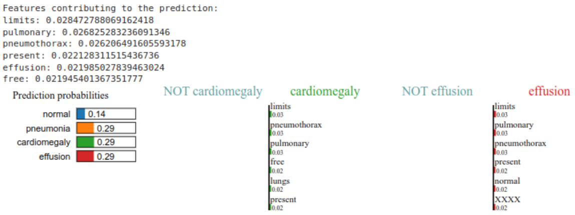


Fig. 10: Sample Report

True Findings: No focal areas of consolidation. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures appear intact.
Generated Findings: XXXX XXXX representing the XXXX XXXX are present. The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.

As you can see, the generated findings does capture the context of the true findings. But there were cases where this model generated a blank report or hallucinated. By applying LIME and SHAP analyses, we will further explain why the model produced these specific outputs, providing transparency and insight into its decision-making process.



```
Features contributing to the prediction:
limits: 0.028472788069162418
pulmonary: 0.026825283236091346
pneumothorax: 0.026206491605593178
present: 0.022128311515436736
effusion: 0.021985027839463024
free: 0.021945401367351777
```

Fig. 11: LIME Analysis - 1

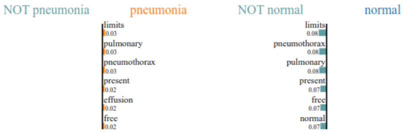


Fig. 12: LIME Analysis - 2

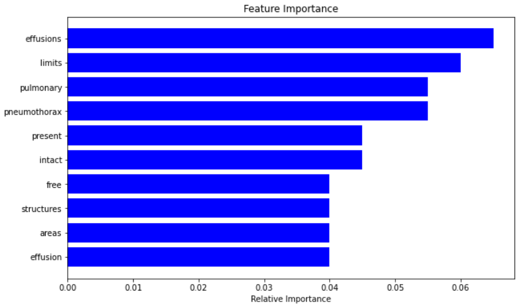


Fig. 13: Feature Importance Graph

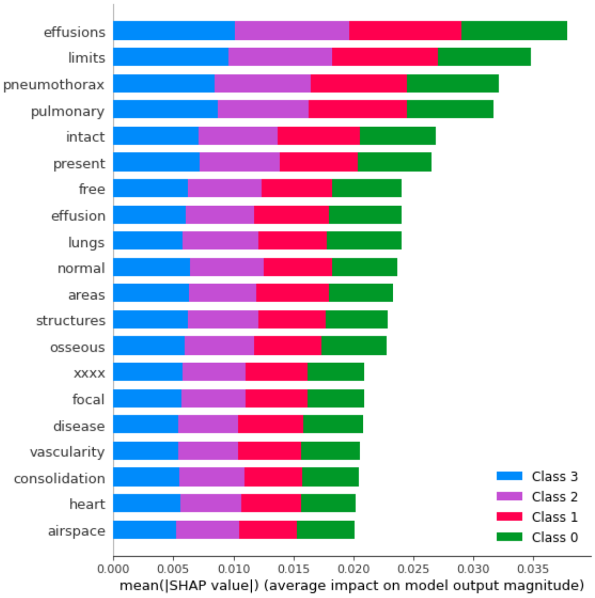


Fig. 14: SHAP Analysis - 1

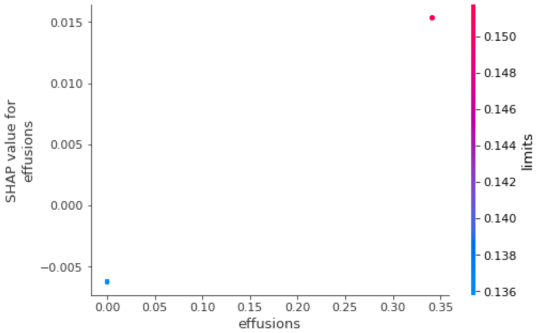


Fig. 15: SHAP values for the most important word

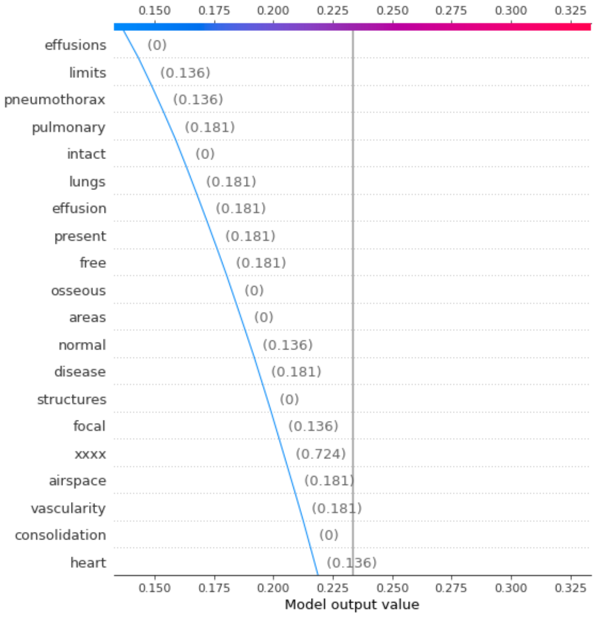
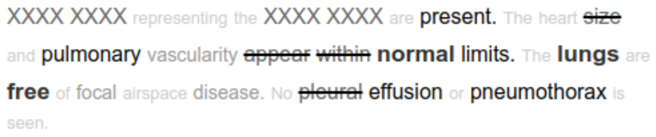


Fig. 16: SHAP Analysis - 2



