iHealth-Chile-3&2 at RRG24: Template Based Report Generation

Oscar Loch [1,2,3], Pablo Messina [1,2,3], Rafael Elberg [1,2,3],
Diego Campanini [2], Álvaro Soto [1,3], René Vidal [4], and Denis Parra [1,2,3]
[1] Department of Computer Science, Pontifical Catholic University of Chile.
[2] Millennium Institute for Intelligent Healthcare Engineering (iHEALTH), Chile.
[3] National Center for Artificial Intelligence (CENIA), Chile.
[4] University of Pennsylvania.
{oscar.loch,pamessina,rafael.elberg}@uc.cl, diego.campanini@ing.uchile.cl,
vidalr@seas.upenn.edu, {asoto,dparra}@ing.puc.cl

Abstract

This paper presents the approaches of the iHealth-Chile-3 and iHealth-Chile-2 teams for the shared task of Large-Scale Radiology Report Generation at the BioNLP workshop. Inspired by prior work on template-based report generation, both teams focused on exploring various template-based strategies, using predictions from multi-label image classifiers as input. Our best approach achieved a modest F1-RadGraph score of 19.42 on the findings hidden test set, ranking 7th on the leaderboard. Notably, we consistently observed a discrepancy between our classification metrics and the F1-CheXbert metric reported on the leaderboard, which always showed lower scores. This suggests that the F1-CheXbert metric may be missing some of the labels mentioned by the templates.

1 Introduction

The generation of radiology reports (RRG) from medical imaging using deep learning represents a significant area of ongoing research (Messina et al., 2022). Successfully implementing this task can help reduce the workload and time spent on administrative duties, such as composing text reports. This efficiency enables physicians to focus more on patient interaction (Topol, 2019) and in identifying anomalies from multiple input images.

There is a pressing need for eXplainable AI (XAI) (Gunning et al., 2019) in critical domains like medicine. In the context of report generation, the explainability aspect remains understudied (Messina et al., 2022). Some models address this issue by generating saliency maps that highlight important pixels, using techniques such as Grad-CAM (Selvaraju et al., 2019) for CNN networks or visualizing attention maps for Transformer networks. However, some authors argue against relying solely on saliency maps as explanations. For instance, Rudin (2019) advocates for using inherently interpretable models that are constrained by domain knowledge, making them transparent and understandable for humans.

To enhance transparency and understandability of our implementation in the Shared task (Xu et al., 2024), we use a simple template-based report generation model. Specifically, we reimplement and modify the template-based strategy proposed by Pino et al. (2021). The team iHealth-Chile-3 focused on meticulously reproducing Pino et al.'s approach, employing DenseNet-121 and a conventional multilabel classification layer for 13 CheXpert classes (excluding "No Findings"), as shown in Figure 2. Meanwhile, team iHealth-Chile-2 developed a different image classifier that combines DenseNet-121 with text embeddings of factual statements, which can be both classified and visually grounded, leveraging very recent work on fact extraction and encoding from radiology reports (Messina et al., 2024). This approach, shown in Figure 3, can be seen as a more general version of stage 1 of CheXfusion (Kim, 2023), the winning method in the ICCV CVAMD 2023 Shared Task on CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays (Holste et al., 2023).

2 Task Description

2.1 Datasets

The data provided by the challenge consists of five datasets: PadChest (Bustos et al., 2020), BIMCV-COVID19 (Vayá et al., 2020), CheXpert (Chambon et al., 2024), OpenI (Demner-Fushman et al., 2016), and MIMIC-CXR (Johnson et al., 2019). Each of these datasets includes radiology reports paired with at least one image. The entire training set comprises $344,394$ reports with at least the Findings section and $366,413$ reports with at least the Impression section. Additionally, the challenge permitted the use of VinDr-CXR (Nguyen et al., 2022), which contains $18,000$ frontal chest X-ray images with labels and bounding box annotations,

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 614–623

but no reports.

In this participation, iHealth-Chile-3 focused on training using only MIMIC-CXR and CheXpert, utilizing the CheXpert labels (Irvin et al., 2019) from both datasets. For training the CNN, this team only used the 13 labels associated with findings (excluding the "No Findings" label) and treated the uncertain label (-1) as negative (0). iHealth-Chile-3 did not employ any additional datasets or data augmentation techniques.

