AUTOMATED TEXT GENERATION FROM RADIOLOGY IMAGES PROJECT REPORT | JUNE - AUGUST 2022

Sarah Hickman
University of Cambridge
Department of Radiology
sh2040@cam.ac.uk

Dan Schofield
NHS England
Transformation Directorate
daniel.schofield1@nhs.net

ABSTRACT

The automation of tasks in the radiology workflow, using machine learning, provides solutions for improvement in both efficiency and diagnostic accuracy of reporting. The current focus of machine learning applications to medical image interpretation is primarily for classification tasks. This differs from the routine clinical output of radiology, where findings are summarised by radiologists in free-text reports of varying levels of detail and complexity depending on the imaging modality and disease findings. Thus, radiology is a multi-modal specialty comprising of free-text and images.

This work was completed as part of an NHS England Data Science PhD Internship project, and the report builds on a previous internship project investigating the developments in multi-modal machine learning techniques to provide free-text as output, and the metrics used to evaluate the performance of such techniques.

1 Introduction

The application of machine learning (ML) to medical imaging has advanced over the past decade due to three major steps. Firstly the development of new model architectures providing an improvement in overall ML performance, secondly an increase in curated clinical data, and lastly an increase availability of computing power. ML algorithms are mainly used as either segmentation algorithms or clinical decision support systems for classification tasks at present, with algorithms focused on a specific task e.g. pneumothorax detection in a chest x-rays (CXRs), to improve diagnostic the accuracy, efficiency or both [1, 2]. Classification models are often binary in their output, with the model set at a specific operating threshold. Thus, models can be easily compared to the ground truth in order to evaluate performance, using commonly applied metrics in medicine; sensitivity, specificity and area under the receiver operating characteristics curve (AUROC). A recent systematic review by Lui et al. demonstrated the performance of ML models for "disease detection" in medical imaging can be considered equivalent to that of "health-care professionals" [3].

Radiology is a multi-modal speciality with each imaging series summarised by a radiologists in a free-text report which is then sent back to clinicians to act upon. Whilst the image captures the entirety of the body area being investigated, a free-text report may only provide details for the clinical diagnosis as well as any pertinent negatives within the image. Reports can vary in length from a few words to many paragraphs depending on the clinical question being asked, imaging modality being used, reporting style of the radiologist and complexity of disease findings.

Natural Language Processing (NLP) is already used in the dictation software radiologists use when reporting scans. However there are many other potential applications of NLP to radiology which require further investigation [4, 5, 6]. One such application is the the use of NLP to generate free-text reports, commonly referred to in the wider ML community as Natural Language Generation (NLG), in order to replicate the routine clinical output of radiology. The reporting of such models performance in clinical practice is limited in the literature and so it is challenging to ascertain the current performance of this technique. This report focuses on the application of NLG to the task of CXR report generation.

CXRs are the most common image modality requested in the UK with $\sim$8.3 million performed each year [7]. Multiple pathologies can be diagnosed from CXRs, however a CXR is often not definitive on its own and is combined with the history of the patient, blood test results, and clinical observations to lead to an overall diagnosis. In addition, further

imaging is often requested, such as a CT, where disease is detected. There are multiple publicly available CXR imaging databases but few databases include the corresponding radiological report alongside the images [8, 9]. This in part is due to the time consuming and often manual process of report anonymisation as well as data governance restrictions surrounding this type of patient data. As CXRs are one of the only imaging modalities with sufficient open database sizes to provide enough volume of both reports and images from which to train and evaluate models, this imaging modality was used to investigate the use of NLG to create free-text reports in this project.

This report summarises the work as part of the second Data Science PhD intern project "Automated Text Descriptions from Imaging". Project 2 extends the work conducted by Dekai Zhang in Project 1 where the `TxtRayAlign` model was developed. This project aims to address the following three questions; 1) what are the clinical applications of ML and NLG models to radiology, 2) how should the performance of NLG models be evaluated, and 3) how does NLG model performance change when adapting hyperparameters and testing on an external dataset?

Section 2 outlines the proposed clinical applications of ML and NLG models to the radiological workflow. Section 3 outlines the metrics used to evaluate NLG models. Section 4 provides and overview of the Indiana University (IU) dataset which is used as an example of an external dataset for testing. Section 5 details the methods for testing and technical requirements for this project. Section 6 details the results from model testing. Section 7 details the limitations and future directions. Section 8 provides an overall conclusion.

2 Background

ML can be used for a variety of clinical applications within the radiological workflow including for; image annotation / classification, alert system triage, image quality control, clinical decision support systems, cohort building for research and image report generation. Figure 1 provides an example of places ML and specifically NLG models could be used in the CXR radiology workflow.

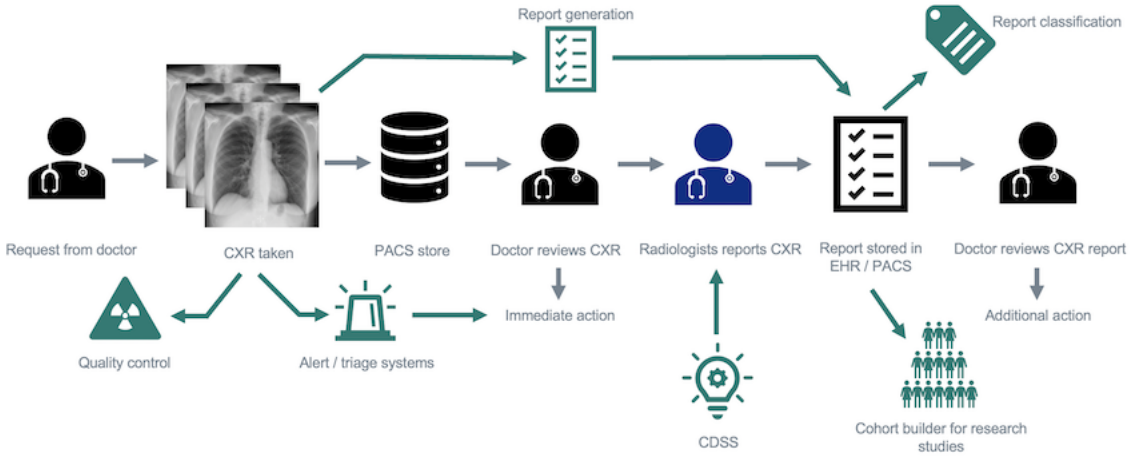


Figure 1: Proposed clinical workflow applications of ML to radiology - using the CXR workflow as an example. [CDSS = Clinical decision support system, CXR = Chest x-ray, EHR= Electronic health record, PACS = Picture archiving and communication system

Radiology reports are provided at different speeds depending on the urgency of the request. Due to the volume of scans produced, it can take anywhere from a few hours to days, for a study to reach the "top of the pile" to be reported. In addition, it can take anywhere from a few minutes to hours to write a report depending on the complexity of the imaging modality being reported (e.g. CT / MRI multi-slice imaging will take longer than a single view plain film x-ray).

