Emergency Department Outcome Prediction Using Multi-modal Modeling Method (Group 7)

Yi Wei (yw5091) Y.WEI@NYU.EDU
Yuanpu Cao (yc4954) YUANPU.CAO@NYU.EDU
Sixing Zhou (sz3704) SZ3704@NYU.EDU
Matteo Rossi (mr6744) MR6744@NYU.EDU

1. Introduction

The demand for emergency department (ED) services is increasing globally, particularly during the COVID-19 pandemic. This growing demand has led to ED crowding and delays in care delivery, which further causes increased morbidity and mortality. Currently, a vast amount of clinical data collected from the widespread-used Electronic Health Records (EHRs) allows us to develop prediction models for improving the efficiency and effectiveness of emergency care.

This project aims to build a machine learning model that predicts patient hospital admittance using emergency department (ED) data. EDs represent the largest source of hospital admissions (Hong et al., 2018). This task of predicting patient hospital admittance has been studied many times in the past, and several machine learning models were built on different datasets, including a benchmark on MIMIC-IV data (Xie et al., 2022). However, most of the previous models underutilized certain hard-to-encode variables such as patient medicine reconciliation, and do not include free-text and X-ray information from patients' records in prediction (Hong et al., 2018; Raita et al., 2019). This project, by contrast, utilizes additional information by extracting embeddings from X-Ray images, X-Ray free notes, and medicine reconciliation. This new model proposed in this project outperforms the previous benchmark, especially on the patinets with X-Ray images and notes. The code is public available on GitHub [1].

2. Related Work

We identify various studies that are relevant to our work. (Hong et al., 2018) also attempts to predict hospital admission by utilizing both information collected at triage as well as patient history. It is a binary classification problem where the label is admission or discharge and 972 relevant variables are extracted for each of the 560K patients. It utilizes Logistic Regression, XGBoost, and Deep Neural Networks, and the results highlight the importance of incorporating historical information like outpatient medication, historical labs, or vitals as opposed to only triage information as it provides a considerable boost in performance. A shortcoming of this method is that it does not use free text data from EHRs and NLP techniques for medical notes could significantly boost performance.

On the other hand, (Qin et al., 2021) uses clinical text for the early detection of sepsis. Their model incorporates structured data in the form of patient measurements like vital

1. https://github.com/45628andy/ER-Prediction

signs, laboratory tests, and medications taken as well as textual notes on the patient. The model uses ClinicalBERT, which is pre-trained on clinical notes, together with other specialized NLP models to represent these notes and they are able to improve the standard utility score for sepsis prediction.

We also review the past work involving developing and validating Multi-Modal Models. In (Sellergren, 2022), by using embeddings for the image extracting from MIMIC-CXR, the model require less computation resources for training. It also applies noisy abnormality labels in NLP for building the CXR-specific image classifier. Another attractive work is (Nemati S, 2018)'s interpretable machine learning model for predicting sepsis in the ICU. Based on the free text notes (MIMIC-IV-NOTES), the model created an estimate of future sepsis severity in ICU patients using simple EMR features such as white blood cell count, heart rate, and APACHE 2 score, which has guiding significance for our work.

Finally, (Wu et al., 2022) uses features from both the structured EHRs and chest X-ray imaging data for 30-day mortality prediction from hospital admission for confirmed COVID-19 patients. The multimodal model achieves better performance on three datasets compared to a model trained on single modality (EHR or CXR alone). The authors also stress the importance of fine-tuning on local datasets for generalizability purposes.

