# Catching Misdiagnosed Limb Fractures in the Emergency Department Using Cross-institution Transfer Learning

**Filip Rusak**[1*] **Bevan Koopman**[1] **Nathan J. Brown**[2]
**Kevin Chu**[2] **Jinghui Liu**[1] **Anthony Nguyen**[1]
[1] The Australian e-Health Research Centre, CSIRO
[2] Emergency and Trauma Centre, Royal Brisbane and Women's Hospital
filip.rusak@connect.qut.edu.au
{nathan.brown3,kevin.chu}@health.qld.gov.au
{bevan.koopman,jinghui.liu,anthony.nguyen}@csiro.au

## Abstract

We investigated the development of a Machine Learning (ML)-based classifier to identify abnormalities in radiology reports from Emergency Departments (EDs) that can help automate the radiology report reconciliation process. Often, radiology reports become available to the ED only after the patient has been treated and discharged, following ED clinician interpretation of the X-ray. However, occasionally ED clinicians misdiagnose or fail to detect subtle abnormalities on X-rays, so they conduct a manual radiology report reconciliation process as a safety net. Previous studies addressed this problem of automated reconciliation using ML-based classification solutions that require data samples from the target institution that is heavily based on feature engineering, implying lower transferability between hospitals. In this paper, we investigated the benefits of using pre-trained BERT models for abnormality classification in a cross-institutional setting where data for fine-tuning was unavailable from the target institution. We also examined how the inclusion of synthetically generated radiology reports from ChatGPT affected the performance of the BERT models. Our findings suggest that BERT-like models outperform previously proposed ML-based methods in cross-institutional scenarios, and that adding ChatGPT-generated labelled radiology reports can improve the classifier's performance by reducing the number of misdiagnosed discharged patients.

## 1 Introduction

When a patient presents to the Emergency Department (ED) with a possible limb fracture, ED clinicians order an X-ray from the radiology department. Following imaging, a radiologist authors a report stating the radiological observations and diagnosis, which is then sent back to the ED clinician requesting the procedure. Unlike radiology images, radiology reports may not be completed before a patient leaves the ED. In such cases, ED clinicians interpret radiological images themselves (Koopman et al., 2015). Occasionally, ED clinicians misdiagnose radiological evidence such as subtle limb abnormalities (e.g., small fractures, dislocations or foreign bodies), resulting in patients being discharged without appropriate treatment (Koopman et al., 2015; Zuccon et al., 2013). As a safety net, ED clinicians retrospectively reconcile radiology report findings with ED discharge diagnoses to detect potential misdiagnoses (Koopman et al., 2015). Since the radiology report reconciliation process is retrospective and performed manually, it may take several days to identify and notify a misdiagnosed patient, exposing them to potentially adverse impacts on their health (Koopman et al., 2015; Masino et al., 2016).

Machine Learning (ML)-based methods for classifying radiology reports (Koopman et al., 2015; Zuccon et al., 2013; de Bruijn et al., 2006; Zhou et al., 2014; Hassanzadeh et al., 2018b) have the potential to streamline and semi-automate the radiology report reconciliation process. However, the development of ML solutions is dependent on the availability of large and diverse labelled datasets from target hospitals for model training (Gligic et al., 2020). While radiology reports may be readily available, labelling them requires domain expertise, is time-consuming and costly (Hassanzadeh et al., 2018b). Therefore, individual departments or hospitals may not have the capacity to collect sufficiently large datasets of labelled radiology reports to conduct their own model training (Li et al., 2021a). Cross-institution transfer learning, in which datasets and model training from one institution are used to start the ML model development at another institution, may solve this problem. However, for cross-institution transfer learning to be useful for developing local ML models for radiology report reconciliation, it must be resilient to interinstitutional variations in reporting styles, lan-

[*] Conducted this research while affiliated with CSIRO.

guage, and verbosity (Hassanzadeh et al., 2018b; Liu et al., 2022).

Many pre-trained Transformer-based language models have achieved state-of-the-art performance in various benchmark datasets (Jia, 2022; Li et al., 2021b), especially Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). In this study, we investigated the benefits of pre-trained BERT-like models on a radiology report classification task in cross-institutional environments, where labelled data from the target institution are unavailable. Although domain-specific pre-training appears to be effective for in-domain applications (Peng et al., 2019), little is known about the impact of pre-training with different corpora on a radiology report classification task in cross-institution settings. Therefore, we focused on answering the following research questions:

**RQ1 - What is the impact of pre-training on the abnormal radiology report classification task?**

To answer RQ1, we chose six different BERT models pre-trained using the medical and biomedical corpora and evaluated them on cross-institution radiology report classification, based on data from three Australian hospitals. We then used the best performing model – PubMedBERT – to explore:

**RQ2 - Can we train a model that is transferable between institutions without relying on samples from the target institution?**

To answer RQ2, we compare PubMedBERT with previously proposed SVM and CNN-based radiology report classification models. Our observations indicate that fine-tuned PubMedBERT models are more transferable in cross-institutional settings than previously proposed SVM and CNN-based solutions. Since labelled radiology reports needed for fine-tuning are scarce (Li et al., 2022), one of the remedies to mitigate the lack of labelled samples is to utilise synthetic radiology reports, generated according to the desired class condition, to diversify the fine-tuning set and boost classification performances. In particular, the recently released ChatGPT shows impressive text generation capabilities and high potential to generate discharge summaries (Patel and Lam, 2023). In contrast to the proposal that ChatGPT be used in the context of generating high-quality discharge summaries to offload junior doctors (Patel and Lam, 2023), we investigate the benefit of using ChatGPT as an additional source of data to fine-tune abnormal radiology report classification (BERT) models. Then we aim to answer the following research question:

**RQ3 - What is the impact of using ChatGPT as an additional data source of synthetic data for classification model fine-tuning?**

Note that we only use ChatGPT to supplement fine-tuning data; our empirical evaluation was still performed with a carefully curated set of real radiology reports by clinicians. We found that including ChatGPT-synthesised radiology reports in fine-tuning improves abnormal radiology record classification performances.

Lastly, we examine the practical application of using pre-trained BERT-like models, fine-tuned on real and synthetic radiology reports, to classify and reconcile radiology reports with ED discharge diagnoses in a clinical environment. Reconciling radiology reports with the corresponding ICD-10 discharge diagnoses from the ED system can result in four outcomes: 1) Both Abnormal; 2) Both Normal; 3) Radiology Abnormal, ED normal; and 4) Radiology Normal, ED Abnormal. By doing so, we are answering the following research question:

**RQ4 - How does adding ChatGPT synthesised radiology reports to fine-tuning dataset impact the downstream reconciliation task?**

When answering RQ4, we particularly pay attention to the difference in reconciliation performance (confusion matrix) between models that included ChatGPT-synthesised reports in fine-tuning and the models that did not. We observed that the PubMedBERT classifier, fine-tuned in real and synthetic radiology records, improves the detection of misdiagnosed patients at the expense of a higher number of records that require manual clarification.

The contribution of this work is fourfold: 1) we demonstrated that pre-trained models generalise better in the case of abnormal radiology report classification in cross-institution settings; 2) we highlighted the impact of ChatGPT on fine-tuning abnormal radiology report classification; and 3) we extended the impact of ChatGPT-generated synthetic report on a downstream reconciliation task.

## 2 Related Work

Common challenges of supervised ML models that support clinical decisions arise from limited clinical data and the lack of their labels, especially when the model is trained with data from a single hospital (Li et al., 2021a). The lack of labelled samples from a target hospital has previously been addressed by using transfer learning

(Gligic et al., 2020), leveraging training sets with labelled (Koopman et al., 2015; Li et al., 2021a) and unlabelled (Hassanzadeh et al., 2018a) data from multiple institutions. However, there are no prior examples of scenarios in which there is no training data at all from the target hospital. In this study, we investigated whether ML automation of the radiology report reconciliation process in a target ED could rely on a training data set that originated from an entirely different hospital.

For our purposes, the model architecture of choice must be able to generalise well across institutions. Methods relying on feature engineering, such as support vector machine (SVM), naïve bayes, or random forest, are not suitable for cross-institution settings since features engineered for a dataset collected at one institution may not be the best fit for data collected at another institution (Xiao et al., 2018). This was also observed by Koopman et al. (2015) who found a significant reduction in performance (F1-Score) of up to 10–12% in SVM-based radiology report classifiers, when the training source institution was different from the target, the test institution. Hassanzadeh et al. (2018b) further demonstrated the dependency on pre-defined feature engineering by showing improved F1-score of 5-10% across hospitals when employing self-feature-extracting CNNs with feature adoption transfer. However, to achieve such improvements in performance still required training data from the target institution. Unlike SVMs and CNNs, Transformer models take advantage of the attention mechanism capable of extracting textual features (location, context, syntactic structure, and semantics), which leads to better performance (Jia, 2022). Transformer-based models, such as pre-trained BERT models, are some of the most successful deep learning (DL) models for natural language processing (NLP) across domains (Zaheer et al., 2020). Therefore, we chose pre-trained BERT-like models for the current study.