In the hospital setting CXRs are routinely reviewed in the first instance by the requesting clinician, as shown in the workflow of Figure 1, thus decisions are often made by the clinician on the ward before the radiologist report is available. Therefore the generation of reports by NLG systems may not speed up this particular workflow but could act as a safety net to support junior clinicians reviewing the CXR, as well as ease the burden on radiologists by providing a preliminary report to be reviewed.

2.1 Report generation

The use of NLG to create structured or unstructured free-text reports has been explored through a variety of methods, these include; image captioning and report retrieval. The former captioning approach looks to generate a free-text report as an output of the model whereas the latter retrieval based approach generates the report from a corpus of reference reports. This second approach maybe more likely to retain the readability of a report, as the sentences have already been written by radiologists, as well as for this reason are more likely to pertain clinically relevant findings. However, there is also an increase in possibly including erroneous information relevant to the original study but not the current study. Using the full text report also allows for costly and time consuming labelling steps in classification tasks to be bypassed.

Two models model architectures are commonly used for NLG, these include:

- **Encoder-decoder models** - These are often used to generate image captions and labels, however they can also be applied to generate the full radiology report [10, 11]. The most common structure is a convolution neural network (CNN) architecture used as the encoder to extract image features in combination with a Recurrent Neural Network (RNN) architecture, usually a long short term memory (LSTM), as the decoder which then generates the report.
- **Contrastively trained models** - These models commonly use two encoders, one for the text and one for the image to extract features and then align these two inputs. Cosine similarity is then used to match the embeddings of the images and text to generate a report [12, 13].

The reports created by NLG models should aim to read like radiologists reports so that the language used and flow is familiar to clinicians reading reports. Reports can contain both the findings (the section of the report outlining all relevant positive and negative findings, which can be thought of as the workings out of a radiologist) and the impression (a summary of the main findings, likely diagnosis and recommendations), or potentially the report could just contain one section. Often "out of reach" information is used by the radiologist when writing reports, such as; electronic health records (EHR), the imaging request from the clinicians treating the patient, and other imaging studies from the same patient. Thus, reports can contain a conclusion which has been drawn from the summation of all this information to provide a diagnosis as well as recommendations for additional imaging and follow up. This can potentially lead to a models performance being penalised if it does not have access to and has not been trained to take into account this additional information when comparing to the ground truth of the original radiologist report.

2.2 Image annotation / classification

Key classification tasks could also be performed by ML models to provide helpful alert flags for clinicians as well as automated triage workflows to remove normal exams. However, these alerts need to be set at a clinically relevant specificity and sensitivity operating thresholds to the task being performed. Such that if a triage system is implemented for likely normal exams that are then either not immediately reviewed by a doctor or never reviewed, then the threshold for such a triage systems needs to be placed at a very high sensitivity to ensure false negatives (a disease is missed) do not occur. The opposite is true for alert systems where a high specificity should be set to avoid inundating clinician's with alerts. As CXRs can contain multiple pathologies multiple category labels can be produced for each disease, however this can lead to a potentially more complex multi-label output that is difficult to interpret compared to a free text report which clinicians are use to.

The ground truth from which to determine if a classification is correct or not can be provided by annotations from radiologists. Ground truth labelling could also be performed by an NLP model using the original radiology report to reduce the need for time and resource intensive labelling by radiologists [6].

2.3 Other uses

Alternative uses for ML models in the radiological workflow include:

- **Cohort building for research** - In order to identify cases for a specific research study radiology information systems (RIS) can be queried to retrieve a list of cases with a specified condition. NLP can then also be used to extract case labels from free-text reports to provide a ground truth from which to compare model performance to as discussed in the previous section. Models can also be used to review trends in radiology reports to provide evidence for the prevalence of conditions or radiological findings in epidemiological studies [4, 6].
- **Clinical decision support systems (CDSS)** - Traditionally CDSS provide prompts to radiologist on the image, also known as computer aided detection (CAD). CDSS could also fill in gaps where a condition has potentially been missed in the report by a radiologists and detected by the model, acting as a safety net to add this sentence

to the report. Alternatively, models can provide prompts to the radiologists with additional resources and information for the radiologist to interact with and choose whether to review and accept or not [4].

- **Visual question and answer (VQA)** - Similar to the report generation task but instead specific questions can be asked to the model from which a natural language response is generated [14]. Such that a clinician could query "Is there a nasograstric tube in this CXR?" and the model could respond "yes" or "no". The main limitation with this approach is the current lack of curated datasets for such a VQA model training and evaluation to be performed on in the medical domain. As patients records become digitised with the use of EHRs there is the potential for these VQA systems to allow doctors to search across the medical record, not just imaging, in a quicker and easier manner. This might be necessary as the volume of information in medical records continues to increase.

- **Quality control systems** - These systems could automate the quality control that exists in radiology to ensure the images taken are of a sufficient quality from which to provide an accurate report. Such as ensuring the key anatomical structures are included in the image. These systems could prompt radiographers to review images that need to be retaken before the patient leaves the department, avoiding delays to care from having to await repeat imaging. Such quality control systems could also be applied to radiology reports, to audit the consistency of reporting as well as use of reporting guidelines [15].

3 Evaluation metrics

There are a number of ways of evaluating ML model performance. For traditional classification tasks sensitivity, specificity and AUROC provide the foundation of most results and are easily interpreted by both clinicians and engineering teams. Specific NLG metrics were developed for the evaluation of models generating natural text or for machine translation (MT) tasks [16]. Previous studies have shown that when using these metrics for radiology report generation there are multiple limitations, including; they are unable to measure diagnostic accuracy, a good MT metric score can be achieved by a report with the opposite meaning from changing a single word, and MT scores decrease when using synonyms [10, 17]. Alternative metrics have been designed such as; CheXpert [18] and MIRQI [19] to obtain more clinically accurate scores. Thus, the evaluation of NLG models which produce radiology free-text reports to ensure the outputs are readable, clinically relevant, and accurate is challenging and there is no single agreed metric that provides this score.

3.1 Machine translation metrics

Metrics for MT tasks that have been applied to evaluate model performance for the generation of radiology reports include [20]:

- **BLEU (Bilingual Evaluation Understudy)** [21] - This is the most common MT metric used which measures the degree of difference between the original text and the ML generated text. It uses n-grams and looks for exact matches in the texts. The BLEU score focuses on precision and does not consider semantics or word order. A brevity penalty is applied to overcome the high scoring of short reports.

