The University of Melbourne - School of Computing and Information Systems

COMP90089 - Machine Learning Applications in Health

Enhancing Chest X-ray Diagnosis for Respiratory Diseases using Transfer Learning

Team 13
Tuan Khoi Nguyen (1025294)
Clemence Mottez (1486585)
Xiaochen Hou (1067898)
Ivy Liang (1397138)
Qichi Liang (1392005)

(Word count: 2062)

Abstract

Chest X-ray (CXR) is common for detecting Chronic Respiratory Diseases (CRD). Deep learning has been prevalent in CXRs, but has questionable reliability due to ambiguous mechanisms. Electronic health records (EHR) contain useful information, yet unusable due to sparsity. This study proposes the combination of CXR with EHR using Transfer Learning to improve diagnosis for 8 CRDs in both performance and interpretability. We used DenseNet architecture to extract CXRs into 18 latent features, and combine them with 12 EHR features that are crucial in CRD diagnosis, which will be fed into 2 models: XGBoost and Gaussian Naive Bayes. Our approach is validated by a comprehensive evaluation framework, emphasising accuracy, precision and recall to address data imbalance. We found that both models improved positive diagnosis predictions, and inspecting XGBoost shows that the model's decision-making process closely matches human experts. (138 words)

1 Introduction (304 words)

1.1 Background & Related Works

Chronic Respiratory Diseases (CRD) is a group of disorders that severely affect airway organs [1], which is responsible for at least 5% of global morbidity and mortality [1, 2]. The associated risk of CRDs, however, can be minimised with early detection [3].

Chest X-rays (CXR) are common and non-invasive for diagnosing CRDs by assessing any abnormalities such as white spots [4]. However, this method requires expert knowledge given the complicated definitions of anomalies [5]. Recently, using Machine Learning (ML), specifically, using Deep Learning (DL) for CXR has been a common technology in detecting CRD with high or even human-level accuracy [6, 7]. While these techniques have made substantial progress in improving diagnosis speed [8] and reducing human error [4], sole dependency on models has caused inherent challenges as how DL models make its predictions are unknown, hindering any underlying bias [9] or misconceptions [10, 8]. Most recent attempts at improving DL explainability, such as heatmaps [11], still cannot fully explain a model's rationale [9]. Data scarcity is also a common problem, as proportion of positive CRD diagnoses is very small, leading to severe data imbalance where ML models cannot learn sufficient information [12, 13].

Electronic Health Records (EHR) are patient medical information stored electronically, with the aim of improving patient care and building predictive tools [14]. However, they suffer from frequent missing information [15, 16], which is unusable in DL models as they strictly cannot work on missing data [17]. Imputation is a common approach, but with the sparse nature of EHR, it cannot guarantee the correctness of the data [18, 19].

1.2 Research Aim

In this work, we propose a multi-faceted predictive system that uses the concept of Transfer Learning (TL) [20] to infuse EHR data into the ML-based CRD diagnosis using CXR (Fig.1). The system is expected to maintain the highly accurate nature of DL and the informativeness of EHR and classical ML models. Rigorous testing and validation will be carried out in 2 stages: the first stage will evaluate the performance of the proposed system, and the second stage will delve deeper into investigating the model's decision-making process. Through this study, we expect to answer the following research question: Does incorporating EHR data into CXR diagnosis for CRD help in improving the ML model's performance, transparency and trustworthiness in its decision-making processes?

2 Methods (961 words)

2.1 Design Choices

2.1.1 Data Sourcing

The work in this report uses 2 sources of dataset: The Medical Information Mart for Intensive Care (MIMIC) and CheXpert.

MIMIC-IV MIMIC-IV [21, 22] holds 299,712 detailed records of patient visits at Beth Israel Deaconess Medical Center (BIDMC). This database is used for extracting patient EHR, as well as the ground truth of this task: the International Classification of Diseases (ICD) code that contains the disease detected on patient discharge.

MIMIC-CXR MIMIC-CXR [23] features 227,835 imaging studies de-identified for 64,588 patients in BIDMC and MIMIC-IV. This database is used to retrieve CXR images of patient.

CheXpert CheXpert [24] is another CXR dataset used in this project. It contains 224,316 CXRs of 65,240 Stanford Hospital patients, independent from BIDMC. In this project, the dataset is used to pre-train the chosen DL model to perform TL on MIMIC CXR data. This ensures the generalisation nature of real-world data, where new instances are not known to the trained model.