Data synthesis is one technique that can mitigate the shortage of labelled training/fine-tuning data. We determined whether ChatGPT-generated synthetic radiology reports could be used to augment training or fine-tuning datasets for the purpose of reconciliating radiological findings. Additionally, we evaluate the impact of ChatGPT-generated reports on the performance of the BERT-based abnormal radiology report classifier when ChatGPT-synthesised reports are included in the fine-tuning dataset. Although ChatGPT has already been explored for data augmentation (Dai et al., 2023), little has been studied to evaluate the impact of ChatGPT-generated *radiology reports* on increasing the performance of the BERT-based classification model, fine-tuned on real samples with and without synthetic reports.

## 3 Materials and Methods

**ChatGPT**. ChatGPT [1], recently developed by OpenAI, is one of the largest language models to date (about 175 billion parameters) based on GPT-3 (Brown et al., 2020). ChatGPT is a generative language model that is designed to generate natural language according to some input prompt. The quality of its generated language is driven in part by the extensive text it was provided as part of the training process.

**Data**. In this study, we used four datasets of free-text limb structure radiology reports; three acquired from the ED of three Australian public hospitals (2378 reports), and a synthetic dataset created using ChatGPT (100 reports). The hospital-acquired radiology reports comprise anonymised adult, children, and mixed (adult and children) reports from three hospitals located in southeast Queensland, Australia. Ethical approval for the acquisition of these data was granted by the Human Research Ethics Committee of the Royal Brisbane and Women's Hospital.

Real free-text radiology reports were manually assessed by two emergency medicine physicians as either "normal" (no fractures, dislocations, or foreign bodies present) or "abnormal" (fractures, dislocations, or foreign bodies present). A software tool was developed to help physicians record their interpretations and highlight the relevant portions of text in the reports. Initially, the assessors agreed on the annotations of 2,215 out of 2,378 reports. A senior physician was then asked to act as a third assessor and resolve disagreements. The dataset distribution from three hospitals (RBWH, RCH and GCH), including the number of reports, the proportion of normal and abnormal cases, the average length of words and the number of unique words in the dataset, are presented in Table 1. The Fleiss kappa ($\kappa$) of 0.85 was calculated from the initial annotations of the first two assessors, indicating a high level of inter-rater reliability.

The 100 synthetic radiology reports – 50 normal and 50 abnormal – were generated using ChatGPT

[1] https://openai.com/blog/chatgpt

| Dataset | Description | #Reports | Normal | Abnormal | Avg. Doc. | #Unique words |
|---|---|---|---|---|---|---|
| RBWH | Royal Brisbane & Womens' Hospital (adult) | 1480 | 58% | 42% | 52 words | 1944 |
| RCH | Royal (Brisbane) Childrens' Hospital (child) | 498 | 66% | 34% | 50 words | 1100 |
| GCH | Gold Coast Hospital (adult 62% & child 38%) | 400 | 62% | 38% | 27 words | 558 |
| ChatGPT | Synthetic reports generated by ChatGPT (adult) | 100 | 50% | 50% | 76 words | 201 |

Table 1: Four different datasets of radiology reports, the number of normal and abnormal cases as identified through our annotation process or conditional generation, and document length for free-text reports document-wise.

prompts listed in Appendix Table 5. To ensure the variability between the synthetic radiology reports generated, we followed the initial with additional prompts. Synthetic reports with only minimal changes (e.g., patient name, age) and the same diagnosis were discarded.

**RQ1 - What is the impact of pre-training on the abnormal radiology report classification task?** We evaluated the six pre-trained BERT-based models on the free-text radiology report classification task to identify abnormalities of limb structures (normal vs abnormal). Six pre-trained models were selected based on their score on the Biomedical Language Understanding and Reasoning Benchmark (BLURB) [2] at the time of conducting experiments. BLURB includes a comprehensive benchmark for PubMed-based biomedical NLP applications and a leaderboard for tracking community progress. We evaluated the following six pre-trained BERT-like models on the cross-institutional radiology report classification task: PubMedBERT (Gu et al., 2021), BERT (Devlin et al., 2019), LinkBERT (base and large) (Yasunaga et al., 2022), BioClinicalBERT (Alsentzer et al., 2019), BlueBERT (base and large) (Peng et al., 2019) and BioELECTRA (base and large) (Kanakarajan et al., 2021). These models are pre-trained on different corpora from different domains (Appendix Table 6). The difference between base and large BERT models is in the number of layers (12 vs 24), hidden layer size (768 vs 1024) and the number of self-attention heads (12 vs 16). PubMedBERT (Gu et al., 2021) is pre-trained from scratch on biomedical article corpora, including both abstracts and full-text articles, from PubMedCentral [3]. LinkBERT is a BERT-based model pre-trained on a large corpus of documents and their links (e.g., hyperlinks, citation links) to incorporate knowledge spanning across multiple documents. BioClinicalBERT is pre-trained in all MIMIC III notes. BlueBERT models were trained on pre-processed PubMed texts extracted from the PubMed ASCII code version, containing approximately 4000 million words. BioELECTRA models were pre-trained on PubMed abstracts only with biomedical domain vocabulary.