- **ROUGE (Recall Oriented Understudy for Gisting Evaluation)** [22] - There are multiple different types of ROUGE score. Many scores again use n-gram matching. In this project we used the ROUGE-L which uses the longest common subsequence (LCS) to calculate the model precision, recall and F1 score. The ROUGE score like BLEU does not consider semantics.

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** [23] - Takes into account unigram matching. The precision, recall, semantic similarity, order matching and a chunk penalty are applied to calculate an overall F-score.

BLEU, ROUGE and METEOR provide scores from 0-1, with 1 being the highest score. The output is compared to sentences from multiple references in MT tasks, however in medicine there is often only one report by one radiologist from which to compare [19]. In addition, these MT metric scores are unable to take the gravity of errors made by the ML models into account in the scores. Tools such as Datasets, NLTK and Microsoft COCO provide python packages from which to calculate these metrics. For this project all scores were calculated using the Datasets packages for each MT metric.

3.2 Clinical metrics

Two clinical metrics have been identified from the literature that allow for the reports to be scrutinized based on the inclusion of key radiological findings and thus evaluate the clinical accuracy of reports generated. These metrics include:

- **CheXpert** [18] - The CheXpert labeler can be used to classify 14 common radiological findings from the reports generated by the radiologist and NLG model into; positive, negative, uncertain and unmentioned. Comparing the results of these classifications the; precision, recall, and F1 score can be generated for models. Please see section 4.2 of the report for Project 1 for further details. This approach was used to evaluate the model performance in this project.
- **MIRQI (Medical Image Report Quality Index)** [13, 19] - Builds off the above CheXpert method. This was adjusted to include additional parameters for 20 common CXR radiological findings as well as account for modifiers, e.g. size, severity and body part affected. An adjustable weighting is applied for positive and negative findings. This method also takes into account synonyms. Labels are applied for positive, negative and uncertain findings from which the precision, recall, and F1 scores are calculated.

These two metrics are specific for the CXR report generation tasks and additional labeling lists would have to be created for alternative imaging modalities and body areas of investigation.

3.3 Clinical Scoring System

As outlined in the previous sections the metrics provide an overview for either the closeness to the original report in language used or the clinical accuracy, however both approaches are limited in being able to take into account the severity of errors and importance of relevant negations. A review of all generated reports by radiologists would be a timely exercise to carry out, however this could be improved with fixed scoring criteria from which the radiologist could score performance and an overall quality score can be calculated. A suggested Clinical Scoring System (CSS) is shown in Table 1.

Table 1: Quantitative radiologist report "Clinical Scoring System"

(ID) Measure	Description	Score
(1) Clinical correctness	Has the key condition/s been identified?	0 = No, 2 = Yes
(2) Pertinent negatives	Are the negatives correct and relevant?	0 = No, 1 = Yes
(3) Accurate recommendations	Has an appropriate recommendation been made?	0 = No, 1 = Yes
(4) Critical error present	Could this error lead to actual harm of the patient?	0 = Yes, 2 = No
(5) Minor error present	Has a minor error been made that is unlikely to lead to harm?	0 = Yes, 1 = No
(6) Erroneous information	Has additional unnecessary information been added?	0 = Yes, 1 = No

What this proposed CSS does not capture is readability and interpretabilty of the generated report. However, for the reports generated through retrieval, as the sentences have been written by radiologists, these are already written in understandable medical language. In future it could be that a 'Turing Test' is added to the CSS so that the radiologists first has to score which report they think was generated by the NLG model, which could be a proxy score for human interpretabilty [24].

This CSS could be used to compare to the other metrics to see which is best correlated with the radiologists interpretation CSS. It is possible multiple metrics should be used in combination to give an overview of model performance for the closeness to original report, clinical accuracy and readability.

However, the CSS still requires subjective and costly radiologists input to calculate. Thus, a hybrid approach incorporating the CSS into the CheXpert / MIRQI scores could be created to overcome this. A proposed incremental implementation below describes such an approach with decreasing radiologists input to no radiologists input required as all the categories are included in the automated metric.

- **MIRQI + CheXpert + CSS [1+4]** - Using the weighting that exists already in MIRQI to preference "clinical correctness" and ensure there is no "critical error" increasing the positive weighting and decreasing the negative weighting in the metric. In addition, using twelve of the fourteen labels from CheXpert that capture the key CXR findings / conditions (minus enlarged cardiomediasninum (*as covered in part by cardiomegaly) and

fracture (*as CXR should not be used for rib fracture diagnosis, the main purpose of the CXR is to look for other complications from rib fractures).

- **MIRQI + CheXpert + CSS [1+4+2]** - As mentioned in the example Case 1 and 3 in Appendix A there are many ways to say an exam is "normal". Firstly ensuring the labeling method is capturing all of the "normal" phrases. Secondly if the overall conclusion from an exam is a "normal" study then this should be taken as the overall negative weighting, ignoring the other negative finding statements in the calculation. "Pertinent negatives" that often differentiate depending on the radiologists reporting and more specifically the indication from the clinician are thus ignored and the model not penalised for not reporting the same negatives. Lastly, it would be of added benefit as an extra step to weight differently the negative finding that matches that of the indication in the request. For example if the clinician has stated in the request the patient has "New onset chest pain and shortness of breath ?pneumothorax?", and the pertinent negative in the report answers this question "Normal study. No pneumothorax" as appose to "Normal study. No pleural effusion" the former should be scored higher.
- **MIRQI + CheXpert + CSS [1+4+2+5+6]** - The presence of a "minor error" maybe detected in the modifiers already included in the MIRQI score. The additional conditions in the MIRQI not captured in the CheXpert labeler could be added here, e.g. hernia and calcinosis, to included the diseases that are not as critical as those outlined in the twelve labels to include in part 1. A checker for repetition could also be included to try to capture some of the "erroneous information" added in report generation.

Two additional aspects to include in this metric could be;

- **MIRQI + CheXpert + CSS [1+4+2+5+6+3]** - Include "Accurate recommendations" which would require the use of local guidelines (e.g. to request a six week follow up CXR for a new pneumonia diagnosis) and clinical pathway information which could be difficult to apply in a standardised metric to be used across different countries.
- **MIRQI + CheXpert + CSS [1+4+2+5+6+3]** - Including "New finding" which would only be possible to detect if taking into account previous films as well as information from EHRs. However, an increased weighting in the metric to a new finding / change in condition (e.g. in Appendix A Case 2 where there is resolution of the effusion) over an existing known condition that is stable could be applied to demonstrate the acute safety net aspect of such a model.