3. Methods

3.1. Overview

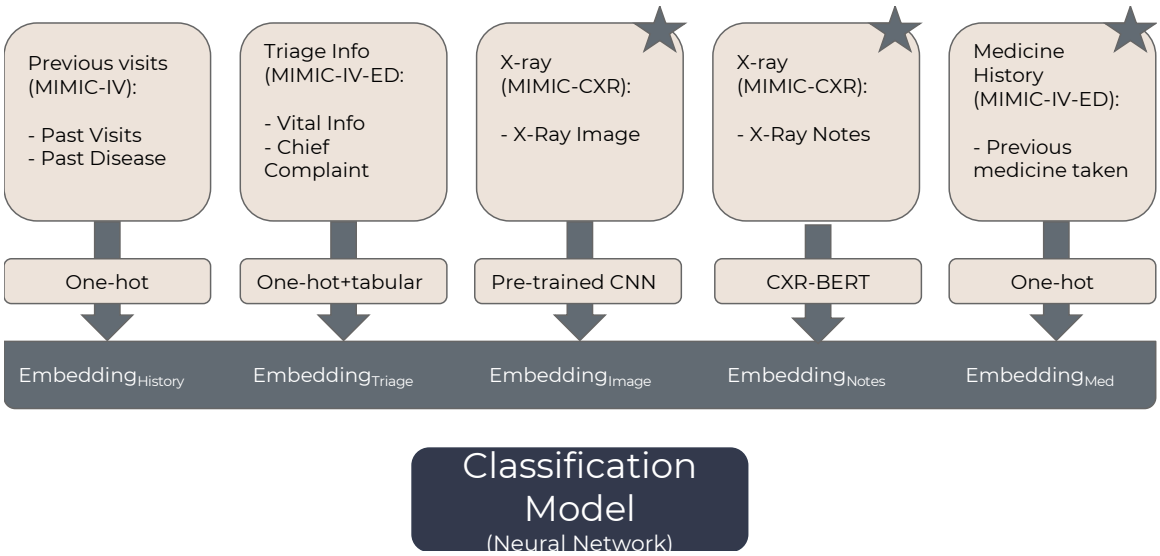


Figure 1: The framework of our multi-modal method.

To summarize, this project first extracts relevant variables from the MIMIC-IV-hosp datasets and mapped previous history and X-Ray results to current visit. Then, each sub-part of the dataset is converted to relevant embedding to be used by the classification model. Numerical variables, such as temperature, are directly used as embedding. Categorical variables, such as admission in the last 60 days, are one-hot-encoded. X-Ray images

are converted to embedding using a CNN model. Free text notes are converted to embedding using a pre-trained language model fine-tuned on this dataset. At last, a multi-layer perceptron model is trained on the embedding to predict whether the patient is admitted to the hospital or discharged after the ED visit. Compared to the baseline, we have added the X-Ray information, which we assume is able to signficantly boost the performance of our model.

3.2. Extract Embeddings from Chest X-ray Reports

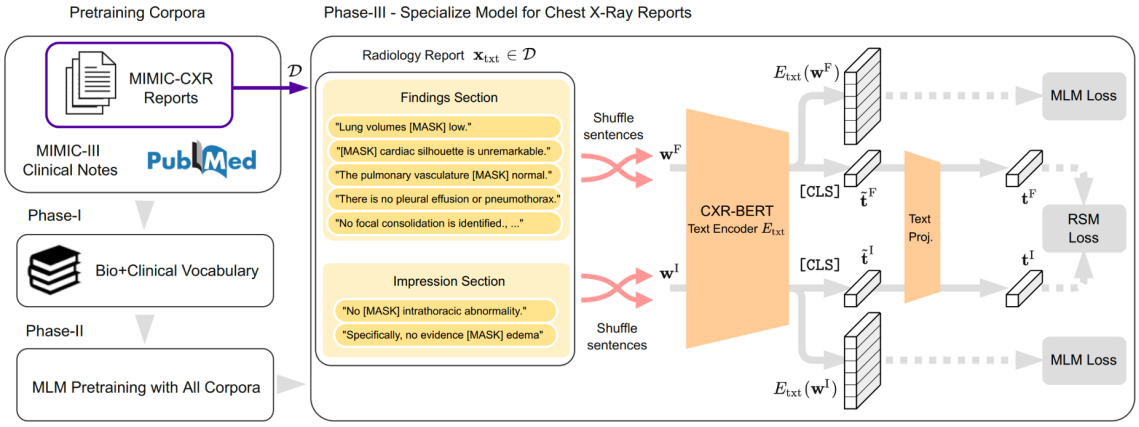


Figure 2: The framework of CXR-BERT (Boecking et al., 2022).