Data access control Acknowledging ethical aspects of medical data, the authors have performed all the necessary training to access the databases. To ensure secure access,

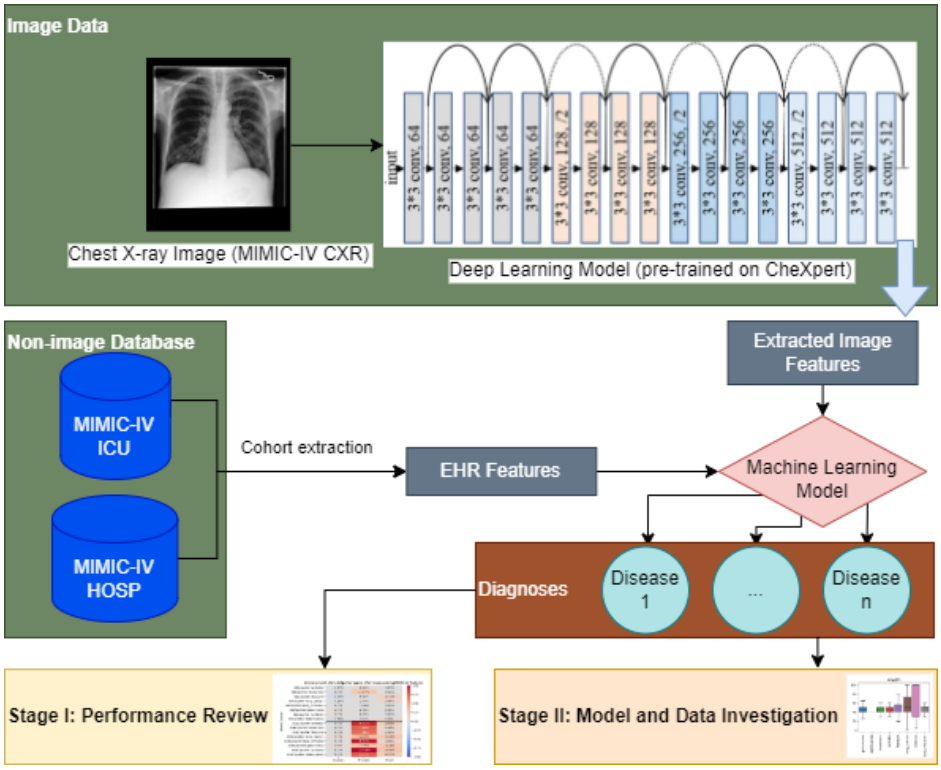

Figure 1: The overall workflow diagram of the proposed system

the data used in this project will not be provided publicly. Programming scripts are provided instead, where logging in is required to confirm one's training status before accessing and generating the datasets.

2.1.2 Data formation

CRDs to predict As many possible CRDs make exhaustiveness infeasible, we selected 8 representative diseases that are expected to have good diversity in characteristics and data distribution [25, 26, 27, 28, 29]. A summary table of the diseases can be found in Table 1A. Patients were determined to have been diagnosed with a CRD based on description of ICD codes assigned during discharge, obtained directly from MIMIC-IV.

EHR features Based on the CRDs chosen, 12 features were chosen to be extracted from MIMIC-IV. Each feature is a key factor in detecting at least one of the 8 diseases chosen, based on diagnosis research related to the chosen CRDs (Table 1C) [25, 26, 27, 28, 29]. For these features, we disregarded all timing information to minimise patient information leakage.

DL image features DenseNet is a deep neural network architecture that can maintain information at various detail levels [30]. This model is chosen to retrieve the most useful details from CXR image. This project will implement CheXpert pre-trained model available on `torchxrayvision` package [31]. The second-last dense layer will be chosen for extracting features, which produces 18 latent image features from each CXR.

2.2 Data collection & organisation

2.2.1 Cohort identification

Patients in MIMIC-IV were excluded if they did not have CXR. The remaining patients will have their ICD codes checked to see if they match 1 of 8 CRDs chosen, which will divide the patients into 2 groups. All patients with matching ICD will be included in the cohort, while the rest, having significantly larger population, will be randomly selected with quantity matching the first group. This process is summarised in Fig.2C. Next, 12 EHR features will be extracted from MIMIC-IV databases, as illustrated in Fig.2A.

Each patient in the selected cohort will then have their most recent CXR before discharge selected. This helps to simplify the model and avoid duplicating data, as one patient can have multiple CXR results in MIMIC [23]. In the final cohort dataset, each patient is expected to have 18 image features and 12 EHR features, totalling 30 features.