The ultimate aim of the NLG model is to generate a report that would lead to the same clinical outcome of the patient. This could be achieved by directly reproducing the same report, however as detailed there is significant subjectivity in reporting and multiple ways to describe the same abnormality and reach the same outcome. It is only possible to measure such a clinical outcome through prospective implementation of models, which first require benchmarking with a comparable robust metric to ensure clinically acceptable retrospective performance.

Appendix A details four example reports and corresponding metric scores from all the methods described in this section of the report.

4 IU dataset

The IU dataset consists of data from two USA hospital sites in the outpatient setting [9] and is available from the National Library of Medicine (NLM) image retrieval service (Open-i). There is only one study per patient and each patient can have one or more images, including the frontal and lateral x-rays with a patient on average having both views available in the majority of cases, as shown in Figure 2. In total there are 3955 cases with reports in XML format and 7471 images in PNG format. Thus, this dataset is significantly smaller than the MIMIC-CXR dataset (67,000 patients, 377,000 CXR images) which is described in the report from Project 1, and used in both Project 1 and 2 [8].

The IU dataset reports contain the indication, comparison, findings, and impression sections as well as MEDLINE manual Medical Subject Heading (MeSH) and Medial Text Indexer (MTI) labels. As part of processing the data, CheXpert labels for each case using the usual CheXpert labeller. A python script was applied to curate the data with the following steps to get the data into the same format as the MIMIC dataset used for Project 1:

- remove images without reports
- replace missing information in reports with 'no record available'
- replace XXX with underscores
- convert all XML reports into a single CSV file

- generate CheXpert labels for impression only, and impression and findings combined reports
- remove cases where NA CheXpert label was generated
- convert images from PNG to JPEG
- use the code from the `TxtRayAlign` to resize the images

An example of a case following processing is shown in Figure 3. The majority of cases in the IU dataset are "normal" cases. The three most common radiological findings are "cardiomegaly", "pulmonary atelectasis" and "calcified granuloma" according to the MeSH headings. Using the Chexpert labels the most common finding was again "No findings", with the three most common radiological findings as; "Lung Opacity", "Cardiomegaly" and "Support Devices". The most common negatives reported include; "Pleural Effusion", "Pneumothorax" and "Enlarged Cardiomediastinum".

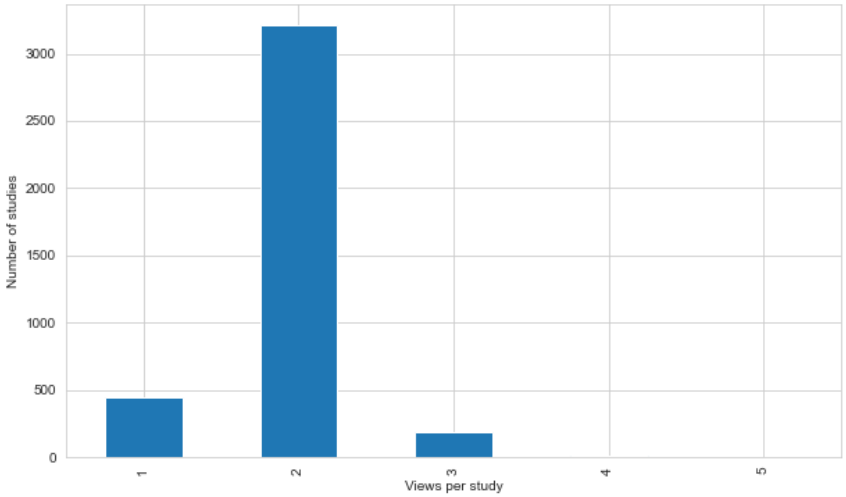


Figure 2: IU dataset number of images per patient episode

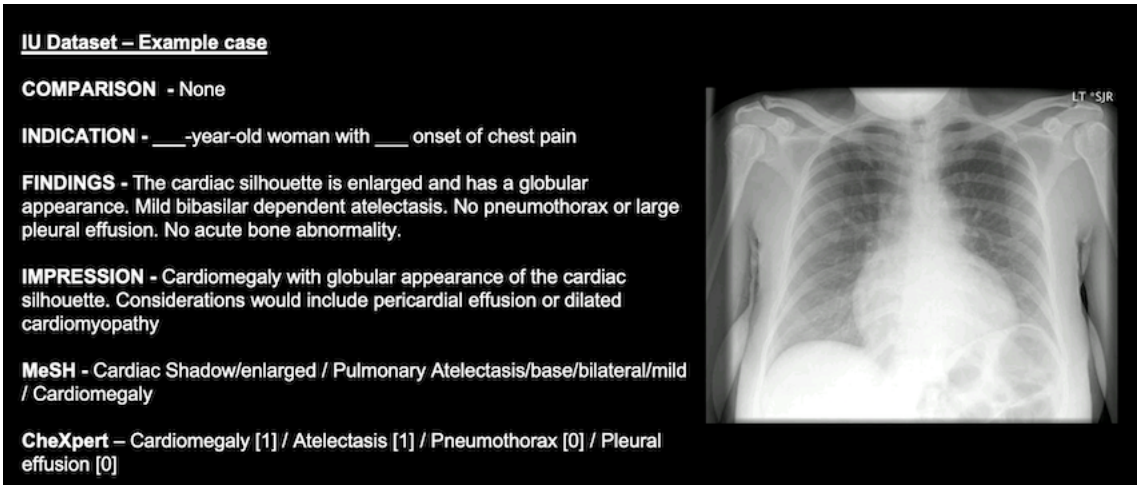


Figure 3: Example report from IU dataset

5 Methods and Requirements

The experiments carried out in Project 2 were conducted using the following methods and technical setup:

- **Data** - MIMIC dataset (Project 1 report, Section 3) and IU dataset (Project 2 report, Section 4)
- **Models** - The top three performing model combinations from Project 1 (ResNet50+Declutr, Resnet50+Distilbert and CLIP+CLIP) were re-trained and tested on 20% of MIMIC data at the patient level. The top performing algorithm from this test was then taken forward for all further experiment testing.
- **Evaluation metrics** - CheXpert label metrics were used for the evaluation of model performance for these experiments (Project 1 report Section 4.2).
- **Compute capability** - A machine hosted by Microsoft Azure with 1 x NVIDIA Tesla T4 GPU, 4 vCPUs from an AMD EPYC 7V12 processor, running Windows Server 2019.

6 Results

The results for each experiment are shown below with each study described in detail alongside the results.