While each model was pre-trained on different corpora, we benchmarked the mentioned models to determine the impact of model pre-training on a classification task on our mixed datasets (RBWH, RCH, and GCH). Since our dataset is relatively small, consisting of only 2378 radiology reports from all three hospitals, we chose to evaluate the pre-trained models under test with 5-fold cross-validation. Each model was fine-tuned for ten epochs per fold, with a learning rate of 9e-6 and randomly selected seed of 112. We compared F scores, precision, recall, and Matthew's correlation coefficients (MCC) between the models.

**RQ2 - Can we train a model that is transferable between institutions without relying on samples from the target institution?** We compare a Transformer-based PubMedBERT model with the SVM and CNN models on the abnormality classification task, in a cross-institutional setting, previously reported in Koopman et al. (2015) and Hassanzadeh et al. (2018b), respectively. We selected PubMedBERT since it achieved slightly higher, but not significantly better, performance across all four metrics as a result of answering RQ1. To compare PubMedBERT with previously proposed methods (Koopman et al., 2015; Hassanzadeh et al., 2018b), we trained PubMedBERT models on data from two out of three hospitals and tested them in the remaining one. In other words, we considered the three fine-tuning/testing splits, namely 1) fine-tuning on RBWH + RCH, testing on GCH, 2) fine-tuning on RBWH + GCH, testing on RCH, 3) fine-tuning on RCH + GCH, testing on RBWH. PubMedBERT was fine-tuned for ten epochs, with the learning rate of 9e-6 and the seed value of 112, to keep it consistent with the experimental set-up in RQ1. We compared F1 scores between PubMedBERT in the current

[2] https://microsoft.github.io/BLURB/
[3] https://www.ncbi.nlm.nih.gov/pmc/

study, and SVM and CNN models from previous studies (Koopman et al., 2015; Hassanzadeh et al., 2018b).

**RQ3 - What is the impact of using ChatGPT as an additional data source of synthetic data for classification model fine-tuning?** We investigate the benefits of including synthetic reports generated by ChatGPT while fine-tuning the PubMedBERT on the radiology report abnormality classification task. We fine-tuned six PubMedBERT models on three datasets (RBWH, RCH and GCH) separately with and without synthetic reports generated by ChatGPT. The model fine-tuning was performed in consistence with the experimental setup of RQ1 and RQ2, where each model was fine-tunned for ten epochs, with the learning rate of 9e-6 and the seed value of 112. We evaluated each fine-tuned model on the remaining two real datasets (e.g., the model trained on RBWH we evaluated on RCH and GCH datasets). The model evaluation consists of an F1 score and a confusion matrix, including the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) per fine-tuned model.

**RQ4 - How does adding ChatGPT synthesised radiology reports to fine-tuning dataset impact the downstream reconciliation task?** To assess the impact of ChatGPT-generated reports (used in fine-tuning) on patient data reconciliation, the radiology report classification results of both models – PubMedBERT fine-tuned with and without ChatGPT-generated reports (RQ3) – were cross checked with the patient's ICD-10 discharge diagnosis of the ED. Since some of the ICD-10 codes were unavailable or missing from the data received from the ED, we performed patient data reconciliation on available 1429/1480 RBWH, 495/498 RCH and 329/400 GCH records. Following the experimental design used to address RQ3, we evaluated two groups of PubMedBERT models, fine-tuned on records from a single hospital with and without ChatGPT-generated reports, on the downstream task of automatic reconciliation of radiology reports and discharge diagnoses. The evaluation of these two fine-tuned PubMedBERT model groups was performed on the datasets from the remaining two hospitals. Based on the classification results, there were four possible combinations of the radiology report classification / ED discharge diagnosis results: 1) *Both Abnormal*; 2) *Both Normal*; 3) *Radiology Abnormal but ED Normal*; and 4) *Radiology Normal but ED Abnormal*.

| Datasets | Methods | RBWH | RCH | GCH |
|---|---|---|---|---|
| RBWH + RCH | SVM | - | - | 0.84 |
| | CNN | - | - | 0.9294 |
| | PubMedBERT | - | - | **0.9416** |
| RBWH + GCH | SVM | - | 0.88 | - |
| | CNN | - | 0.9367 | - |
| | PubMedBERT | - | **0.944** | - |
| GCH + RCH | SVM | 0.80 | - | - |
| | CNN | 0.9085 | - | - |
| | PubMedBERT | **0.9086** | - | - |

Table 2: Results (F1 scores) for a transferred SVM, CNN without transfer learning, and PubMedBERT trained on multiple sources and evaluated on a different target source. Bold numbers represent the highest F score for each target test set.