We adopt a pre-trained CXR-BERT (Boecking et al., 2022) as the encoder to extract embeddings from CXR reports. As a specialized Chest X-Ray language model, CXR-BERT exactly fits our scenario. Specifically, the model is pre-trained by three phases as shown in Figure 2: (1) Boecking et al. (2022) constructed a Bio-clinical corpora using the data from PubMed abstracts, MIMIC-III clinical notes, and MIMIC-CXR radiology reports. (2) Following the configurations of RoBERTa (Liu et al., 2019), Boecking et al. (2022) trained a randomly initialized BERT model by Masked Language Modelling (MLM) on the constructed corpora. The MLM loss refers to the cross-entropy for predicting dynamically masked tokens. (3) In order to further specialize the model to the Chest X-Ray domain, Boecking et al. (2022) fine-tuned the CXR-BERT on MIMIC-CXR dataset. Since the Chest X-Ray report contains two sections: *Finding* section and *Impression* section as shown in Figure 3, Boecking et al. (2022) additionally proposed a Radiology Section Matching (RSM) task to make the model match the two sections from the same report. The MLM loss is also used to fine-tune the model, thus the resulting total loss in phase (3) is $\mathcal{L}_{III} = \mathcal{L}_{RSM} + \lambda \mathcal{L}_{MLM}$.

Since we focus on extracting embeddings from Chest X-Ray reports which aligns with the design of the CXR-BERT, we directly adopt the CXR-BERT as the encoder. Specifically, we merge the *Finding* section and *Impression* section for each report and take it as the input of the CXR-BERT to obtain a 128-dimensional embedding.

3.3. Extract Embeddings from Chest X-ray Images

For the Chest X-ray images, we extract 1376 dimensional embeddings for a total of 21,914 instances. For the same image, we evaluate two types of embeddings, which are computed with similar approaches. This allows us to evaluate the quality of the representation and whether these have an impact on the overall classification performance of our multi-modal model. The first type of embedding was extracted from the "Generalized Image Embeddings for the MIMIC Chest X-Ray dataset" Sellergren (2023). These embeddings are generated using the approach described in "Simplified Transfer Learning for Chest Radiography Models Using Less Data" Sellergren (2022). We employed them to train baseline models as they didn't require computing resources and were a fast way to prototype. For the second type of embedding, we were granted access to the pre-trained "CXR network" from Google Health Uddin (2022), which is trained with the transfer learning method mentioned above, and we used the layer before the projection head to generate the embeddings. Before feeding the images to the network we also applied random image transformations. The network is trained in a two-step process. It first uses a backbone network pre-trained on generic natural images and then uses 821,544 chest radiographs from India and the United States Wang X. (2017) in a supervised contrastive learning fashion in which the representation of images with the same label is pulled together and vice versa.

Due to the structure of the datasets, we are unable to link all the data from the ED visits to the medical notes and CXR images. For patients with only one visit, we randomly sample one CXR image for each patient and concatenate it with the other information available for a given patient. However, for patients with multiple hospital visits, hence multiple CXR images, we only considered those who had the same classification label (admitted or discharged) across images, as we assume their representations wouldn't be too distant. However, we made no distinction whether an image was from a frontal or lateral view. Overall, the embeddings were generated with 100+ hours of compute on CPU.

To compare the representation quality of the two embedding types we concatenate them with the embeddings from the medical notes and train two separate ANNss with the same architecture to compare their performance. Table 1 provides a brief overview of the results on relevant metrics.