Experiment 1. The first study took the top three performing architectures from Project 1 and re-trained the models on 20% of the data partitioned at the patient level with only one image AP / PA per patient using the impression from the reports. Keeping all hyperparameters the same as per Project 1 (Section 6.2, Table 3). The results are shown in Table 2 as well as Figure 4. The top performing algorithm (Resnet50+Declutr) in this experiment was taken forward for further testing.

Table 2: 20% MIMIC data at the patient level

Model	Hits	Duplicates	Flat Hit	Precision	Recall	F1
Resnet50+Declutr	22	0.10	0.44	0.21	0.27	0.23
Resnet50+Distilbert	21	0.12	0.42	0.19	0.27	0.22
CLIP+CLIP	16	0.14	0.32	0.15	0.21	0.18

Experiment 2. For the second study the top performing model was re-trained, adjusting the number of sentences sampled from the report each time during training from 1, to 2, or 3. The results are shown in Table 3. The results show that the model trained on two / three sentences instead of the usual one sentence performs better across all metrics, except duplicates.

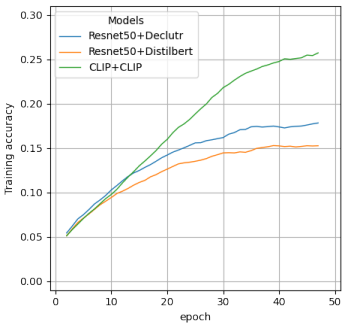
Table 3: N sentences used for training adjustment using 20% MIMIC data at the patient level

Model	Hits	Duplicates	Flat Hit	Precision	Recall	F1
Resnet50+Declutr, $N = 1$	22	0.10	0.44	0.21	0.27	0.23
Resnet50+Declutr, $N = 2$	29	0.14	0.58	0.28	0.33	0.30
Resnet50+Declutr, $N = 3$	28	0.15	0.56	0.25	0.29	0.27

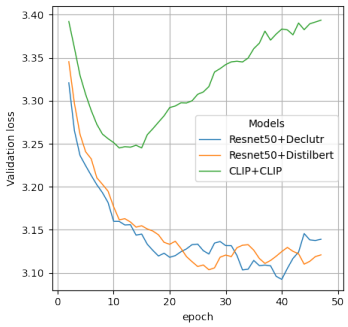
Experiment 3. The top performing model was then tested using IU data with impressions only and then impressions and findings to compare performance using IU and MIMIC data. The results are shown in Table 4. The model maintains / improves some performance metrics when tested using MIMIC images and IU text reports, however performance decreases when tested with MIMIC text and IU images, or both IU text and images.

Table 4: Performance on MIMIC and IU datasets for model trained using 20% MIMIC data at the patient level

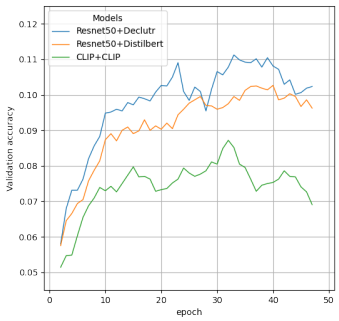
Model	Text	Image	Hits	Duplicates	Flat Hit	Precision	Recall	F1
IU Impression only	-	-	-	-	-	-	-	-
Resnet50+Declutr	MIMIC	MIMIC	22	0.10	0.44	0.21	0.27	0.23
Resnet50+Declutr	MIMIC	IU	11	0.09	0.22	0.09	0.16	0.11
Resnet50+Declutr	IU	MIMIC	24	0.11	0.48	0.21	0.27	0.24
Resnet50+Declutr	IU	IU	09	0.06	0.18	0.07	0.13	0.09
IU Impression + Findings	-	-	-	-	-	-	-	-
Resnet50+Declutr	MIMIC	MIMIC	22	0.10	0.44	0.21	0.27	0.23
Resnet50+Declutr	MIMIC	IU	18	0.08	0.36	0.14	0.08	0.10
Resnet50+Declutr	IU	MIMIC	24	0.13	0.48	0.24	0.26	0.25
Resnet50+Declutr	IU	IU	13	0.12	0.13	0.13	0.08	0.10



(a) Experiment 1 - Training Accuracy



(b) Experiment 1 - Validation Loss



(c) Experiment 1 - Validation Accuracy

Figure 4: Experiment training and validation graphs

7 Limitations and future directions

This project provided further exploration of potential applications of ML to radiology (including clinical concern), specifically how NLG models could be used to automate the generation of radiology reports for CXR images. Three potential uses for NLG models in the clinical radiological workflow highlighted include; 1) use as a safety-net for radiologists to auto-fill positive findings if not included in the report by the radiologist, 2) provide preliminary reports for acute CXRs to support junior doctors interpreting scans on the wards in the first instance whilst awaiting the radiologists report communicating critical findings, 3) automate follow up oncology scans, e.g. CT, reporting to provide a faster indication if a malignancy has progressed / quantifying response to therapy.

Some questions raised when carrying out this project included:

- Should only the impression section of the report be generated or both the impression and findings?
- Should both views be used together (PA / AP and lateral) by the model?
- Should a variable length of report be used and not the fixed output of two sentences?

Looking to the future, a proposal for an iteration of this work responding to these questions, that the most clinically useful model for the above proposed workflow applications, would be to; 1) generate the impression section of the report only, 2) use both views if available together to generate the report and 3) provide a variable length report to avoid repetition and the inclusion of erroneous information if a normal exam, as well as only include the key clinical diagnosis if a positive exam. In addition as this project focused on further investigating the retrieval based approach, future studies using an encoder-decoder model for a generation based approach to allow for the comparison of performance would be of interest.

As demonstrated by the examples in Appendix A, free text reports pose a challenging task from which to compare NLG outputs due to the "ambiguity, syntax, synonymy, medical abbreviations", use of negation, reference to "out of reach"

information, linking of associated findings, and overall individual variation in reporting style seen between different radiologists as highlioghted in [25]. For this project we described six different evaluation metrics, one of which was developed as part of this project (CSS). In the next project a comparison on a large scale of reports using the metrics should be carried out to see which metric correlates best with the CSS.

This project also highlights when building large radiological imaging databases that both the images and reports should be stored. A standardised anonymisation protocol may have to be developed so that as much information is retained as possible whilst ensuring no patient identifiable data is included.

Here we highlight some of the limitations to this project. Firstly, the data was from open access datasets from one country (USA). The healthcare system in the USA differs from the UK, and thus the requesting of radiological investigations by clinicians can differ also. In addition, as this data is from the USA, the language used will also differ from the UK in the general and medical terms used. However, as training and testing took place using data only from the USA this did not effect this project directly. Secondly, there was only a limited number of datasets that contain both the radiological report and CXR images. There is an additional database, PadChest, which could be considered for this project as it also contains both the images and reports. However, PadChest reports are in Spanish, which again may limit the applicability of the results to the NHS [26]. Lastly, the project focused on the CXR workflow, but this project provides scope from which these NLG methods can be developed for other workflows and imaging modalities.