	A. Information		B. Statistics		C. Diagnosability using features												
Disease	Characteristics	Keywords (for ICD lookup)	Number of cases in cohort	% of cases in cohort	chest x-ray	diabetes	age	hiv	oxygen	heart rate	temperature	hemoglobin	red blood cell count	white blood cell	smoke	cough	sputum culture
arthritis	Infection, physical symptoms	*arthritis*	1191	4.10%	■		■					■	■	■	■		
bronchitis	Infection, low mortality, physical symptoms	*bronchitis*	1488	5.12%	■					■	■					■	■
fracture	Physical disorder	*fracture, broken*	2074	7.14%	■				■								
lung_cancer	Tumor, physical symptoms, high mortality, common	*tumor, cancer*	7775	26.76%	■		■								■	■	■
lung_infection	Infection, also covers unspecified/unclear diseases	*infection, infectious*	949	3.27%	■			■		■	■	■	■	■			■
pneumonia	Infection, medium mortality, physical symptoms, comn	*pneumonia*	7983	27.48%	■				■		■	■	■	■		■	■
scoliosis	Physical disorder	*scoliosis, curvature*	150	0.52%	■					■							
tuberculosis	Infection, high mortality, low infection rate	*tuberculosis*	770	2.65%	■	■		■									■

Table 1: Description table of chosen CRDs. Green-coloured cell indicates that feature can be used to diagnose the corresponding CRD in literature

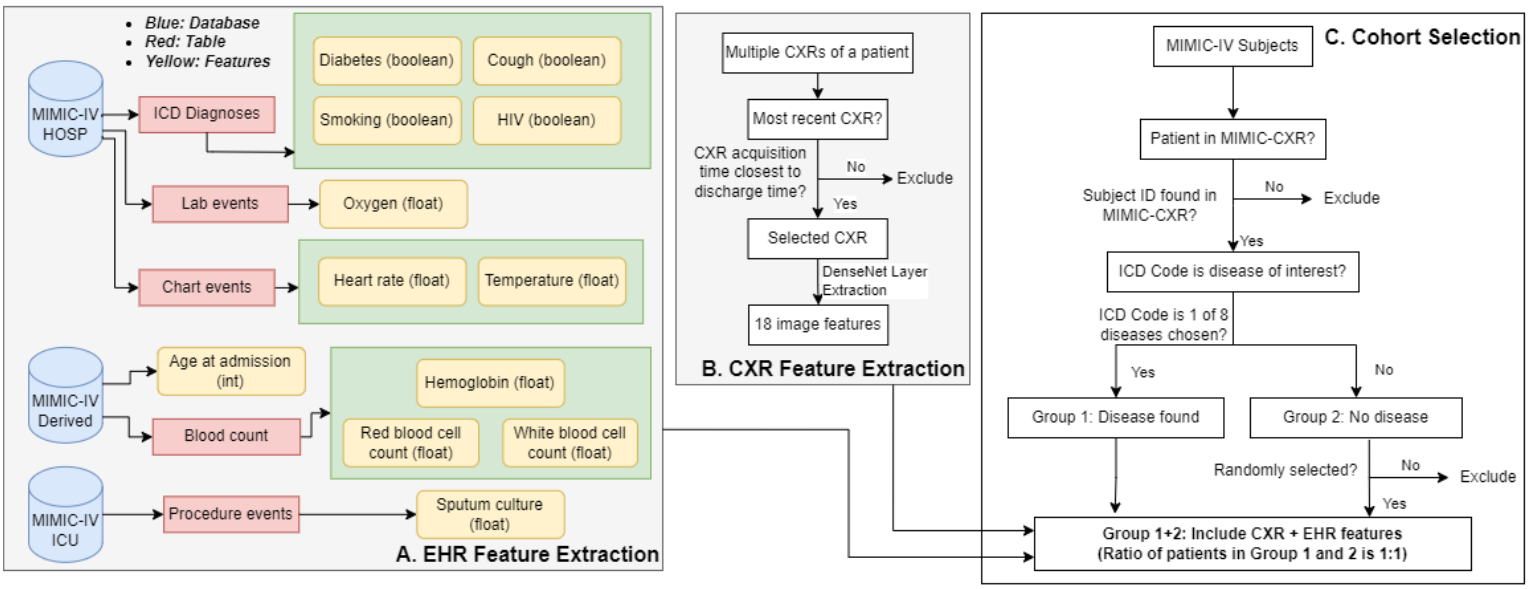

Figure 2: Procedure outline of the cohort extraction process

2.2.2 Data cleaning

While the image features do not require further processing as data protocol and cleaning in MIMIC-CXR is well-handled [23], several irregular measurements were found in EHR data, such as human temperature of 150^oC. Therefore, to ensure data integrity for continuous features, we removed any measurements that are outliers in feature distribution, i.e. more than 3 standard deviations away from the mean.

2.3 Pipeline

The assessment procedure of this project will be divided into 2 stages: Evaluation and Experimentation.