On the other hand, iHealth-Chile-2 leveraged concurrent work on fact extraction and encoding from radiology reports, which includes 591,920 factual statements extracted from MIMIC-CXR radiology reports. A representative subset of these facts was sampled, and with the assistance of a Natural Language Inference (NLI) system (the explanation of which is beyond the scope of this paper), negative facts were identified for all the reports. Furthermore, by combining 78 classes from the Chest ImaGenome dataset (Wu et al., 2021) and the 26 classes from the CXR-LT 2023 challenge (Holste et al., 2023) and removing the overlap, a total of 93 classes were exhaustively annotated by the same NLI system, providing more standardized supervision for MIMIC-CXR. iHealth-Chile-2 also utilized CheXpert, with the 14 classes adapted as short factual statements, VinDr-CXR, with its 28 classes adapted for fact classification, and the 22 bounding box classes used for visual grounding supervision. OpenI was also adapted for fact classification by converting its manual and automatic tags into short sentences with the assistance of GPT-4.

The results reported in this work are measured using the challenge's hidden test set, which contains $1,063$ samples for the generation of the Findings section.

3 Methodology

3.1 Model Architecture

The approaches followed by both teams are summarized in Figure 1. Essentially, an image classifier is trained for multi-label classification. This classifier is then used to make predictions over one or more views, which are processed by a rule-based algorithm to build the final report. Both teams used the PyTorch implementation of DenseNet-121 (Huang et al., 2017) as the visual backbone of their models, outputting 1024-D feature vectors.

The specific implementation by iHealth-Chile-3 is shown in Figure 2. This approach strictly follows Pino et al.'s straightforward implementation (Pino et al., 2021). A fully connected layer predicts 13 classes. For each classified label, there is a pair of fixed sentences: one for when the label is classified as present and another for when it is absent. These sentences are then concatenated to form the final report.

In contrast, iHealth-Chile-2 replaces the fully connected layer with a more sophisticated attention-based pooling mechanism conditioned on a fact embedding, as shown in Figure 3. This approach has the added advantage that the attention can be supervised with ground-truth visual grounding annotations if available, such as bounding boxes in the case of VinDr-CXR. Furthermore, its use of text embeddings to indicate the fact to classify allows the model to work as an *open-vocabulary* multi-label classifier, which can be easily applied to an arbitrary number of datasets with different number of classes or factual statements.

3.2 Training Strategy and Implementation Details

iHealth-Chile-3. This team trained models on MIMIC-CXR and CheXpert using CheXpert labels, selecting the first image in the array of images associated with each medical report, which was generally a frontal view.

To address class imbalance, a Weighted Binary Cross Entropy Loss was employed. The model was optimized using Adam with a learning rate of 0.0001 and a weight decay of 0.00001. Additionally, a learning rate scheduler reduced the learning rate by a factor of 0.1 if the monitored metric did not improve for three consecutive epochs. This dynamic adjustment helps refine the training process and achieve better convergence based on the model's performance. The input images were resized to 256×256 and normalized with a mean and standard deviation of 0.5.

The model was trained for 12 epochs with a batch size of 110, using an NVIDIA RTX A6000 GPU, with an estimated training time of 42 hours.

iHealth-Chile-2. This team utilized the MIMIC-CXR, CheXpert, VinDr-CXR, and OpenI datasets. To ensure a more balanced sampling of all datasets in subsequent batches, a multi-dataset dataloader was implemented. This dataloader sampled from each dataset with a weight of 5.0 for MIMIC-CXR and 1.0 for each of the other datasets, giving