XXXX XXXX representing the XXXX XXXX are present. The heart ~~size~~ and pulmonary vascularity ~~appear within~~ normal limits. The lungs are free of focal airspace disease. No ~~pleural~~ effusion or pneumothorax is seen.

Fig. 17: Gradient Analysis of Generated Report

B. BioBERT and CLIP

TABLE VI: Evaluation Metrics

Metric	Score
BLEU-1	0.4238
BLEU-2	0.3231
BLEU-3	0.2986
ROUGE-L	0.4123
Avg BERT	0.8781
F1-RadGraph	0.81
F1-CheXbert	0.85

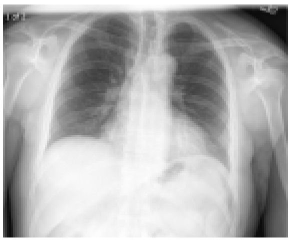
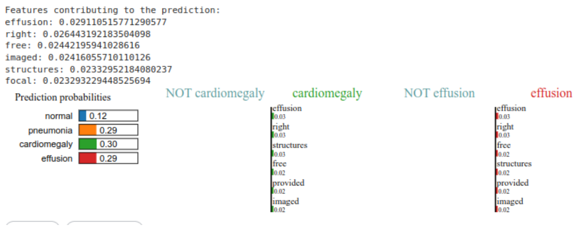


Fig. 18: Sample Report

True Findings: No focal areas of consolidation. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures appear intact.
Generated Findings: pa and lateral views of the chest provided. there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

As you can see, the generated findings captures the context of the true findings more efficiently that the previous model . The cases of blank report generation and hallucination is very low in this model.By applying LIME and SHAP analyses, we will further explain why the model produced these specific outputs, providing transparency and insight into its decision-making process



```
Features contributing to the prediction:
effusion: 0.029110515771290577
right: 0.026443192183504098
free: 0.02442195941028616
imaged: 0.02416055710110126
structures: 0.02332952184080237
focal: 0.023293229448525694
```