2.3.1 Stage I: Evaluation

The evaluation stage aims to assess the performance of the ML models chosen on the cohort dataset, which helps to clarify whether using TL and EHR data can improve CRD diagnosis performance. In each process, the data is split into train and validation data at 65-35 ratio, which is expected to capture good data diversity for assessment. 5 models will be run concurrently: the original DenseNet model implemented to perform binary classification on each CRD, and each of 2 chosen ML models will perform prediction on data with and without EHR features. For each model-CRD pair, this process will run 5 times to capture sufficient variability.

Models With high data sparsity, we selected models that can skip over missing data without imputation. We also favoured models with good interpretability. As a result, 2 models were chosen: XGBoost and Gaussian Naive Bayes (GNB).

XGBoost [32] combines multiple simple models, which utilises accurate learning by ensembling diversity. XGBoost uses Gain importance metric to measure feature importance during training [32]. Not only improving interpretability, this helps the model to handle well irrelevant features and class imbalance, as well as being able to ignore missing features in prediction.

GNB is a simple ML model that predicts using Bayes theorem and Gaussian distribution. It simplifies the prediction process by assuming feature independence and normal distribution of features, but usually does not perform well as this assumption rarely holds in real-world [33]. This model is chosen as the baseline model to define standards of a good model outcome in this task.

Metrics For model evaluation, we selected accuracy, precision and recall as performance metrics to understand the model's predictive capabilities. These metrics are useful to evaluate our model's capacity to balance precise predictions and reduce minority miss-outs, especially when False Negatives are shown to be less desired due to more severe consequences [34].

Hyperparameter setting GNB does not have hyperparameter, while XGBoost's default hyperparameters have been optimised to best fit with most data cases [32]. Therefore, we decided not to proceed with tuning.

2.3.2 Stage II: Experimentation

To further understand the model mechanism, experimentation stage aims to investigate and describe XGBoost's rationale by reviewing feature importance using its internal parameter Gain. This stage is expected to help answer whether EHR was really considered alongside image features when performing CRD diagnosis.

3 Results (352 words)

3.1 Stage I: Evaluation

Table 2A shows that the accuracy is very high for the models. However, precision and recall (PnR) are almost 0 for all models, suggesting they are likely to have predicted all instances as negative. Comparing with the distributions in Table 1B, it becomes more evident when most of the accuracy for each CRD is actually equal to the proportion of negative diagnosis in the dataset. The only exception to this are the precision results for lung infection and pneumonia - diseases that have less severe class imbalance. This highlights that class imbalance can indeed affect the predictions negatively.

When TL is applied to image data, directly comparing Table 2A and Table 2B shows that recall started to improve for few diseases, even in the simple GNB model. For both models, precision went up minimally, but recall for GNB saw an interesting case where lung cancer and pneumonia went up by approximately 40%.

More significant results are observed when EHR features are incorporated into the data. First, for GNB, the 40% recall improvement that was observed in Table 2B no longer hold in Table 2C. However, compared to the original DL model, GNB models still showed little improvement for most PnR metrics, with most significant recall increase being 9% and precision increase being 36%, and a precision reduction of 10% was observed for lung cancer (Table 2D). XGBoost, on the other hand, saw a great increase in PnR at every CRD, with Table 2E showing precision improvement ranging from 15% to 100%, and recall improvement ranging from 11% to 37% when comparing against the original model.

3.2 Stage II: Experimentation

Fig.3 showed interesting insights on the features of interest for XGBoost via Gain. First, the majority of expert criteria that we have researched in Table 1.C shows a good match with the heatmap, displaying XGBoost's resemblance to human experts in the decision-making process. Some interesting details outside our research are also observed when comparing the differences: cough has significant role in diagnosing arthritis, but not for other CRDs, and diabetes is also significant in detecting pneumonia.

All CRDs show that all image features are considered almost equally, while there are differences in favouring EHR features, which makes sense for the intended inclusion of irrelevant EHR features. Nevertheless, EHR and CXR image features are all considered across the diseases, showing that EHR features help in decision-making, but do not cause over-reliance.