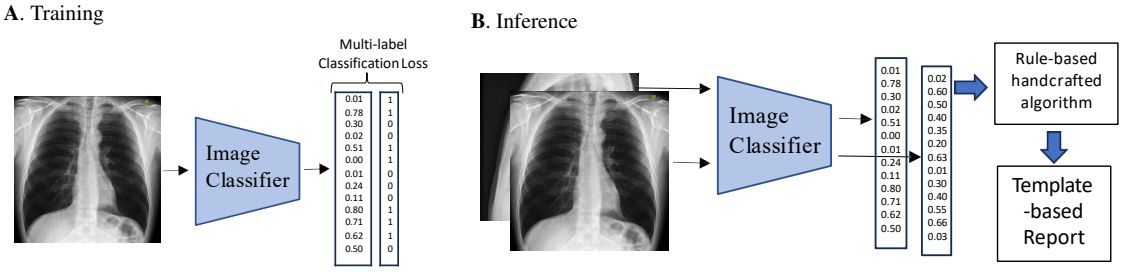

Figure 1: Overview of the template-based approach followed by both teams. During training, a single-view image classifier is trained for multi-label classification. During inference, the image classifier is used to predict labels for one or all the views associated with a given report to generate. These classification predictions are then processed by a handcrafted rule-based algorithm that builds the final report.

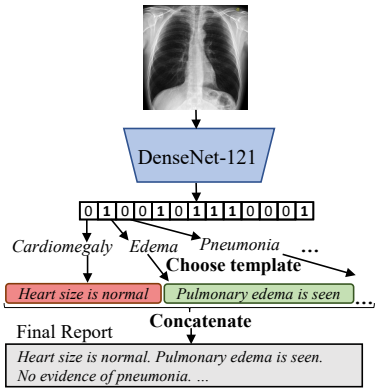

Figure 2: Template-Based Architecture of iHealth-Chile-3. The DenseNet is trained to classify the 13 labels as shown in the scheme. After the training is complete, in inference, the DenseNet is frozen and classifies the 13 labels for an input image. For each label, a template sentence is chosen depending on the absence or presence of the label. Finally, the chosen template sentences are concatenated into a final report.

MIMIC-CXR more weight due to its larger number of facts to classify, as discussed in Section 2.1.

For CheXpert and VinDr-CXR, a hybrid loss combining standard BCE, Weighted by Class BCE, and Focal Loss was used because these datasets have a fixed number of classes. For MIMIC-CXR and OpenI, BCE + Focal Loss was employed. In the case of VinDr-CXR, the Mean Absolute Error (MAE) between the predicted attention map and the ground-truth bounding boxes is used as attention supervision loss for visual grounding of the classified facts. The AdamW optimizer (Loshchilov and Hutter, 2019) was used with a cyclic exponential learning rate varying from 1e-4 to 1e-6 over 8 epochs. Each epoch consisted of approximately 800 batches. The model was trained for about 20 hours, after which no significant gains in validation metrics were observed. The batch size was 13 images per batch, with about 40 facts sampled per image. Combined with 10 gradient accumulation steps, the effective batch size was 130 images. Images were resized to 416×416.

All experiments were implemented using Python 3.10.10 with PyTorch version 1.13.1+cu117 (Paszke et al., 2017). The experiments were conducted on a computing node equipped with a 20-core Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz, three NVIDIA GPUs—two GeForce RTX 2080 Ti with 11GB memory and one GeForce RTX 3090 with 24GB memory. The system was complemented by 125GB of RAM.

3.3 Report Generation using Templates

For report generation, curated sets of two sentences per abnormality were manually selected to indicate presence and absence. These sets are categorized into different types of templates (Pino et al., 2021): *Mimic Style*, *Ambiguous*, *Fusion*, and *Fusion + Groups*.

The *Mimic Style* sentences correspond to a simple template shown in Appendix Table 5, while the *Ambiguous* sentences correspond to the template shown in Appendix Table 6. On the other hand, the *Fusion* template combines the absent template sentences from *Mimic Style* with the present template sentences from *Ambiguous*.

The *Fusion + Groups* template functions differently from the other templates. Instead of replacing a sentence for each label, it groups labels together. If a group of labels matches the value of abnormalities specified in a grouped template (see Appendix Table 7), that template is added to the final report. After iterating through all grouped templates, the remaining abnormalities are addressed using the *Fusion* template for each individual disease, thus giving the template its name *Fusion + Groups*.