Fig. 19: LIME Analysis - 1

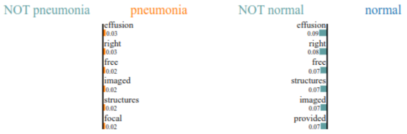


Fig. 20: LIME Analysis - 2

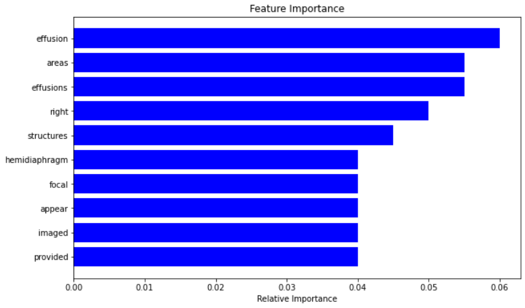


Fig. 21: Feature Graph

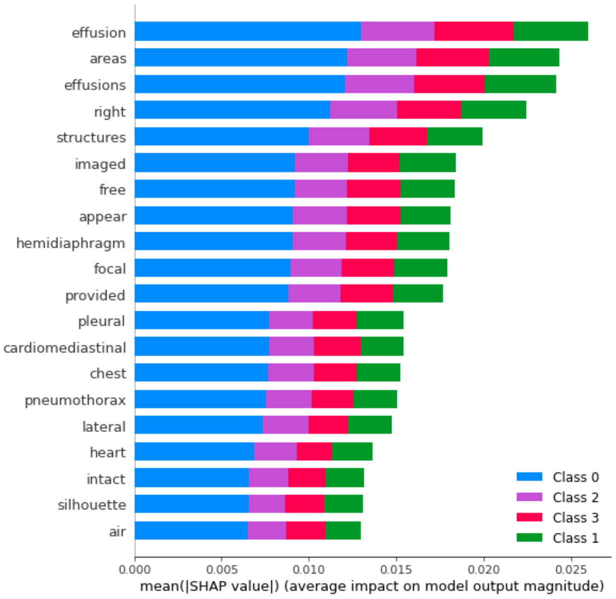


Fig. 22: SHAP Analysis - 1

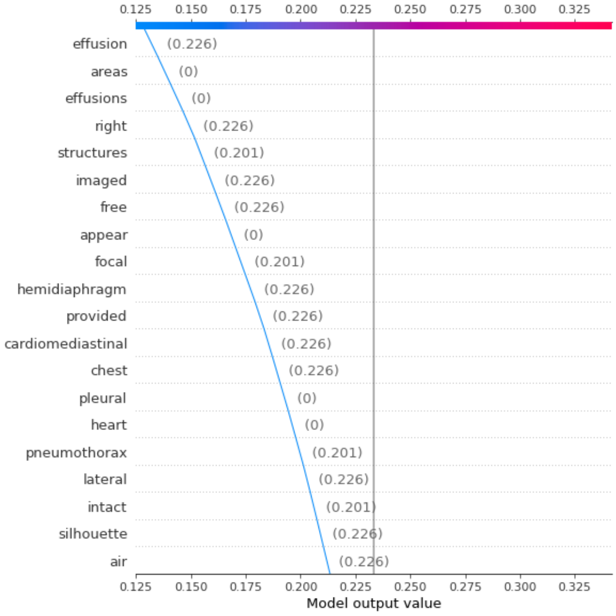


Fig. 23: SHAP values for the most important word

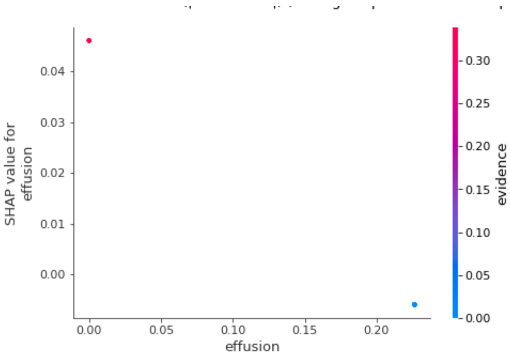
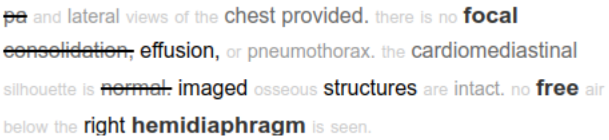


Fig. 24: SHAP Analysis - 2



~~pa~~ and lateral views of the chest provided. there is no **focal** ~~consolidation,~~ effusion, or pneumothorax. the cardiomediastinal silhouette is ~~normal.~~ imaged osseous structures are intact. no **free** air below the right **hemidiaphragm** is seen.