		A. Original			B. TL applied (image features only)			C. TL applied (with EHR features)			D. Performance change C-B			E. Performance change C-A		
Model	**Disease**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Gaussian Naïve Bayes	arthritis	95.9%	0.0%	0.0%	95.9%	0.0%	0.0%	93.5%	2.7%	2.1%	-2.4%	2.7%	2.1%	-2.5%	2.7%	2.1%
	bronchitis	94.8%	0.0%	0.0%	94.8%	0.0%	0.0%	94.2%	31.6%	2.8%	-0.6%	31.6%	2.8%	-0.6%	31.6%	2.8%
	fracture	92.9%	0.0%	0.0%	89.7%	9.4%	4.9%	89.6%	13.8%	8.8%	-0.1%	4.4%	3.9%	-3.3%	13.8%	8.8%
	lung_cancer	73.1%	36.3%	0.2%	59.8%	31.9%	44.1%	71.2%	26.4%	3.7%	11.4%	-5.6%	-40.3%	-1.9%	-9.9%	3.5%
	lung_infection	96.8%	0.0%	0.0%	96.6%	2.2%	0.3%	95.9%	2.1%	0.6%	-0.7%	0.0%	0.3%	-0.9%	2.1%	0.6%
	pneumonia	72.5%	43.9%	0.2%	58.4%	31.7%	44.5%	73.2%	77.4%	3.7%	14.8%	45.7%	-40.8%	0.7%	33.5%	3.5%
	scoliosis	99.5%	0.0%	0.0%	93.0%	0.5%	7.5%	99.0%	0.0%	0.0%	6.0%	-0.5%	-7.5%	-0.5%	0.0%	0.0%
	tuberculosis	97.2%	0.0%	0.0%	94.9%	5.6%	3.6%	96.5%	1.4%	0.4%	1.6%	-4.3%	-3.1%	-0.7%	1.4%	0.4%
XGBoost	arthritis	95.9%	0.0%	0.0%	95.9%	2.9%	0.0%	96.4%	96.0%	14.3%	0.6%	93.1%	14.2%	0.6%	96.0%	14.3%
	bronchitis	94.7%	0.0%	0.0%	94.7%	13.8%	0.3%	95.2%	82.8%	11.1%	0.5%	68.9%	10.8%	0.5%	82.8%	11.1%
	fracture	92.9%	0.0%	0.0%	92.7%	8.9%	0.3%	93.5%	78.3%	11.3%	0.8%	69.4%	11.0%	0.6%	78.3%	11.3%
	lung_cancer	73.1%	27.3%	0.1%	71.5%	36.3%	8.4%	78.0%	68.3%	33.8%	6.6%	32.0%	25.4%	4.9%	41.0%	33.6%
	lung_infection	96.8%	0.0%	0.0%	96.8%	4.0%	0.1%	97.2%	97.1%	12.4%	0.4%	93.1%	12.3%	0.4%	97.1%	12.4%
	pneumonia	72.5%	55.8%	0.1%	70.3%	34.5%	9.0%	78.5%	70.7%	37.3%	8.2%	36.2%	28.4%	6.0%	14.9%	37.2%
	scoliosis	99.4%	0.0%	0.0%	99.4%	0.0%	0.0%	99.5%	100.0%	21.2%	0.1%	100.0%	21.2%	0.1%	100.0%	21.2%
	tuberculosis	97.4%	0.0%	0.0%	97.4%	3.3%	0.1%	97.9%	96.9%	19.0%	0.5%	93.6%	19.0%	0.5%	96.9%	19.0%

Table 2: Performance table of different model-CRD-feature set combinations, averaged over 5 runs (for variability, see Appendix Table A.1 for standard deviation measurements)

Diseases / Features	arthritis	bronchitis	fracture	lung_cancer	lung_infection	pneumonia	scoliosis	tuberculosis
age	0.061	0.054	0.053	0.074	0.055	0.034	0.058	0.048
cough	0.072	0.009	0.020	0.000	0.000	0.011	0.000	0.000
diabete	0.044	0.050	0.055	0.068	0.063	0.086	0.038	0.068
heart_rate	0.043	0.035	0.040	0.037	0.042	0.031	0.111	0.048
hemoglobin	0.054	0.045	0.045	0.054	0.055	0.032	0.061	0.043
hiv	0.057	0.055	0.033	0.060	0.063	0.042	0.035	0.052
oxygen	0.026	0.038	0.042	0.036	0.051	0.111	0.041	0.028
rbc	0.056	0.061	0.055	0.053	0.058	0.034	0.067	0.048
smoke	0.051	0.061	0.046	0.047	0.046	0.042	0.054	0.055
sputum_culture	0.025	0.050	0.035	0.028	0.035	0.153	0.000	0.079
temperature	0.040	0.035	0.055	0.039	0.049	0.043	0.049	0.048
wbc	0.047	0.042	0.046	0.045	0.051	0.047	0.060	0.050
Image feature 0	0.021	0.029	0.030	0.023	0.018	0.016	0.019	0.027
Image feature 1	0.026	0.026	0.027	0.022	0.023	0.018	0.022	0.021
Image feature 2	0.019	0.025	0.024	0.025	0.022	0.017	0.022	0.023
Image feature 3	0.022	0.024	0.026	0.031	0.023	0.023	0.027	0.024
Image feature 4	0.024	0.028	0.023	0.021	0.023	0.021	0.025	0.024
Image feature 5	0.027	0.024	0.027	0.024	0.020	0.020	0.023	0.021
Image feature 6	0.024	0.023	0.026	0.020	0.027	0.017	0.032	0.020
Image feature 7	0.022	0.031	0.023	0.025	0.025	0.019	0.031	0.023
Image feature 8	0.022	0.024	0.026	0.034	0.022	0.023	0.016	0.028
Image feature 9	0.024	0.026	0.023	0.025	0.027	0.016	0.019	0.023
Image feature 10	0.025	0.024	0.028	0.031	0.028	0.019	0.019	0.024
Image feature 11	0.023	0.029	0.025	0.021	0.023	0.016	0.024	0.023
Image feature 12	0.025	0.027	0.029	0.022	0.023	0.016	0.022	0.022
Image feature 13	0.020	0.023	0.027	0.027	0.020	0.018	0.028	0.022
Image feature 14	0.022	0.029	0.026	0.030	0.032	0.020	0.031	0.027
Image feature 15	0.023	0.023	0.028	0.023	0.026	0.018	0.021	0.023
Image feature 16	0.032	0.025	0.033	0.023	0.029	0.020	0.013	0.031
Image feature 17	0.021	0.026	0.027	0.031	0.022	0.018	0.033	0.025