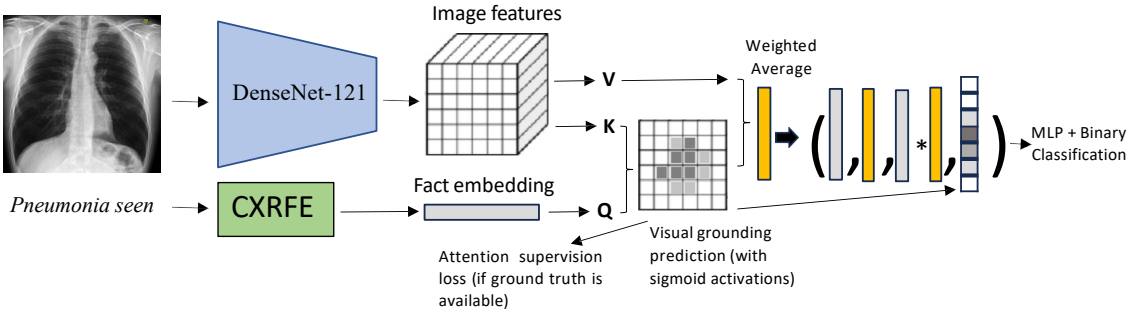

Figure 3: Fact Classifier architecture of iHealth-Chile-2. CXRFE stands for **C**hest **X**-**R**ay **F**act **E**ncoder, an improved version of CXR-BERT (Boecking et al., 2022) via several NLP tasks, as outlined in a concurrent publication (Messina et al., 2024). This Fact Classifier is an experimental architecture, that seeks to generalize the stage-1 classifier of CheXfusion (Kim, 2023). Unlike iHealth-Chile-3, the Fact Classifier is trained on all views, and during inference the predictions from all views are aggregated via max-pooling.

Table 1: Classification metrics on the MIMIC-CXR and CheXpert validation sets using the CNN trained by iHealth-Chile-3.

Precision	Recall	F1-Micro	F1-Macro
0.36	0.74	0.48	0.36

4 Experiments and Results

iHealth-Chile-3. After training the CNN, we obtained the classification results shown in Table 1. We achieved a precision of 0.36, which, being relatively low, immediately impacts our performance on the NLP metrics discussed later in this section. Furthermore, the significantly lower value of F1-Macro compared to F1-Micro suggests that the model performs notably weaker on specific labels, likely due to class imbalance.

Table 2 presents the results of report generation on the findings and impression hidden test sets. The metrics detailed are BLEU4 (B4 Papineni et al., 2002), ROUGE-L (RL Lin, 2004), BERTScore (BS Zhang et al., 2019), F1-CheXbert (chX Smit et al., 2020), and F1-RadGraph (RG Delbrouck et al., 2022a). All values were calculated using the official leaderboard web page with the VilMedic framework (Delbrouck et al., 2022b). By examining Table 2, we can observe that the Template Type which most increases the F1-RadGraph score is the *Ambiguous* Template type, improving this score by at least 2 points compared to the *Mimic Style* Template. This improvement is likely due to the inclusion of location-specific terms like "left" and "right." However, there is a corresponding decrease in BLEU4, possibly because the ground-truth report specifies the location of the disease, and the addition of terms like "left" and "right" might introduce inaccuracies.

Additionally, Table 2 reveals that the best Template for the findings section, based on F1-RadGraph, is the *Fusion + Groups* template, while for the impression section, the best is the *Fusion* Template.

On the other hand, the F1-CheXbert score is lower than the F1-Macro and F1-Micro scores for the classification of CheXpert labels. This suggests that the BERT model used for the F1-CheXbert metric may not accurately detect some of the labels encoded in the template-generated sentences, even if they are simple, making this metric potentially unreliable for this task. A similar issue is observed with BERTScore, which does not consistently align with the other metrics.

iHealth-Chile-2. Table 3 presents the classification and template-based report generation metrics on the MIMIC-CXR and CheXpert validation sets. We highlight two notable results from this Table: (1) The Fact Classifier achieves significantly higher scores when evaluated with labels produced by the same tool used to annotate the training set (i.e., VisualCheXbert for CheXpert and the NLI labeler for MIMIC-CXR); and (2) The performance drops when the CheXpert labeler and CheXbert evaluate a template-based report built from the classifications, particularly with F1-CheXbert (macro and micro). This provides further evidence that the metric may be missing some of the labels mentioned in the templates.