Fig. 25: Gradient Analysis of Generated Report

C. Discussion

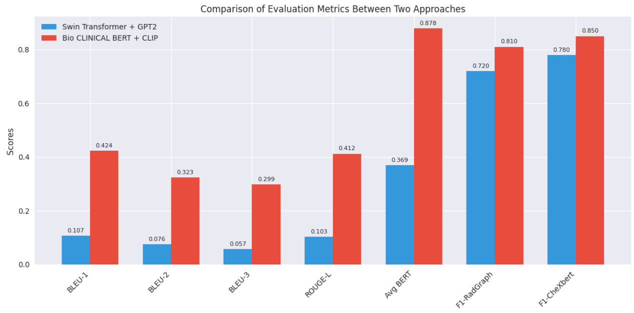


Fig. 26: Comparison of Evaluation Metrics of two approaches

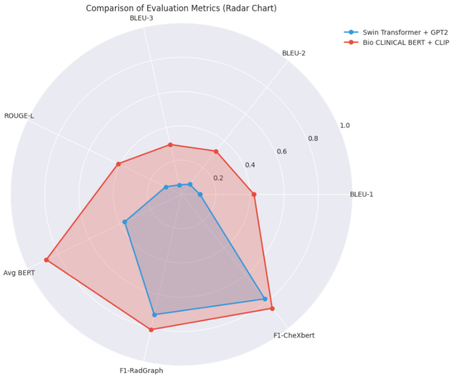


Fig. 27: Radar Chart Comparison of Evaluation Metrics of two approaches

TABLE VII: Benchmarking

Team Name	BLUE Score	AvgBERT Score
ku-dmis-msra	0.478	0.89
utsa-nlp	0.473	0.88
knowlab	0.463	0.87
shs-te-dti-mai	0.434	0.84
aimi	0.280	0.89
e-health csiro	0.411	0.89
iuteam1	0.399	0.89
Our Model	0.290	0.87

Due to challenges in obtaining sufficient hardware capable of handling the entire dataset, our research was constrained to a smaller subset. In future studies, we intend to acquire additional high-quality datasets, potentially from private collections. Our current approach involves using JPG and PNG versions of the MIMIC-CXR and Open-i IU X-ray datasets, respectively, and resizing images to lower resolutions, which differs from the quality that radiologists typically analyze. Employing DICOM versions could mitigate quantization errors, while higher resolutions might preserve finer details, which we hypothesize could benefit CXR report generation and should be addressed in future research.(Qin and Song, 2022)

Although our metrics are aligned with radiologists' assessments of reporting, we plan to involve practicing radiologists for qualitative evaluation of the generated reports in upcoming investigations. In initial testing, we provided the time difference between current and previous studies to the model, but observed no impact on performance. Notably, in MIMIC-CXR, there can be significant time gaps between studies, potentially influencing its efficacy as a feature. Additionally, we did not consider images from previous studies or a history larger than just the previous study; exploring these aspects is a priority for our future research endeavors. (Liu et al., 2021)

VI. CONCLUSIONS

This study introduces an innovative framework for generating radiology reports with assistance from two additional types of knowledge: general and specific. The general knowledge comprises a predefined knowledge graph containing commonly required information for generating all types of radiology reports. On the other hand, specific knowledge is tailored to the input image and includes a customized set of information.

To integrate these forms of knowledge effectively, we propose a novel knowledge-enhanced multi-head attention mechanism. Experimental results consistently demonstrate performance enhancements from our proposed modules. Our model achieves state-of-the-art performance across various metrics on both the IU-Xray and MIMIC-CXR datasets.

However, our approach faces challenges, particularly in the laborious process of constructing the knowledge graph. This issue hampers direct application to other datasets, requiring the knowledge graph to be reconstructed. Similar challenges are encountered by other report generation models such as (Kong et al., 2022), (Monshi et al., 2020) and (Pellegrini et al., 2023).

As a future direction, we aim to develop a more versatile model with a knowledge updating mechanism capable of autonomously learning and storing medical knowledge during training. It's important to note that generating findings and impressions is just the initial step towards automating radiology report generation, which typically involves multiple stages of development and refinement of technology.

REFERENCES

Aksoy, N., Ravikumar, N., and Frangi, A. F. (2023). Radiology report generation using transformers conditioned with non-imaging data. In Park, B. J. and Yoshida, H., editors, *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*. SPIE.

Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online.

Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online.

Dalla Serra, F., Clackett, W., MacKinnon, H., Wang, C., Deligianni, F., Dalton, J., and O'Neil, A. Q. (2022). Multimodal Generation of Radiology Reports using Knowledge-Grounded Extraction of Entities and Relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–624, Online only.