## 4 Experiments and Results

**RQ1 - What is the impact of pre-training on the abnormal radiology report classification task?** Figure 1 shows the fine-tuned means and standard deviations for F-score, precision, recall and MCC across 5-folds for each of the six pre-trained BERT-based models. Both BioELECTRA models, the base and the large models were excluded from the comparison since the models did not converge and always predicted the same (abnormal) class. Figure 1 shows that the PubMedBERT model achieves the highest performance across all four metrics (F1-score, Precision, Recall and MCC). To determine the significance of the difference in performance between models, we calculated two-sided 95% Wilson confidence intervals (Figure 1 - right). Models with confidence intervals that do not overlap are regarded significantly different at $p < 0.05$. Overlapping of the Wilson confidence intervals suggests that the performances of PubMedBERT, BERT, BioClinicalBERT, BlueBERT-base and LinkBERT (base and large) were not significantly different from each other; however, all those models performed significantly better than the BlueBERT-large model.

**RQ2 - Can we train a model that is transferable between institutions without relying on samples from the target institution?** The results of the abnormal report classification performance achieved by PubMedBERT models and their comparison with the earlier reported performance of the SVM and CNN models are presented in Table 2. PubMedBERT achieved comparable or higher F1-score compared to SVM and CNN in all three cross-institution fine-tuning/testing splits. In the case of the data split, where the models were fine-tuned on RBWH + RCH and tested on GCH,

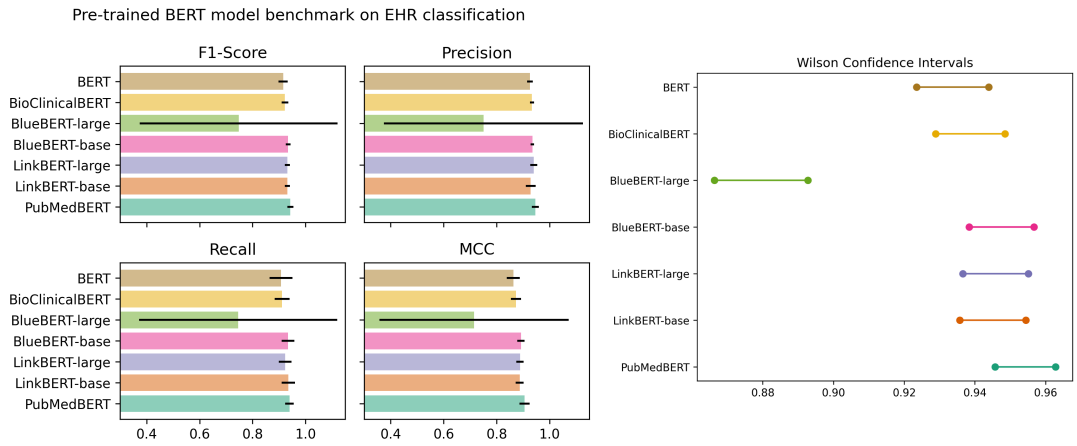


Figure 1: F1-score, Precision, Recall and Matthew's Correlation Coefficient (MCC) computed over 5-fold cross-validation (mean and standard deviation) of six pre-trained BERT models fine-tuned on a mixed dataset (RBWH, RCH, GCH). Two-sided, 95% Wilson confidence intervals for each model.

| | | F1-score | | | TP | | | TN | | | FP | | | FN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RBWH | RCH | GCH | RBWH | RCH | GCH | RBWH | RCH | GCH | RBWH | RCH | GCH | RBWH | RCH | GCH |
| RBWH | No ChatGPT | - | 0.9477 | **0.9431** | - | 154 | 141 | - | **327** | **242** | - | **4** | **7** | - | 13 | 10 |
| | ChatGPT | - | **0.9619** | 0.9255 | - | **164** | **149** | - | 321 | 227 | - | 10 | 22 | - | **3** | **2** |
| RCH | No ChatGPT | 0.8769 | - | 0.9037 | 520 | - | 136 | 814 | - | **235** | 48 | - | **14** | 98 | - | 15 |
| | ChatGPT | **0.9016** | - | **0.9201** | **545** | - | **144** | **816** | - | 231 | **46** | - | 18 | **73** | - | **7** |
| GCH | No ChatGPT | **0.8835** | **0.9358** | - | 508 | 153 | - | **838** | **324** | - | **24** | **7** | - | 110 | 14 | - |
| | ChatGPT | 0.8627 | 0.8595 | - | **550** | **159** | - | 755 | 287 | - | 107 | 44 | - | **68** | **8** | - |