Additional evidence of the impact of the labeling tool on the evaluation is provided in Appendix Table 8. One evaluation considers 78 classes from the

Table 2: iHealth-Chile-3's metrics on the hidden test sets. All metrics are calculated using Vilmedic on the official challenge web page. The abbreviations used are: B4 (BLEU4), RL (ROUGE-L), BS (Bertscore), cXb (F1-cheXbert), and RG (F1-RadGraph).

Template Type	Findings Hidden Test Set					Impression Hidden Test Set				
	B4	RL	BS	cXb	RG	B4	RL	BS	cXb	RG
Mimic Style	4.74	16.17	**47.28**	27.44	13.08	**1.72**	9.41	36.18	24.55	8.30
Ambiguous	3.58	14.65	44.99	**29.35**	15.85	1.64	9.84	**37.38**	**26.84**	10.34
Fusion	**4.80**	16.88	46.73	28.20	18.70	1.66	**10.21**	37.21	25.82	**11.58**
Fusion + Groups	4.18	**17.05**	42.91	27.20	**19.42**	1.42	10.13	33.01	24.91	11.53

Table 3: Classification and Template-based Report Generation results on the validation sets of MIMIC-CXR and CheXpert. The classes considered are the 14 classes of the CheXpert dataset. On MIMIC-CXR we consider two sources of ground-truth labels for evaluation: the CheXpert labeler and our own NLI labeler. In the case of CheXpert, we use the labels produced by VisualCheXbert (Jain et al., 2021) that were released with the dataset. The reports were produced with the *Fusion + Groups* technique.

Classification: CheXpert labeler / VisualCheXbert				**Classification: NLI labeler (ours)**				**Template-based Report Generation**			
F1 (micro)	**F1 (macro)**	**PRC-AUC (micro)**	**PRC-AUC (macro)**	**F1 (micro)**	**F1 (macro)**	**PRC-AUC (micro)**	**PRC-AUC (macro)**	**F1-CheXp (micro)**	**F1-CheXp (macro)**	**F1-CheXb (micro)**	**F1-CheXb (macro)**
MIMIC-CXR validation set (9178 images)											
0.491	0.405	0.418	0.416	0.628	0.519	0.668	0.557	0.510	0.424	0.430	0.372
CheXpert validation set (5468 images)											
0.679	0.554	0.719	0.717	-	-	-	-	0.539	0.417	0.442	0.358

Table 4: iHealth-Chile-2's metrics on the findings-hidden-test-set and impression-hidden-test-set.

Dataset	**Method**	**B4**	**RL**	**BS**	**cXb**	**RG**
findings-hidden-test-set	Fact Classifier + Templates (*Fusion + Groups*)	**4.81**	**15.96**	**44.03**	**33.69**	**18.41**
findings-hidden-test-set	Fact Classifier + BART (findings, v1)	2.33	14.22	43.39	28.00	14.48
findings-hidden-test-set	Fact Classifier + BART (findings, v2)	2.78	14.29	43.40	31.00	14.74
impression-hidden-test-set	Fact Classifier + BART (impression)	2.28	11.33	35.98	20.87	7.59

Chest ImaGenome dataset (Wu et al., 2021), while the other considers the 26 classes from the CXR-LT 2023 challenge (Holste et al., 2023). Noticeably, the performance drops significantly when evaluated with the original labels compared to the labels generated by our NLI system. This discrepancy suggests that either our NLI system is incorrect, or the labels provided by the original datasets, which were also extracted from reports, are inaccurate. This issue warrants further investigation in future work.

Lastly, Table 4 presents all submissions by iHealth-Chile-2 to the hidden test set (findings and impression). The best approach is clearly based on templates. However, for completeness, we also include unsuccessful attempts at producing reports generatively using BART (Lewis et al., 2020), a sequence-to-sequence model, by training it to generate reports from templates. This approach degraded performance, so we advise against it.