Figure 3: Heatmap displaying degree of linking between features and diseases

4 Discussion (373 words)

4.1 Result remarks

At a glance, from the results, we can see that incorporating EHR features into CXR diagnosis via TL can indeed help improve CRD detection. Evaluation results using XGBoost showed significant improvement in PnR: up to 100% for precision and up to 37% for recall, meaning that it was able to make significantly more True Positive predictions. Even the

simple GNB model also shows observable improvement for many CRDs, showing how crucial additional data can be to the model predictions, even when they are sparse.

From the experimentation stage, we can see that for all diseases, the models considered a good mixture of image features and EHR features. Inspection shows that XGBoost's decision-making procedure favours the same features as human experts, which displays better model reliability. Some insightful differences were observed, hinting that our diagnosis research was not exhaustive, which is expected under limited timing.

With the methodologies aiming to make our problem have best resemblance to real-world scenarios, the results for both stages can prove to be meaningful towards ML application in health. The proposed system not only shows significant improvement in performance and interpretability, but it also employs what is already available to a patient, thus maximising efficiency. These benefits will put a step in addressing the ML problems in trust that are prevalent in literature.

4.2 Limitations & Recommendations

Although the model met human standards, the study design was imperfect. 8 CRDs chosen were not exhaustive of all possible CRDs due to timing constraints, which may affect generalisability. With only 1 CXR chosen per patient, any temporal details such as symptom development would be removed, despite being a crucial detail in CRD diagnostics [35]. We decided not to proceed with more CXRs due to the storage limitation on operating machines, and all-inclusion might result in more data discrepancy, given that earlier CXRs of a diagnosed patient may not show evident disease symptoms [35].

Study evaluation also has limitations, as ground truth solely relies on the ICD code assigned at discharge, which may subject to human error as high as 6.1% [36] despite being considered 'gold standard' [37]. Nevertheless, we believe the system can still be useful for practitioners as an objective tool.

This project was done without availability of experts. Expert knowledge can better guide TL in both performance and interpretability as DL model can be guided towards predicting vital signs in CXR, such as white spots or fractures, rather than arbitrarily zoning features in the images. Furthermore, while this study process was only trained on CheXpert and MIMIC-IV, which were all collected in USA clinics [24, 22], our programming implementation can be replicated to accommodate different medical conventions across the world. With the rapid development and usage of EHR, this can even be further developed into commercialisation software.

5 Conclusion (72 words)

This study, through implementing a comprehensive CRD diagnosis system, can show that incorporating EHR features into DL-based CXR diagnosis can indeed help improve performance and interpretability. Study procedures, while comprehensive, have potential for further quality improvements and expansions. While several limitations can be addressed, the system is a useful diagnosis system that can effectively aid clinicians in detecting CRDs. Further development is encouraged to increase the positive impact of such systems in medical clinics.

CRediT authorship contribution statement

Tuan Khoi Nguyen: Project administration, Conceptualization, Writing – review & editing, Formal analysis. **Clemence Mottez**: Writing – original draft, Methodology, Conceptualization, Data curation, Investigation. **Ivy Liang**: Writing – original draft, Methodology, Conceptualization, Data curation, Formal analysis. **Xiaochen Hou**: Writing – original draft, Software, Investigation. **Qichi Liang**: Writing – original draft, Data curation, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Anthony L. Byrne, Ben J. Marais, Carole D. Mitnick, Leonid Lecca, and Guy B. Marks. Tuberculosis and chronic respiratory disease: a systematic review. *International Journal of Infectious Diseases*, 32:138–146, March 2015.