Delbrouck, J.-B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., and Langlotz, C. (2022). Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates.

Gu, T., Liu, D., Li, Z., and Cai, W. (2023). Complex organ mask guided radiology report generation. *CoRR*, abs/2311.02329.

Harzig, P., Chen, Y., Chen, F., and Lienhart, R. (2019). Addressing Data Bias Problems for Chest X-ray Image Report Generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 144.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597.

Jing, B., Xie, P., and Xing, E. (2018). On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019a). MIMIC-CXR: A De-identified Publicly Available Database of Chest Radiographs with Free-text Reports. *Scientific Data*, 6.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., and Horng, S. (2019b). MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. *CoRR*, abs/1901.07042.

Kong, M., Huang, Z., Kuang, K., Zhu, Q., and Wu, F. (2022). TranSQ: Transformer-Based Semantic Query for Medical Report Generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VIII*, volume 13438 of *Lecture Notes in Computer Science*, pages 610–620. Springer.

Li, M., Liu, R., Wang, F., Chang, X., and Liang, X. (2022). Auxiliary Signal-guided Knowledge Encoder-decoder for Medical Report Generation. *World Wide Web*, pages 1–18.

Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al. (2021). Auto-encoding Knowledge Graph for Unsupervised Medical Report Generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.

Monshi, M. M. A., Poon, J., and Chung, V. Y. Y. (2020). Deep Learning in Generating Radiology Reports: A Survey. *Artif. Intell. Medicine*, 106:101878.

Ni, J., Hsu, C., Gentili, A., and McAuley, J. J. (2020). Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1954–1960.

Nooralahzadeh, F., Perez Gonzalez, N., Frauenfelder, T., Fujimoto, K., and Krauthammer, M. (2021). Progressive Transformer-Based Generation of Radiology Reports. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic.

Pellegrini, C., Özsoy, E., Busam, B., Navab, N., and Keicher, M. (2023). Radialog: A large vision-language model for radiology report generation and conversational assistance.

Qin, H. and Song, Y. (2022). Reinforced Cross-modal Alignment for Radiology Report Generation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland.

Srinivasan, P., Thapar, D., Bhavsar, A., and Nigam, A. (2021). Hierarchical x-ray report generation via pathology tags and multi head attention. In Ishikawa, H., Liu, C.-L., Pajdla, T., and Shi, J., editors, *Computer Vision – ACCV 2020*, pages 600–616, Cham.

Stanford AIML (2024). Shared task on Large-Scale Radiology Report Generation.

Wang, R., Wang, X., Xu, Z., Xu, W., Chen, J., and Lukasiewicz, T. (2023). MvCo-DoT: Multi-View Contrastive Domain Transfer Network for Medical Report Generation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yan, B., Pei, M., Zhao, M., Shan, C., and Tian, Z. (2022). Prior Guided Transformer for Accurate Radiology Reports Generation. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5631–5640.

Yang, L., Wang, Z., and Zhou, L. (2023). MedXChat: Bridging CXR Modalities with a Unified Multimodal Large Model.

You, D., Liu, F., Ge, S., Xie, X., Zhang, J., and Wu, X. (2021). AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III*, volume 12903 of *Lecture Notes in Computer Science*, pages 72–82.