5 Conclusions and Future Work

We have presented the results of the iHealth-Chile-3 and iHealth-Chile-2 teams in the Large-Scale Radiology Report Generation shared task. Both teams used a template-based method, where an image classifier predicts specific classes, which are then used to generate a report with predefined templates. The performance in the challenge was modest. Interestingly, despite the templates being tailored for CheXpert classes, the F1-CheXbert metrics were consistently lower than the classification metrics.

Based on these results, future work should focus on: (1) Thoroughly evaluating report generation metrics to identify and address limitations in existing ones; (2) Improving chest X-ray image classifiers, particularly for long-tail classes; and (3) Developing more advanced report generation systems that surpass rigid templates while preserving classifier accuracy for long-tail classes.

6 Limitations

The iHealth-Chile-3's approach has several limitations that warrant discussion. Firstly, this approach is restricted in its ability to specify the location of detected abnormalities. It can only confirm the presence or absence of these abnormalities without providing detailed localization within the images. This spatial limitation may affect clinical applicability, where precise localization is often critical.

Secondly, the overall performance of the reports generated by this approach is inherently tied to the performance of the multi-label classifier employed. Any deficiencies or inaccuracies in the classifier directly impact the quality and reliability of the generated reports. Moreover, even if the multi-label classifier were to achieve perfect performance, the scope of the reports would still be confined to the 13 specific labels used in this approach. This means that any abnormalities outside these predefined categories would go unreported, potentially missing other clinically significant findings.

Additionally, the resolution of the images used in this study, limited to 256x256 pixels, could further constrain the performance. Lower resolution images may lack the necessary detail for accurate detection and classification of certain abnormalities, leading to potential misclassification or oversight. Future work could explore the impact of using higher resolution images to determine if this enhances the diagnostic accuracy and overall utility of the approach.

The strategy adopted by iHealth-Chile-2 has notable limitations as well. Firstly, it is based on an experimental architecture still under development and unpublished at the time of this writing. It also depends on an auxiliary Natural Language Inference (NLI) system that is being developed concurrently, with significant involvement of GPT-4. As discussed in Section 4, the discrepancies between the original labels from source datasets and our NLI-based labels highlight the need for further investigation. We aim to elaborate on these aspects in future publications.

The Fact Classifier tested by iHealth-Chile-2 may also be limited by its use of DenseNet-121 as its visual backbone. Given the advances in architectures based on vision transformers, such as the Swin Transformer (Liu et al., 2021), DenseNet-121 might not be the optimal choice. This limitation is also shared by iHealth-Chile-3.

Lastly, a significant limitation in the classification approach itself followed by both teams is the lack of a clear strategy for translating classifications into a final natural language report. Even if an optimal open-vocabulary classifier were to accurately identify a comprehensive list of abnormalities with good visual grounding, it remains unclear how to convert these predictions into a report that scores well according to the challenge metrics. This gap between classification/visual grounding and report generation warrants further investigation.

Acknowledgements

This work has been funded by Millenium Science Initiative Program ICN2021_004 (iHEALTH); IMFD ICN17 002, and Fondecyt grants 11230762 and 1231724; National Center for Artificial Intelligence (CENIA) FB210017, Basal ANID; Fondecyt 1221425; and the National Agency for Research and Development (ANID) through the Scholarship Program / Doctorado Becas Chile / 2019 - 21191569.

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. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations, pages 23–34.

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

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120.

Gregory Holste, Song Wang, Ajay Jaiswal, Yuzhe Yang, Mingquan Lin, Yifan Peng, and Atlas Wang. 2023. Cxr-lt: Multi-label long-tailed classification on chest x-rays. *PhysioNet*, 5:19.

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

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. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. 2021. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Dongkyun Kim. 2023. Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2702–2710.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40.

Pablo Messina, René Vidal, Denis Parra, Álvaro Soto, and Vladimir Araujo. 2024. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. 2022. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *Machine Learning in Medical Imaging*, pages 654–663, Cham. Springer International Publishing.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Eric Topol. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, 1st edition. Basic Books, Inc., USA.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Joy T Wu, Nkechinyere Agu, Ismini Lourentzou, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest imagenome dataset for clinical reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Templates used by Health-Chile-3's approach

The *Mimic Style* template set, shown in Table 5, corresponds to sentences which simply indicate presence or absence of the labels. This template set was named *Mimic Style* because the sentences were chosen manually to imitate the sentences found in the MIMIC-CXR reports.