Key aspects in any ML healthcare project were discussed and considered throughout this project, including;

- **Bias** – such as bias in the data used for testing, model bias which could occur from data used for training, and automation bias from an over-reliance of clinicians on ML systems.
- **Explainability (XAI)** - NLG models provide explainable and interpretable outputs for clinicians and radiologists in a free text format which they are use to interacting with.
- **Handling of uncertainties and disagreements** - if a clinician disagrees with a ML model or there is uncertainty about a ML models decision it is important that arbitration mechanisms are built into the workflow. For routine tasks this could be managed by the point below of keeping a "human in the loop" with sufficient training to override the models decision. It is also important that cases where there are disagreements or uncertainties are reviewed and evaluated for why such a disagreement took place and for the frequency of such instances.
- **Human in the loop / on the loop** - whilst it has been proposed that ML models could operate as entirely stand-alone systems in the first instance keeping a trained clinician within the same workflow to check ML decision is important as outlined in the point above.
- **Generalisability** - as highlighted in the limitations section of this report the data was primarily from one country (USA) and thus the generalisability of results to the UK NHS workflow needs to be considered in future testing.
- **Training of medical staff and expert de-skilling** - the aim of these models is not to replace radiologists but to support radiologists and provide mechanisms to improve accuracy and efficiency. Time consuming tasks that do not require the expertise of a radiologists could be supplemented with ML to free up the time of a radiologists to utilise their skills in specialist areas. However, it is important radiologists are still trained to interpret these scans without the assistance of ML systems so as not to result in de-skilling.

8 Conclusion

The use of NLG for the auto generation of radiology reports has the potential to provide multiple radiology workflow applications. The metrics used to evaluate the performance of models for clinical tasks require further refinement to ensure clinical accuracy is captured. The effect on model performance from adapting model training as well as performance on external dataset was also conducted.

9 Code

The link to the project GitHub can be found here: `https://github.com/nhsx/txt-ray-align/`

References

[1] Brendan S. Kelly, Conor Judge, Stephanie M. Bollard, Simon M. Clifford, Gerard M. Healy, Awsam Aziz, Prateek Mathur, Shah Islam, Kristen W. Yeom, Aonghus Lawlor, and Ronan P. Killeen. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *European Radiology*, 2022.

[2] Alice C. Yu, Bahram Mohajer, and John Eng. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiology: Artificial Intelligence*, 4(3):1–9, 2022.

[3] Xiaoxuan Liu, Livia Faes, Aditya U. Kale, Siegfried K. Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R. Ledsam, Martin K. Schmid, Konstantinos Balaskas, Eric J. Topol, Lucas M. Bachmann, Pearse A. Keane, and Alastair K. Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

[4] E Pons, L M.M Braun, M.M.G Hunink, and J.A Kors. Natural Language Processing in Radiology: A Systematic Review. *Radiology*, 279(2), 2016.

[5] Tianrun Cai, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. Natural language processing technologies in radiology research and clinical applications. *Radiographics*, 36(1):176–191, 2016.

[6] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu, and Beatrice Alex. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21(1):1–18, 2021.

[7] B. J. Stevens, L. Skermer, and J. Davies. Radiographers reporting chest X-ray images: Identifying the service enablers and challenges in England, UK. *Radiography*, 27(4):1006–1013, 2021.

[8] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):1–8, 2019.

[9] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[10] Zaheer Babar, Twan van Laarhoven, and Elena Marchiori. Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines. *PLoS ONE*, 16(11 November):1–20, 2021.

[11] William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for Chest X-Ray Report Generation. In Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 126–140. PMLR, 13 Dec 2020.

[12] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model. *Proceedings of Machine Learning Research*, 158:209–219, 2021.

[13] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. pages 1–15, 2020.

[14] Eric Lehman, Vladislav Lialin, Katelyn Y. Legaspi, Anne Janelle R. Sy, Patricia Therese S. Pile, Nicole Rose I. Alberto, Richard Raymund R. Ragasa, Corinna Victoria M. Puyat, Isabelle Rose I. Alberto, Pia Gabrielle I. Alfonso, Marianne Taliño, Dana Moukheiber, Byron C. Wallace, Anna Rumshisky, Jenifer J. Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. Learning to Ask Like a Physician. 2022.

[15] Emma M. Davidson, Michael T.C. Poon, Arlene Casey, Andreas Grivas, Daniel Duma, Hang Dong, Víctor Suárez-Paniagua, Claire Grover, Richard Tobin, Heather Whalley, Honghan Wu, Beatrice Alex, and William Whiteley. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Medical Imaging*, 21(1):1–13, 2021.

[16] Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. Deep learning in generating radiology reports : A survey. *Artifical Intelligence in Medicine*, 2020.

[17] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. *Multimodal recurrent model with attention for automated radiology report generation*, volume 11070 LNCS. Springer International Publishing, 2018.

[18] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference,*

IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pages 590–597, 2019.

[19] Pablo Pino. *X-rays : Analysis of NLP metrics and clinically correct template-based model.* PhD thesis, 2022.

[20] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Computing Surveys*, 55(2):1–39, mar 2023.

[21] K Papineni, S Roukos, T Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation LK - https://rug.on.worldcat.org/oclc/204431620. *Annual Meeting- Association for Computational Linguistics Ta - Tt -*, 40(July):311–318, 2002.

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[23] Satanjeev Banerjee and Alon 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*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[24] Hanqiang Ouyang, Fanyu Meng, Jianfang Liu, Xinhang Song, Yuan Li, Yuan Yuan, Chunjie Wang, Ning Lang, Shuai Tian, Meiyi Yao, Xiaoguang Liu, Huishu Yuan, Shuqiang Jiang, and Liang Jiang. Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test. *Frontiers in Oncology*, 12(March):1–12, 2022.

[25] Lane F. Donnelly, Robert Grzeszczuk, and Carolina V. Guimaraes. Use of Natural Language Processing (NLP) in Evaluation of Radiology Reports: An Update on Applications and Technology Advances. *Seminars in Ultrasound, CT and MRI*, 43(2):176–181, 2022.

[26] Aurelia Bustos, Antonio Pertusa, Jose Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:1–35, 2020.

A Appendix - Example metrics

Below are four examples of cases from which a CXR report was generated by a NLG model and compared to a radiologist report using the metrics outlined in the Section 3 of this report. All the cases in this Appendix are from the IU dataset.