Table 1: Comparison of the performance between the generalized and generated image embeddings.

Method	Accuracy	Precision	Recall	Specificity
CXR Embs + Notes	0.7295	0.75	0.73	0.9034
CXR Imgs + Notes	0.7746	0.77	0.77	0.8127

Overall, the results suggest that training on our image-generated embeddings can provide a better performance, although the first approach does not classify as many discharged patients as admitted, as the specificity scores show. Further steps to improve the representation quality would be to fine-tune the pre-trained network on the same classification task

we are tackling on similar CXR datasets. Finally, a better approach should be employed to handle both frontal and later CXR views.

4. Dataset and Experimental Setup

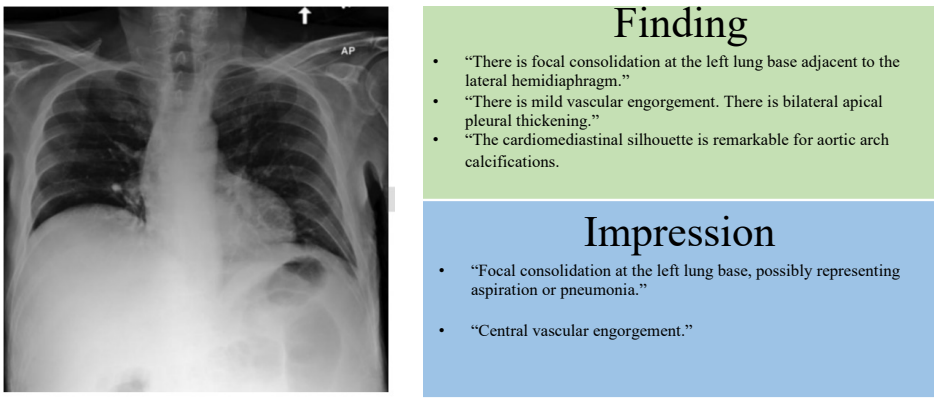
4.1. Dataset

This project mainly uses the publicly available MIMIC-IV Emergency Department (MIMIC-IV-ED) dataset, which contains over 400,000 ED visit episodes from 2011 to 2019. For cohort selection criteria and preprocessing, we follow the benchmarkXie et al. (2022). The benchmark is a series of standard preprocessing methods designed specifically for this dataset and problem. We decide to follow the benchmark to allow others to compare our method and results with others. In addition, we extract past medical history from the MIMIC-IV-Hosp dataset and incorporate chest X-Ray images and text report results from the MIMIC-CXR dataset. Specifically, we consider varying variables as follows:

- **Variables contained in the MIMIC-IV-ED dataset** patient information, demographic, medicine history, chief complaints, and variables collected at triage (e.g. arrival method, temperature, heart rate, resp-rate, o2sat, sbp, dbp, pain)

- **Variables contained in the MIMIC-CXR :** Chest X-ray images, Chest X-Ray reports. We also provide specific examples as shown in Figure 3, where Chest X-Ray report contains two sections: The *Finding* section and the *Impression* section.

- **Variables contained in the X-RAY and Hops dataset:** Past diseases, past hospital/ICU admission, and medical history



(*a*) CXR image (*b*) CXR report

Figure 3: Examples of CXR image and CXR report

After preprocessing raw data. We totally have 418100 ED visits, and we randomly split the ED visits entries into 80% training data and 20% test data. A summary of the data is included in the appendix 5.

4.2. Experimental Setup

Baselines. We compare our method with five baselines: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Emergency Severity Index (ESI), and Multilayer Perceptron (MLP). LR, RF, and GB are traditional machine learning methods. ESI is a traditional clinical scoring-based method a registered nurse assigns based on clinical judgments (Eitel et al., 2003). MLP is a deep-learning method comprised of three dense layers separately followed by non-linear activation. Note that all the previous baselines are designed on tabular data.