Table 3: Confusion matrix computed for testing cases of PubMedBERT fine-tuned on a dataset containing radiology reports from RBWH, RCH, and GCH, with and without synthetic 100 radiology reports generated by ChatGPT. The models are evaluated on the corresponding two remaining hospital radiology reports datasets by computing F1-Score, the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

PubMedBert achieved a 1.3% F1-score increase compared to CNN and a 12% F1-score increase compared to SVM. When fine-tuned on RBWH + GCH and tested on RCH, PubMedBERT achieved a 0.8% F1-score increase compared to CNN and a 7% F1-score increase compared to SVM. When the models under test were fine-tuned on GCH + RCH and tested on RBWH, PubMedBERT was similar to CNN but obtained a 14% increase in F1 compared to SVM. The F1 scores achieved by PubMedBERT follow the same trend as SVM and CNN, where higher F1 were achieved in training scenarios where the train set involved substantially more samples than the test set (e.g., fine-tuning on RBWH+RCH and testing on GCH). Overall, according to the obtained results presented in Table 2, PubMedBERT generalises better in the cross-institutional setting than previously proposed SVM and CNN-based models.

**RQ3 - What is the impact of using ChatGPT as an additional data source of synthetic data for classification model fine-tuning?** The evaluation results (F1 score, TP, TN, FP and FN) of PubMedBERT fine-tuned with and without ChatGPT-generated reports are detailed in Table 3. Compared with models without ChatGPT data, those fine-tuned with ChatGPT data resulted in more true positives (reports for which both the ML-classifier and the expert labeler indicated an abnormality was present) but also more false positives (reports for which the ML-classifier indicated an abnormality when an abnormality was not present). Conversely, the models fine-tuned with ChatGPT data resulted in fewer true negatives and fewer false negatives compared with models without ChatGPT data. This pattern appeared in all training scenarios except when the model was trained on reports from RCH and tested on real reports from RBWH, whereby the models with ChatGPT data resulted in more true negatives and fewer false positives.

Although these trade-offs do not manifest as a clear improvement in metrics such as the F1-score (Table 3), the observed trade-off trend has important implications on the downstream task considered here of automated abnormality classification from radiology reports. The role of an ML-based classifier in practice would be to automatically shortlist or highlight all reports that indicate the

presence of a radiological abnormality. This would allow ED clinicians to focus on the "abnormal" reports and conduct a more efficient reconciliation process. A model that generates high numbers of true positives and true negatives, while keeping the number of false negatives (potential missed abnormalities) and false positives low is desirable, and our output is consistent with this. Despite relatively low numbers of false positives and false negatives, the high true positive and true negative cases could help to significantly reduce the manual report reconciliation burden on ED clinicians. According to Table 3, fine-tuning PubMedBERT on real reports plus ChatGPT-generated synthetic leads to much lower number of FN than using real reports alone. For example, we saw a 25.5% reduction in FN when *training on RCH and testing on RBWH* and a 76. 9% reduction when *training on RBWH and testing on RCH*.

**RQ4 - How does adding ChatGPT synthesised radiology reports to fine-tuning dataset impact the downstream reconciliation task?** The results obtained, detailed in Table 4, suggest the same trend observed when answering RQ3. Table 4 reveals the trade-off between the ability of the models to reconcile discharge diagnosis with greater disagreement between abnormal radiology classification outcome and normal ED discharge diagnosis (PubMedBERT fine-tuned with ChatGPT-generated reports); or normal radiology classification outcome and abnormal ED discharge diagnosis (PubMedBERT fine-tuned without ChatGPT-generated reports). The consequences of the reconciliation disagreement between these two model groups impact patients in the retrospective review process of the ED differently. The automatic classification outcomes from models fine-tuned with real radiology report only result in a lower number of reports that require manual processing by a clinician but a higher number of misdiagnosed discharged patients. In contrast, automatic classification results from models fine-tuned with real and ChatGPT-generated reports result in a higher number of radiology reports that require manual processing by a clinician and a lower number of misdiagnosed discharged patients. On average, across the six testing scenarios, for a 48.38% higher number of reconciliation disagreements between the abnormal radiology model classification outcome and normal ED discharge diagnosis (requiring manual review), the number of actual misdiagnosed reconciliation cases is 15.35% lower. This implies a lower number of disagreements between normal radiology model classification outcome and abnormal ED diagnosis. Since the severity and cost of misdiagnosis in undiscovered patients can be higher than the cost of a manual retrospective review of radiology reports, PubMedBERT models fine-tuned on the combination of real and ChatGPT-generated reports achieve higher performance than PubMedBERT models fine-tuned on real reports only.