In our experiments two sentences were output as the results. These can be compared to the impression and findings sections of the radiologists report. However, as highlighted in Project 1, a variable length report generator could be of use to adapt to the length of reports shown in the original radiologists outputs.

Case 1 (Figure A.1) - shows a normal CXR which is captured in the radiologists and generated reports. There are many ways to say something is normal e.g, "normal", "no findings", "no acute process", "no cardiopulomnary findings", etc. A sentence noting the lack of sensitivity of CXRs for the diagnosis of rib fractures is added to the generated report, which would not lead to harm and is a reminder for clinicians. However, in adding this sentence this has led to the incorrect allocation of labels by both CheXpert and MIRQI for a fracture being present. The results using the different scoring metrics are shown in Table A.1 and CheXpert and MIRQI labels are also shown below.

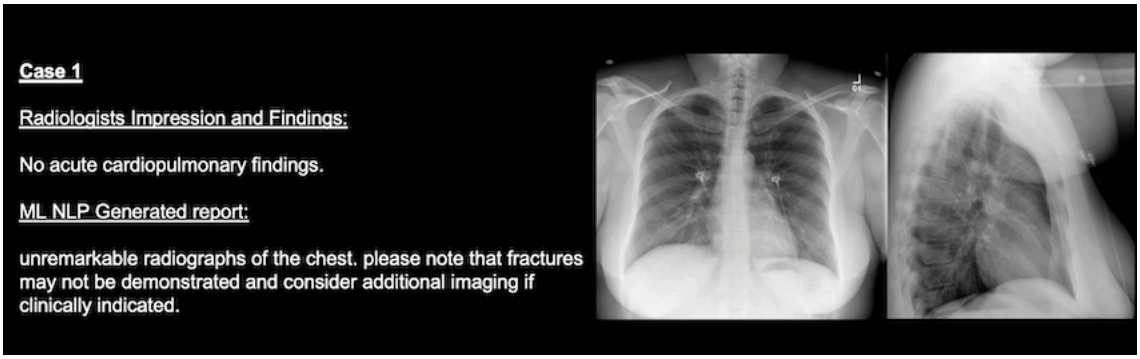


Figure A.1: Case 1 - Radiologist original report and ML NLP model report

Table A.1: Case 1 - Example results from one generated report

Measure	Score
BLEU	0.04, 0.02, 0.01, 0.01
ROUGE	Precision = 0.0 Recall = 0.0 F1 = 0.0
METEOR	0.39
CheXpert	Precision = 0.0 Recall = 0.0 F1 = 0.0
MIRQI	Precision = 0.0 Recall = 0.0 F1 = 0.0
CSS	8

CheXpert

- Radiologists = No Finding [1]
- Generated = Fracture [1]

MIRQI

- Radiologists = NA
- Generated = 'fracture', 'Fracture', 'POSITIVE'

Case 2 (Figure A.2) shows a normal CXR following the resolution of a plural effusion mentioned by the radiologists by comparing with a previous CXR film. The generated report reports a "worsening of known tumour" as well as a left sided effusion which has "increased" compared to the previous film, which is the opposite of the radiologists report. The reporting of a tumour which is not present is also a critical error which could lead to harm, both through the exposure to further radiation from CT and biopsies, as well as patient psychological harm through anxiety whilst awaiting confirmation results. The results using the different scoring metrics are shown in Table A.2 and CheXpert and MIRQI labels are also shown below.

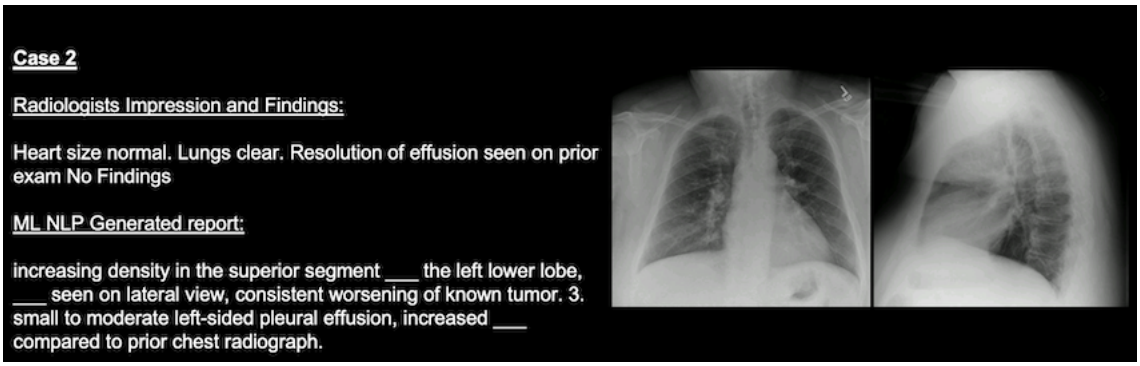


Figure A.2: Case 2 - Radiologist original report and ML NLP model report

Table A.2: Case 2 - Example results from one generated report

Measure	Score
BLEU	0.14, 0.02, 0.01, 0.01
ROUGE	Precision = 0.1 Recall = 0.21 F1 = 0.13
METEOR	0.35
CheXpert	Precision = 0.0 Recall = 0.0 F1 = 0.0
MIRQI	Precision = 0.0 Recall = 0.0 F1 = 0.0
CSS	1

CheXpert

- Radiologists = No Finding [1] / Cardiomegaly [0] / Pleural Effusion [0]
- Generated = Lung Lesion [1] / Pleural Effusion [1]

MIRQI

- Radiologists = 'heart size', 'Cardiomegaly', 'NEGATIVE', 'normal' / 'effusion', 'Pleural Effusion', 'NEGATIVE'
- Generated = 'tumor', 'Lung Lesion', 'POSITIVE', 'known' / 'density', 'Other Finding', 'POSITIVE', 'increasing' / 'effusion', 'Pleural Effusion', 'POSITIVE', 'moderate/left-sided/pleural'

Case 3 (Figure A.3) shows a normal CXR. As noted in Case 1 there are many ways to say an exam is normal. Sometimes the radiologist will expand on the no findings statement to report the pertinent negatives as to how they have reached the conclusion of "normal", such as "no pneumothorax or pleural effusion". Similarly in the generated a normal exam is reported, however the generated report uses different phrasing for normal for certain findings e.g. "clear lungs" = "no alveolar consolidation". However, the CheXpert and MIRQI labels have not picked up the overall "no findings" / negatives from the generated report. The results using the different scoring metrics are shown in Table A.3 and CheXpert and MIRQI labels are also shown below.