Evaluation Metrics. We use the widely adopted AUROC (Area Under the Receiver Operating Characteristic curve) and AUPRC (the Area Under the Receiver Operating Characteristic curve) to measure the prediction performance of our method and baselines. We also report sensitivity and specificity measures under the optimal threshold, defined as the point nearest to the upper-left corner of the ROC curves.

5. Results

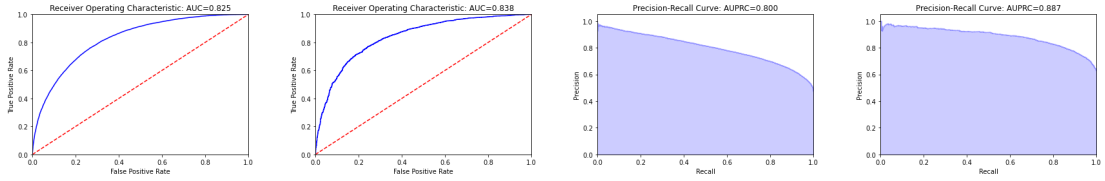
We compare different baselines for emergency department outcome prediction: LR, RF, GB, ESI, MLP with our multi-modal method. Note that no previous methods on this task use the multi-modal solution. First, we present the evaluation results on all patients in Table 2, we can observe that our method slightly outperforms other baselines in terms of AUROC and AUPRC, while the clinical scoring-based ESI underperforms machine learning-based methods. However, partial patients in the dataset don't have all types of data. To fully verify the effectiveness of our method, we additionally compare the best baseline MLP and our method only on patients with all types of data as shown in Table 3. We can observe that the performance in terms of AUROC and AUPRC increase more significantly. Additionally, we also provide the ROC curves and PR curves of our method in Figure 4.

Table 2: Comparison of the performance of different baselines and ours, obtained on all patients.

Method	AUROC	AUPRC	Sensitivity	Specificity
Logistic Regression	0.804	0.765	0.745	0.722
Random Forest	0.819	0.784	0.736	0.750
Gradient Boosting	0.818	0.792	0.748	0.733
Emergency Severity Index	0.709	0.629	0.583	0.780
Multi-Layer Perceptron	0.822	0.796	0.763	0.727
Ours (Multi-Modal)	**0.825**	**0.800**	0.758	0.735

Table 3: Comparison of the performance of MLP and ours, obtained on patients with all types of data.

Method	AUROC	AUPRC	Sensitivity	Specificity
Multi-Layer Perceptron	0.809	0.864	0.750	0.719
Ours (Multi-Modal)	**0.838**	**0.887**	0.771	0.756



(a) (b) (c) (d)

Figure 4: ROC Curve and Precision-Recall Curve of our method. (a) and (c) are obtained from all patients. (b) and (d) are obtained from patients with all data types

6. Fairness

In addition to building the model, we analyze the fairness of emergency department outcome prediction models. Marginalized or racialized groups often lack access to regular healthcare, making emergency department the only form of care available. Therefore, the fairness of clinical outcome prediction model is very important. Moreover, a fairness evaluation supplements previous literature, as in previous literature, the focus is mostly on the model performance overall, not on the fairness and potential bias of the model.

To begin, we define three main metrics that we will focus on: False positive rate, false negative rate, and statistical parity difference. Statistical parity difference refers to the percentage of patients in a protective class that are predicted as needing to be admitted, which we define as a favorable outcome here. Such metrics are different from previous performance metrics, as these fairness metrics focus more on the impact of the model.

To examine the fairness, the optimal threshold for each model (base and our model) is computed as the point closest to the upper left corner of a ROC curve. Then, the three metrics are computed for each racial group. For the purpose of this evaluation, the patients are divided into 4 groups: White (58% of the data), Black (22% of the data), Hispanic (8% of the data), and Asian (4% of the data). The table for our model is shown below, and the table for the baseline model is included in the appendix (6).