[2] Christopher J.L. Murray and Alan D. Lopez. Measuring the global burden of disease. *New England Journal of Medicine*, 369(5):448–457, August 2013.

[3] EA Simoes, T Cherian, J Chow, SA Shahid-Salles, R Laxminarayan, and TJ John. Disease control priorities in developing countries, 2006.

[4] Dulani Meedeniya, Hashara Kumarasinghe, Shammi Kolonne, Chamodi Fernando, Isabel De la Torre Díez, and Gonçalo Marques. Chest x-ray analysis empowered with deep learning: A systematic review. *Applied Soft Computing*, 126:109319, September 2022.

[5] B. Van Ginneken, B.M. Ter Haar Romeny, and M.A. Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, 2001.

[6] Adnane Ait Nasser and Moulay A. Akhloufi. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics*, 13(1):159, January 2023.

[7] Prof. Nisha P. Tembhare, Prof. Puneshkumar U. Tembhare, and Prof. Chandrapal U. Chauhan. Chest x-ray analysis using deep learning. *International Journal for Research in Applied Science and Engineering Technology*, 11(1):1441–1447, January 2023.

[8] Catherine M Jones, Quinlan D Buchlak, Luke Oakden-Rayner, Michael Milne, Jarrel Seah, Nazanin Esmaili, and Ben Hachey. Chest radiographs and machine learning – past, present and future. *Journal of Medical Imaging and Radiation Oncology*, 65(5):538–544, June 2021.

[9] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, May 2020.

[10] Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis*, 84:102721, February 2023.

[11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[12] Xuchun Wang, Hao Ren, Jiahui Ren, Wenzhu Song, Yuchao Qiao, Zeping Ren, Ying Zhao, Liqin Linghu, Yu Cui, Zhiyang Zhao, Limin Chen, and Lixia Qiu. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. *Computer Methods and Programs in Biomedicine*, 230:107340, March 2023.

[13] Lorena Álvarez-Rodríguez, Joaquim de Moura, Jorge Novo, and Marcos Ortega. Does imbalance in chest x-ray datasets produce biased deep learning approaches for COVID-19 screening? *BMC Medical Research Methodology*, 22(1), April 2022.

[14] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16(3):371–379, May 2009.

[15] Rolf H. H. Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*, 4(1), July 2020.

[16] Emily Getzen, Lyle Ungar, Danielle Mowery, Xiaoqian Jiang, and Qi Long. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics*, 139:104269, March 2023.

[17] Jie Lin, NianHua Li, Md Ashraful Alam, and Yuqing Ma. Data-driven missing data imputation in cluster monitoring system based on deep neural network. *Applied Intelligence*, 50(3):860–877, October 2019.

[18] Gavin Tsang, Shang-Ming Zhou, and Xianghua Xie. Modeling large sparse data for feature selection: Hospital admission predictions of the dementia patients using primary care electronic health records. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–13, 2021.

[19] Thomas Tsiampalis and Demosthenes Panagiotakos. Methodological issues of the electronic health records' use in the context of epidemiological investigations, in light of missing data: a review of the recent literature. *BMC Medical Research Methodology*, 23(1), August 2023.

[20] Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine Learning*, 90(2):161–189, July 2012.

[21] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2023.

[22] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), January 2023.

[23] 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), December 2019.

[24] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019.

[25] Daniel Aletaha and Josef S. Smolen. Diagnosis and management of rheumatoid arthritis. *JAMA*, 320(13):1360, October 2018.

[26] Guillermo Stegen, Kenneth Jones, and Patricio Kaplan. CRITERIA FOR GUIDANCE IN THE DIAGNOSIS OF TUBERCULOSIS. *Pediatrics*, 43(2):260–263, February 1969.

[27] Joseph A Janicki and Benjamin Alman. Scoliosis: Review of diagnosis and treatment. *Paediatrics & Child Health*, 12(9):771–776, November 2007.

[28] Dawn E. Jaroszewski, Brandon J. Webb, and Kevin O. Leslie. Diagnosis and management of lung infections. *Thoracic Surgery Clinics*, 22(3):301–324, August 2012.

[29] Samuel N. Grief and Julie K. Loza. Guidelines for the evaluation and treatment of pneumonia. *Primary Care: Clinics in Office Practice*, 45(3):485–503, September 2018.

[30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[31] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.