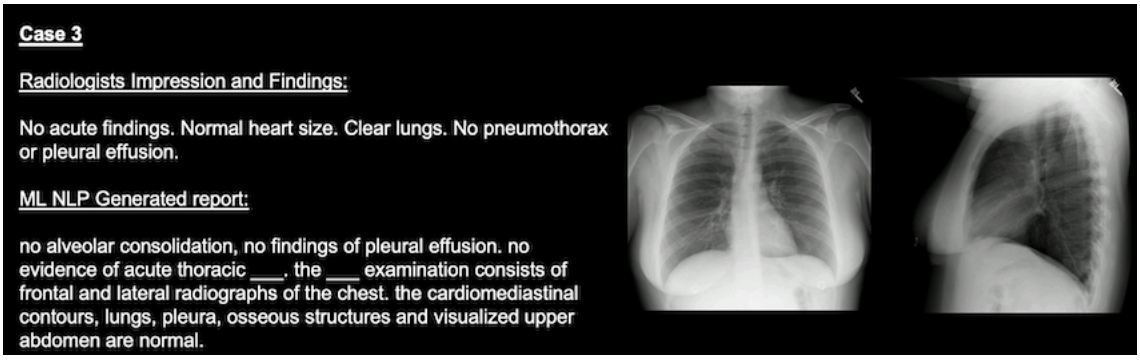


Figure A.3: Case 3 - Radiologist original report and ML NLP model report

Table A.3: Case 3 - Example results from one generated report

Measure	Score
BLEU	0.17, 0.04, 0.02, 0.01
ROUGE	Precision = 0.12 Recall = 0.31 F1 = 0.16
METEOR	0.52
CheXpert	Precision = 0.33 Recall = 0.25 F1 = 0.14
MIRQI	Precision = 0.20 Recall = 0.20 F1 = 0.20
CSS	7

CheXpert

- Radiologists = No Finding [1] / Cardiomegaly [0] / Pneumothorax [0] / Pleural Effusion [0]
- Generated = Enlarged Cardiomediastinum [1] / Consolidation [0] / Pleural Effusion [0]

MIRQI

- Radiologists = 'heart size', 'Cardiomegaly', 'NEGATIVE', 'normal' / 'pneumothorax', 'Pneumothorax', 'NEGATIVE', 'no/" / 'effusion', 'Pleural Effusion', 'NEGATIVE'
- Generated = 'effusion', 'Pleural Effusion', 'NEGATIVE', 'pleural' / 'consolidat', 'Consolidation', 'NEGATIVE', 'no/alveolar/findings' / 'contour', 'Enlarged Cardiomediastinum', 'POSITIVE', 'cardiomediastinal/"

Case 4 (Figure A.4) show a CXR with a left-sided pneumothorax and a catheter in situ for drainage. The generated report misinterprets the CXR and reports a malignancy in the right upper lung which is a critical error on two parts, firstly for missing the pneumothorax and secondly for suggesting a cancer where there is not one present. The results using the different scoring metrics are shown in Table A.4 and CheXpert and MIRQI labels are also shown below.

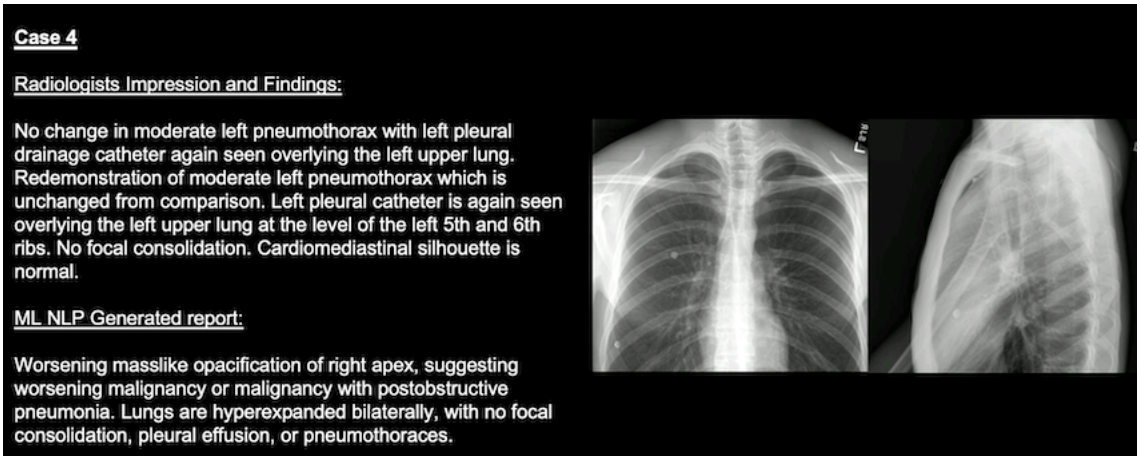


Figure A.4: Case 4 - Radiologist original report and ML NLP model report

Table A.4: Case 4 - Example results from one generated report

Measure	Score
BLEU	0.20, 0.03, 0.02, 0.01
ROUGE	Precision = 0.15 Recall = 0.07 F1 = 0.10
METEOR	0.17
CheXpert	Precision = 0.17 Recall = 0.25 F1 = 0.10
MIRQI	Precision = 0.20 Recall = 0.10 F1 = 0.13
CSS	0

CheXpert

- Radiologists = Enlarged Cardiomediastinum [1] / Consolidation [0] / Pneumothorax [1] / Support Devices [1]
- Generated = Lung Lesion [1] / Lung Opacity [1] / Consolidation [0] / Pneumonia [1] / Pneumothorax [0] / Pleural Effusion [0]

MIRQI

- Radiologists = 'pneumothorax', 'Pneumothorax', 'POSITIVE', 'moderate/left' / 'catheter', 'Support Devices', 'POSITIVE', 'left/pleural/drainage' / 'pneumothorax', 'Pneumothorax', 'POSITIVE', 'moderate/left' / 'consolidat', 'Consolidation', 'NEGATIVE', 'no/focal' / 'mediastinal silhouette', 'Enlarged Cardiomediastinum', 'POSITIVE', 'cardiomediastinal/
- Generated = 'mass', 'Lung Lesion', 'POSITIVE', 'opacification' / 'pneumonia', 'Pneumonia', 'POSITIVE', 'postobstructive' / 'opaci', 'Airspace Opacity', 'POSITIVE', 'worsening/masslike/suggesting' / 'pneumothoraces', 'Pneumothorax', 'NEGATIVE' / 'effusion', 'Pleural Effusion', 'NEGATIVE', 'consolidation/pleural' / 'consolidat', 'Consolidation', 'NEGATIVE', 'no/focal/effusion/'