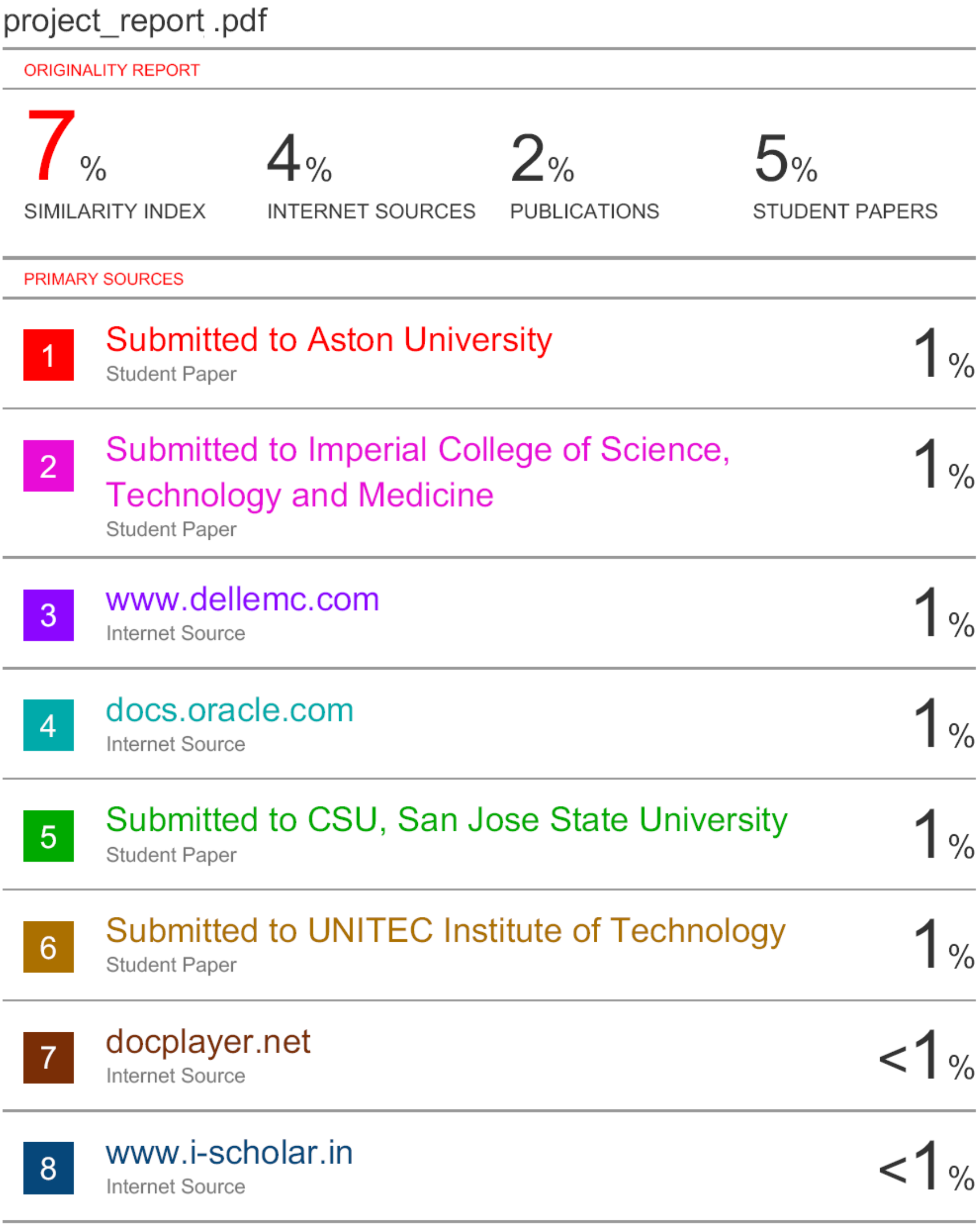
You, J., Li, D., Okumura, M., and Suzuki, K. (2022). JPG - Jointly Learn to Align: Automated Disease Prediction and Radiology Report Generation. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5989–6001.

Zhang, K., Jiang, H., Zhang, J., Huang, Q., Fan, J., Yu, J., and Han, W. (2023). Semi-supervised Medical Report Generation via Graph-guided Hybrid Feature Consistency. *IEEE Transactions on Multimedia*, pages 1–13.

Zhao, R., Wang, X., Dai, H., Gao, P., and Li, P. (2023). Medical report generation based on segment-enhanced contrastive representation learning.

Zhou, Y., Huang, L., Zhou, T., Fu, H., and Shao, L. (2021). Visual-textual Attentive Semantic Consistency for Medical Report Generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3965–3974.

APPENDIX



project_report .pdf

ORIGINALITY REPORT

7%	4%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Aston University Student Paper	1%
2	Submitted to Imperial College of Science, Technology and Medicine Student Paper	1%
3	www.dellemc.com Internet Source	1%
4	docs.oracle.com Internet Source	1%
5	Submitted to CSU, San Jose State University Student Paper	1%
6	Submitted to UNITEC Institute of Technology Student Paper	1%
7	docplayer.net Internet Source	<1%
8	www.i-scholar.in Internet Source	<1%