[32] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[33] Feng-Jen Yang. An implementation of naive bayes classifier. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 301–306, 2018.

[34] Robert D. Duval and Leonard Groeneveld. Hidden policies and hypothesis tests: The implications of type II errors for environmental regulation. *American Journal of Political Science*, 31(2):423, May 1987.

[35] Liqa A. Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulmonary Medicine*, 20(1), September 2020.

[36] Nasir Wabe, Ling Li, Robert Lindeman, Jeffrey J. Post, Maria R. Dahm, Julie Li, Johanna I. Westbrook, and Andrew Georgiou. Evaluation of the accuracy of diagnostic coding for influenza compared to laboratory results: the availability of test results before hospital discharge facilitates improved coding accuracy. *BMC Medical Informatics and Decision Making*, 21(1), May 2021.

[37] E. M. Burns, E. Rigby, R. Mamidanna, A. Bottle, P. Aylin, P. Ziprin, and O. D. Faiz. Systematic review of discharge coding accuracy. *Journal of Public Health*, 34(1):138–148, July 2011.

APPENDIX

A Further results

		TL Applied (image features only)			TL Applied (with EHR features)			Performance change (Adding EHR features)			Performance change (TL + Adding EHR features)		
Model	Disease	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Gaussian Naïve Bayes	arthritis	0.3%	0.0%	0.0%	1.8%	1.6%	1.5%	1.6%	1.6%	1.5%	1.6%	1.6%	1.5%
	bronchitis	0.2%	0.0%	0.0%	0.6%	23.1%	2.0%	0.6%	23.1%	2.0%	0.6%	23.1%	2.0%
	fracture	1.6%	1.1%	1.9%	0.6%	0.9%	1.7%	1.6%	0.6%	2.9%	0.5%	0.9%	1.7%
	lung_cancer	0.5%	0.5%	1.7%	0.9%	3.9%	1.4%	1.0%	3.5%	2.5%	1.0%	8.1%	1.5%
	lung_infection	0.3%	3.3%	0.4%	0.5%	2.3%	0.6%	0.4%	4.0%	0.7%	0.4%	2.3%	0.6%
	pneumonia	0.3%	0.4%	1.2%	0.3%	9.6%	2.3%	0.4%	9.3%	3.0%	0.4%	23.5%	2.3%
	scoliosis	3.8%	0.2%	5.6%	0.5%	0.0%	0.0%	3.6%	0.2%	5.6%	0.5%	0.0%	0.0%
	tuberculosis	1.8%	3.2%	2.0%	0.7%	2.0%	0.8%	2.4%	3.1%	2.7%	0.8%	2.0%	0.8%
XGBoost	arthritis	0.2%	6.4%	0.1%	0.2%	3.2%	1.9%	0.0%	8.4%	1.9%	0.1%	3.2%	1.9%
	bronchitis	0.1%	14.4%	0.3%	0.1%	5.8%	1.4%	0.1%	16.8%	1.3%	0.1%	5.8%	1.4%
	fracture	0.0%	5.2%	0.2%	0.1%	4.6%	1.7%	0.1%	4.6%	1.7%	0.1%	4.6%	1.7%
	lung_cancer	0.2%	1.7%	0.7%	0.4%	0.8%	1.3%	0.2%	1.2%	1.5%	0.3%	8.6%	1.3%
	lung_infection	0.1%	8.9%	0.1%	0.1%	3.0%	0.9%	0.0%	10.2%	0.8%	0.0%	3.0%	0.9%
	pneumonia	0.4%	0.9%	0.4%	0.4%	0.3%	1.0%	0.2%	0.8%	1.2%	0.1%	20.2%	1.0%
	scoliosis	0.1%	0.0%	0.0%	0.1%	0.0%	3.9%	0.0%	0.0%	3.9%	0.0%	0.0%	3.9%
	tuberculosis	0.2%	7.5%	0.2%	0.2%	2.4%	1.0%	0.1%	7.1%	1.0%	0.1%	2.4%	1.0%

Table A.1: Standard deviation of different model-CRD-feature set combinations over 5 runs

B Supplementary materials

Source code repository For the purpose of version control and maintain, the source code of this project is published to GitHub: `https://github.com/clemence-mottez/mimic_iv/tree/main`

Processed data In obligation with ethics and data access, it is required that you have completed mandatory training to gain access to MIMIC. For any data-related enquiries, please email the Chief Researcher (Khoi Nguyen - `tuankhoi@unimelb.edu.au`) with your proof of training completion.

Full results Data in this report has been curated to make display more explainable to audiences. Unprocessed data along with additional plots are published to our source code repository and can be found here: `https://github.com/clemence-mottez/mimic_iv/tree/main/Results`