## 5 Discussion and Conclusion

We determined that PubMedBERT was the best-performing of six pre-trained BERT-like models for classifying free-text radiology reports of X-rays for suspected limb fractures in ED patients. Compared to SVM and CNN models, PubMedBERT had better performance (measured by F1-score) for classifying radiology reports when training data and testing data were from different hospitals, suggesting that PubMedBERT has better transferability in cross-institution settings, especially in a low-data regime where the data from the target hospital is unavailable.

We also found that PubMedBERT models, which included some ChatGPT-generated synthetic radiology reports in fine-tuning, resulted in higher numbers of true positives and false positives and lower numbers of true negatives and false negatives than models without synthetic reports. The trade-off in detecting more true positives, using the model enhanced by ChatGPT data, is that there were also more false positives. While this implies that more patients with misdiagnoses would be identified, it also increases the number of reports that must be manually reconciled. This is an important observation in the reconciliation process since the higher number of FPs has less severe consequences on reconciliation than the higher number of FNs. This is because every FP-classified radiology report would require manual clarification, and every FN-classified report stands for a misdiagnosed case. Nevertheless, if all radiology reports are requried to be reviewed, as is done in current practice, our approach to reconciliation can allow patient cases to be prioritised for clinical follow-up such that suspected misdiagnosed cases would be prioritised for manual review.

To address the issue of data imbalance, it is common in the literature to perform over- or under-sampling when developing prediction models (Hassanzadeh et al., 2014; van den Goorbergh et al.,

| | | Both Abnormal | | | Both Normal | | | Radiology Abnormal, ED Normal | | | Radiology Normal, ED Abnormal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RBWH | RCH | GCH | RBWH | RCH | GCH | RBWH | RCH | GCH | RBWH | RCH | GCH |
| RBWH | No ChatGPT | - | 126 | 109 | - | **302** | **148** | - | **29** | **20** | - | 38 | 52 |
| | ChatGPT | - | **129** | **118** | - | 289 | 140 | - | 42 | 28 | - | **35** | **43** |
| RCH | No ChatGPT | 357 | - | 109 | **806** | - | **148** | **183** | - | **20** | 83 | - | 52 |
| | ChatGPT | **366** | - | **115** | 792 | - | 144 | 197 | - | 24 | **74** | - | **46** |
| GCH | No ChatGPT | 347 | 126 | - | **831** | **299** | - | **158** | **32** | - | 93 | 38 | - |
| | ChatGPT | **356** | **134** | - | 718 | 265 | - | 271 | 66 | - | **84** | **30** | - |

Table 4: Reconciliation results encapsulate the agreement between ED discharge diagnosis and radiology report classification model results, where the agreement between the two falls into one of the four categories: 1) Both Abnormal, 2) Both Normal, 3) Radiology Abnormal, ED Normal, and 4) Radiology Normal, ED Abnormal. Two radiology report classification models were compared, the radiology report classifier where ChatGPT-generated reports were and were not used in fine-tuning. The bold numbers represent the better performing model based on the reconciliation outcome.

2022). Using synthetic data generated from ChatGPT can be viewed as another approach to augment modeling by changing the data distribution. We demonstrate that using synthetic data reduces the number of unwanted predictions, such as false negatives. This shows that augmenting with ChatGPT has a similar effect to balancing the data distribution by increasing the sample size of rare classes, in this case the abnormal diagnoses.

Overall, when developing a solution for automated reconciliation of radiology reports and discharge diagnoses in a setting where labelled radiology reports from the target institution are unavailable, pre-trained transformer models such as PubMedBERT fine-tuned on available labelled reports from partner institutions, together with ChatGPT-synthesised radiology reports can boost the automatic reconciliation performance. As we showed the promise of using NLP models to facilitate diagnosis reconciliation for ED clinicians, more works may investigate similar approaches to streamline the manual review process, flag mismatches, and explore workflow integration.

## Acknowledgements

We thank the anonymised reviewers for their thoughtful comments to help improve the manuscript during the review process.

## References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*, pages 72–78.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are Few-Shot learners. In *NeurIPS*, volume 33, pages 1877–1901.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for text data augmentation.

Berry de Bruijn, Ann Cranney, Siobhan O'Donnell, Joel D Martin, and Alan J Forster. 2006. Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports. *Journal of the American Medical Informatics Association: JAMIA*, 13(6):696–698.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of NAACL*, pages 4171–4186.

Luka Gligic, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2020. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *Neural networks: the official journal of the International Neural Network Society*, 121:132–139.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1):1–23.

Hamed Hassanzadeh, Tudor Groza, Anthony Nguyen, and Jane Hunter. 2014. Load balancing for imbalanced data sets: Classifying scientific artefacts for evidence based medicine. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 972–984. Springer International Publishing.