In terms of fairness, our model is slightly better as the FNRs are more similar across racial group. However, both the baseline model and the new model do not perform well in terms of fairness, mostly due to the dataset. At the optimal threshold, white patients have the lowest false negative rate, and all other patients have high false negative rate. This is problematic because people of color, who are very vulnerable in the health care system, are getting less care than they need when compared to white patients.

Although mitigation is possible, it is impossible to get a model that has the same FNR and FPR and same statistical parity. If needed, one can set multiple thresholds for different

Table 4: Fairness Evaluation of Our Model (MLP on Multi-Modal Data)

Race	FPR	FNR	Statistical Parity	Base Statistical Parity
White	0.319	0.217	0.567	0.533
Black	0.222	0.284	0.416	0.393
Hispanic	0.185	0.338	0.354	0.355
Asian	0.211	0.284	0.399	0.372
Overall	0.265	0.242	0.498	0.473

groups to achieve similar false positive and negative rates. However, this method is not sufficient for other metrics. Thus, more efforts to address the fairness of this model is needed when using this model to make prediction. Domain knowledge from nurses and doctors on different patients could be helpful to understand the situation in the emergency department that leads to unfairness.

7. Discussion and limitation

In this work, we present a state-of-art model that combine NLP and CV technology to predict hospital admittance based on ED data. We extract and integrate embed dings from tabular data, chest X-ray free-text, and image data. What's more, we use proper evaluation metric for evaluating our model, taking fairness into account. In conclusion, our multi-modal model gain better performance in terms of AUROC and AUPRC compared with the baselines.

However, our model still has some limitations. To begin with, the sample size of our study was relatively small and there exists default values for certain types of data, which may have limited the unreliability of the result. In addition, due to privacy concerns, the study-id in the original data set was masked, and we are only able to use a subset of X-Ray data to improve the performance. At last, we use pre-trained CXR-BERT as text encoder and carry out fine-tuning in our work. This may lead to overfitting or a decrease in model performance when the data is insufficient. Also, it may cause difficulty in hyperparameter tuning, which requires a lot of experimentation and adjustment.

In future works, a more comprehensive dataset that includes multi-modal data that are not censored for privacy is needed to overcome the limitation. Further research directions need to be focused on how to preserve the privacy while retaining as much information as possible for prediction. This can benefit not only this project but data science for healthcare in general. Also, a model that is more interpretable and with domain knowledge encoded in the model can potentially be more useful in practice, as it helps overcome fairness issues mentioned before.

Contribution Statement

The contributions from Yi Wei (yw5091) :

- Purposed this project and conducted preliminary literature review.
- Conducted exploratory data analysis on the datasets.
- Extracted medicine embedding from MIMIV-IV-ED.
- Evaluated the fairness of the models.
- Wrote report (6 Fairness evaluation) and edited other sections

The contributions from Yuanpu Cao (yc4954):

- Evaluated all baselines (LR, RF, GB, ESI, MLP) on all patients and evaluated MLP on patients with multi-modal data.
- Extracted embeddings from MIMIC-CXR report using CXR-BERT.
- Implemented our Multi-modal method incorporating tabular, image, text data and evaluated it on all patients and patients with multi-modal data.
- Wrote report (3.2 Extract Embeddings from Chest X-ray Reports, 4.2 Experimental Setup, 5 Results)

The contributions from Sixing Zhou (sz3704) :

- Evaluated the performance of pre-trained CXR-BERT model with tabular, image and free-text notes embeddings as input.
- Reviewed the past work involving multi-modal model and conduct literature review. Sum up the limitations of work and find suggestions about future improvement.
- Wrote report (2 Related Work, 7 Conclusion)

The contributions from Matteo Rossi (mr6744) :

- Conducted literature review for the ED Outcome Prediction task
- Extracted generalized image embeddings and scraped MIMIC Chest X-Ray JPG dataset (0.5TB+) to generate embeddings
- Built and evaluated ANN network on medical notes and image embeddings for hospital admission/discharge task
- Wrote report (2 Related Work, 3.3 Extract Embeddings from Chest X-ray Images)

References

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 1–21. Springer, 2022.