The *Ambiguous* template set, shown in Table 6, corresponds to sentences which when they indicate presence are ambiguous. For example, they can be ambiguous in terms of location, indicating the presence of an abnormality on the left or right side of the image.

Finally, the *Group* template set (not to be confused with the *Fusion + Groups* template approach) serves as an auxiliary template to be combined with the simpler templates that indicate the single presence of labels. This template set is shown in detail in Table 7.

Table 5: Sentences in the Mimic Style template set.

Abnormality	**Absence template**	**Presence template**
Cardiomegaly	Heart size is normal	The heart is enlarged
Enlarged Cardiomed.	The mediastinal contour is normal	The cardiomediastinal silhouette is enlarged
Consolidation	No focal consolidation	There is focal consolidation
Lung Opacity	The lungs are free of focal airspace disease	One or more airspace opacities are seen
Atelectasis	No atelectasis	Appearance suggest atelectasis
Pleural Effusion	No pleural effusion	Pleural effusion is seen
Pleural Other	No fibrosis	Pleural thickening is present
Pneumonia	No pneumonia	There is evidence of pneumonia
Pneumothorax	No pneumothorax is seen	There is pneumothorax
Edema	No pulmonary edema	Pulmonary edema is seen
Lung Lesion	No pulmonary nodules or mass lesions identified	There are pulmonary nodules or mass identified
Fracture	No fracture is seen	A fracture is identified
Support Devices	-	A device is seen

Table 6: Sentences in the Ambiguous template set.

Abnormality	**Absence template**	**Presence template**
Cardiomegaly	no cardiomegaly	the heart is stable, mild, moderate, severe or enlarged in size
Enlarged Cardiomed.	mediastinal contour is normal	the cardiomediastinal silhouette is unchanged, enlarged or widened
Consolidation	no consolidation	there is observed left or right lung consolidation
Lung Opacity	free of focal airspace disease	there are left or right present lung airspace opacities
Atelectasis	no atelectasis	there is observed left or right lung present atelectasis
Pleural Effusion	no pleural effusion	there is an observed left, right or bilateral, small, moderate or large pleural effusion
Pleural Other	no fibrosis	there is present left or right, minimal, mild or severe pleural thickening
Pneumonia	no pneumonia	observed process left or right lung pneumonia
Pneumothorax	no pneumothorax	there is noted left sided or right sided, small, moderate or large pneumothorax in the lung
Edema	no pulmonary edema	there is noted mild, moderate or severe pulmonary edema
Lung Lesion	no pulmonary nodules	there are left or right pulmonary lung nodules observed
Fracture	no fracture	there is a rib or clavicular left or right sided fracture
Support Devices	there is no picc line	there is a noted right sided or left sided picc or tube

Table 7: Sentences for Group Template.

Abnormalities	**Value of labels**	**Template Group Sentence**
'Lung Lesion', 'Lung Opacity', 'Edema', 'Consolidation', 'Pneumonia', 'Atelectasis'	0 (all absent)	the lungs are clear
'Consolidation', 'Pleural Effusion', 'Pneumothorax'	0 (all absent)	there is no focal consolidation , pleural effusion , or pneumothorax .
'Pneumothorax', 'Pleural Effusion'	0 (all absent)	there is no pleural effusion or pneumothorax .
'Pneumothorax', 'Consolidation'	0 (all absent)	there is no focal consolidation or pneumothorax .

Table 8: Fact classification results on MIMIC-CXR test set. These results are shown for illustrative purposes only. The performance achieved by the fact classifier according to the labels produced by our NLI labeler is significantly higher than the performance according to the original labeling tools of the datasets.

Original Labeler		**NLI labeler**	
F1 (micro)	**F1 (macro)**	**F1 (micro)**	**F1 (macro)**
CXR-LT (26 classes)			
0.451	0.306	0.620	0.454
Chest ImaGenome (78 classes)			
0.321	0.261	0.533	0.355