Hamed Hassanzadeh, Mahnoosh Kholghi, Anthony Nguyen, and Kevin Chu. 2018a. Clinical document classification using labeled and unlabeled data across hospitals. *AMIA Annual Symposium proceedings*, 2018:545–554.

Hamed Hassanzadeh, Anthony Nguyen, Sarvnaz Karimi, and Kevin Chu. 2018b. Transferability of artificial neural networks for clinical document classification across hospitals: A case study on abnormality detection from radiology reports. *Journal of biomedical informatics*, 85:68–79.

Keliang Jia. 2022. Sentiment classification of microblog: A framework based on BERT and CNN with attention mechanism. *Computers & Electrical Engineering*, 101:108032.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Bevan Koopman, Guido Zuccon, Amol Wagholikar, Kevin Chu, John O'Dwyer, Anthony Nguyen, and Gerben Keijzers. 2015. Automated reconciliation of radiology reports and discharge summaries. *AMIA Annual Symposium proceedings*, 2015:775–784.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R Andrew Taylor, Harlan M Krumholz, and Dragomir Radev. 2022. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511.

Jin Li, Yu Tian, Runze Li, Tianshu Zhou, Jun Li, Kefeng Ding, and Jingsong Li. 2021a. Improving prediction for medical institution with limited patient data: Leveraging hospital-specific data based on multicenter collaborative research network. *Artificial intelligence in medicine*, 113:102024.

Pengfei Li, Peixiang Zhong, Kezhi Mao, Dongzhe Wang, Xuefeng Yang, Yunfeng Liu, Jianxiong Yin, and Simon See. 2021b. ACT: an attentive convolutional transformer for efficient text classification. *Proceedings of 2021 AAAI Conference*, 35(15):13261–13269.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Aaron J Masino, Robert W Grundmeier, Jeffrey W Pennington, John A Germiller, and E Bryan Crenshaw, 3rd. 2016. Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. *BMC medical informatics and decision making*, 16:65.

Sajan B Patel and Kyle Lam. 2023. ChatGPT: the future of discharge summaries? *The Lancet. Digital health*, 5(3):e107–e108.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *BioNLP Workshop and Shared Task*, pages 58–65.

Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association: JAMIA*, 29(9):1525–1534.

Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1419–1428.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Conference of ACL*, pages 8003–8016, Dublin, Ireland.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: transformers for longer sequences. In *NeurIPS*, 1450, pages 17283–17297.

Yihua Zhou, Per K Amundson, Fang Yu, Marcus M Kessler, Tammie L S Benzinger, and Franz J Wippold. 2014. Automated classification of radiology reports to facilitate retrospective study in radiology. *Journal of digital imaging*, 27(6):730–736.

Guido Zuccon, Amol S Wagholikar, Anthony N Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. 2013. Automatic classification of Free-Text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology. *AMIA Joint Summits on Translational Science proceedings.*, 2013:300–304.

# A Appendix Tables

| Initial prompts | Abnormal case | Normal case |
|---|---|---|
| 1 | "Write an example of a limb x-ray radiology report with an abnormality." | "Write an example of a limb x-ray radiology report without abnormalities." |
| 2 | "Write an example of a limb x-ray radiology report with several abnormalities." | "Write an example of a normal limb x-ray radiology report." |
| 3 | "Write an example of a limb x-ray radiology reports with max 12 abnormalities." | "Write an example of a limb x-ray radiology reports with max 12 normal observations." |
| 4 | "Write an example of a limb x-ray radiology report with an abnormality, use lowercase abbreviations with no explanation, and no full stop after an abbreviation." | "Write an example of a normal limb x-ray radiology report, use lowercase abbreviations with no explanation, and no full stop after an abbreviation." |
| **Auxiliary prompts** | | |
| 1 | "Give me another example." | |
| 2 | "Give me another example with more clinical detail." | |
| 3 | "Give me another example with more specific details." | |
| 4 | "Give me another example with more specific details, but less repetitive." | |
| 5 | "Give me another example. Use abbreviations without explanation." | |

Table 5: Prompts ChatGPT was presented to obtain synthetic radiology report examples used for training. The auxiliary prompts were used to gather more diverse synthetic samples.

| Pre-trained BERT model | Corpora |
|---|---|
| BERT | 3,300 million words from BooksCorpus and English Wikipedia |
| PubMedBERT | PubMedCentral abstracts and full-text articles |
| LinkBERT (base and large) | A large corpus of documents and their links (e.g., hyperlinks, citation links) |
| BlueBERT (base and large) | PubMed texts (about 4000 million words) |
| BioClinicalBERT | All notes from MIMIC III |
| BioELECTRA (base and large) | PubMed abstracts only with biomedical domain vocabulary |

Table 6: Pre-trained BERT models and training corpora.