David R Eitel, Debbie A Travers, Alexander M Rosenau, Nicki Gilboy, and Richard C Wuerz. The emergency severity index triage algorithm version 2 is reliable and valid. *Academic Emergency Medicine*, 10(10):1070–1080, 2003.

Woo Suk Hong, Adrian Daniel Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7): e0201016, 2018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Razmi F Stanley MD Clifford GD Buchman TG. Nemati S, Holder A. An interpretable machine learning model for accurate prediction of sepsis in the icu. 2018.

Fred Qin, Vivek Madan, Ujjwal Ratan, Zohar Karnin, Vishaal Kapoor, Parminder Bhatia, and Taha Kass-Hout. Improving early sepsis prediction with multi modal learning. *arXiv preprint arXiv:2107.11094*, 2021.

Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David FM Brown, Carlos A Camargo, and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Critical care*, 23(1):1–13, 2019.

Chen C. Nabulsi Z. Li Y. Maschinot A. Sarna A. Huang J. Lau C. Kalidindi S. R. Etemadi M. Garcia-Vicente F. Melnick D. Liu Y. Eswaran K. Tse D. Beladia N. Krishnan D. Shetty S. Sellergren, A. B. Simplified transfer learning for chest radiography models using less data. *Radiology. 2022;305(2):454-465. doi:10.1148/radiol.212482*, 2022.

Kiraly A. Pollard T. Weng W. Liu Y. Uddin A. Chen C. Sellergren, A. Generalized image embeddings for the mimic chest x-ray dataset (version 1.0), 2023. URL https://doi.org/10.13026/pxc2-vx69.

Akib Uddin. Simplified transfer learning for chest radiography model development, 2022. URL https://ai.googleblog.com/2022/07/simplified-transfer-learning-for-chest.html.

Lu L Lu Z Bagheri M Summers RM. Wang X., Peng Y. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *IEEE CVPR*, 2017.

Joy Tzung-yu Wu, Miguel Angel Armengol de La Hoz, Po-Chih Kuo, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, Wesley Yeung, Jerry Jurado, Achintya Moulick, Carmelo Milazzo, et al. Developing and validating multi-modal models for mortality prediction in covid-19 patients: A multi-center retrospective study. *Journal of Digital Imaging*, pages 1–16, 2022.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658, 2022.

Appendix

Basic characteristics of the dataset

Table 5: Basic characteristic of the training set and testing set combined. Continuous variables are presented as mean (SD); binary or categorical variables are presented as count (%).

	Overall	Discharge	Hospitalized
Number of Visits	418,100	220,276	197,824
Demographics			
Age	52.83(20.61)	46.33(19.37)	60.07(19.49)
Gender			
Female	227,007(54.3%)	126,755(57.6%)	100,252(50.7%)
Male	191,093(45.7%)	93,251(42.4%)	97,572(49.3%)
Racial Composition			
White	244,093(58.3%)	113,989(51.8%)	130,104(66.0%)
Black	92,168(22.0%)	55,944(25.5%)	36,224(18.2%)
Hispanic	35,205(8.4%)	22,910(10.5%)	12,295(6.0%)
Asian	18,321(4.3%)	12,774(5.8%)	7,124(3.6%)
Other	28,313(6.7%)	11,197(5.0%)	7,583(3.8%)

Additional model fairness evaluation

Table 6: Fairness Evaluation of Baseline Model (LR on tabular data only)

Race	FPR	FNR	Statistical Parity	Base Statistical Parity
White	0.282	0.281	0.525	0.533
Black	0.179	0.382	0.351	0.393
Hispanic	0.155	0.428	0.303	0.355
Asian	0.172	0.382	0.338	0.372
Overall	0.227	0.316	0.443	0.473