# A Dual-View Approach to Classifying Radiology Reports by Co-Training

**Yutong Han**[1], **Yan Yuan**[2], **Lili Mou**[1,3]
[1]Dept. Computing Science, Alberta Machine Intelligence Institute (Amii), University of Alberta
[2]School of Public Health, University of Alberta
[3]Canada CIFAR AI Chair, Amii
yhan22@ualberta.ca, yyuan@ualberta.ca, doublepower.mou@gmail.com

**Abstract**

Radiology report analysis provides valuable information that can aid with public health initiatives, and has been attracting increasing attention from the research community. In this work, we present a novel insight that the structure of a radiology report (namely, the *Findings* and *Impression* sections) offers different views of a radiology scan. Based on this intuition, we further propose a co-training approach, where two machine learning models are built upon the *Findings* and *Impression* sections, respectively, and use each other's information to boost performance with massive unlabeled data in a semi-supervised manner. We conducted experiments in a public health surveillance study, and results show that our co-training approach is able to improve performance using the dual views and surpass competing supervised and semi-supervised methods.

**Keywords:** Radiology report analysis, Co-training, Semi-supervised learning

## 1. Introduction

Radiology report analysis plays an important role in patient diagnosis and monitoring (Carlson et al., 2020; Machitori et al., 2020; Anzai et al., 2023). For example, brain radiology reports—such as those derived from magnetic resonance imaging (MRI) and computed tomography (CT)—are typically in the form of free text, and can be used to determine the presence of brain tumors and track their progression over time. This helps collect surveillance data for public health initiatives (Yuan et al., 2018).

Machine learning methods have been widely applied to the radiology domain, as the ever-growing volume of radiology reports makes it difficult for humans to label every single one. In early work, researchers perform manual feature engineering to construct classifiers such as decision trees (Yadav et al., 2013) and support vector machines (Grundmeier et al., 2016). More recently, deep learning has been a prevailing approach to radiology report analysis, leading to great advancements in the field. Wood et al. (2020) finetune the BioBERT model (Alsentzer et al., 2019) for MRI scan classification. Smit et al. (2020) use the labels produced by a traditional rule-based X-ray classifier (Irvin et al., 2019) to train a BERT model (Devlin et al., 2019), which outperforms the rule-based classifier.

However, we observe that existing methods do not make full use of the internal structures of a radiology report, which typically contains a *Findings* section and an *Impression* section. The former details factual observations made by a radiologist, whereas the latter synthesizes their findings into a summary (Ghosh et al., 2023). Our intuition is that such structural information can provide different views of a radiology report and improve the performance of machine learning systems.

In this paper, we propose a co-training approach to radiology report analysis, framing the *Findings* and *Impression* sections as two different views. Specifically, we train two classifiers for *Findings* and *Impression*, respectively. Then, we use one classifier's predicted labels to train the other in a co-training fashion (Blum and Mitchell, 1998), which makes use of a large unlabeled dataset. These co-trained classifiers can be combined as an ensemble (Dietterich, 2000) to make final predictions. In this way, the model trained on one section is able to glean information from the other in a semi-supervised manner. This allows us to make use of the structure of a typical radiology report as well as unlabeled data to improve overall performance.

We conducted experiments for a brain tumor surveillance project in collaboration with Alberta Health Services (AHS), a Canadian provincial health agency, where we are provided with de-identified historical radiology reports of real patients. The results show that co-training improves each individual model in a semi-supervised manner, and that their ensemble is able to further boost the performance. Our entire approach outperforms both supervised learning based on the small labeled data and self-train, a competing semi-supervised method.[1]

[1]Code available at: https://github.com/MANGA-UOFA/Radiology-Cotrain



*LREC-COLING 2024*, pages 477–483
20-25 May, 2024. 

## 2. Related Work

**Semi-supervised learning** assumes only a small labeled dataset exists, and takes advantage of massive, readily available unlabeled data to improve model performance (Ouali et al., 2020). Two popular frameworks are self-training and co-training. In self-training, a model generates pseudo-labels for the unlabeled data and trains itself (Yarowsky, 1995), whereas in co-training, two models are built on two views (different input information about a data sample) and co-train each other (Blum and Mitchell, 1998). In fact, co-training has been previously used in various NLP applications. Maveli and Cohen (2022) use inside-span and outside-span views to co-train an unsupervised constituency parser; Wang et al. (2022) use a query view and a document view to co-train a selective search system; and Lang et al. (2022) use two different language models' representations to co-train and improve the performance of a prompting system.

**Radiology report analysis** has gained traction in recent years (Karimi et al., 2017; Khanna et al., 2023), as the textual reports provide rich supplementary information to images (Wood et al., 2020; Dalla Serra et al., 2022). While an entire report can be the input to a machine learning system (Drozdov et al., 2020; Di Noto et al., 2021), researchers have realized the value of using the section structure of radiology reports. Peng et al. (2020) use the *Findings* section to extract information about lymph nodes in abdominal MRI reports; Irvin et al. (2019) use the *Impression* section to create a rule-based pathology classifier for chest X-rays. Zhang et al. (2018) train a text generation system to automatically synthesize an *Impression* section from the *Findings* section.

To the best of our knowledge, we are the first to propose a co-training method based on *Findings* and *Impression*, as well as to build model ensembles of the two sections.

## 3. Approach

**Formulation.** Given a radiology report $\mathbf{x}$, our goal is to predict a label $y \in \{0, \cdots, K-1\}$ with $K$ predetermined categories. For example, an important label for brain radiology reports is $y \in \{0, 1\}$, indicating the absence or presence of a brain tumor.

In this work, we need to tackle a common and realistic setting for radiology report analysis: we only have a small set of labeled reports $\mathcal{D}_l = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{L}$, but there exists a large unlabeled dataset $\mathcal{D}_u = \{\mathbf{x}^{(j)}\}_{j=1}^{U}$.

Our intuition is that a typical radiology report has internal structures. In Figure 1a, for example, the report has several sections, namely, *History*, *Technique*, *Findings*, and *Impression*. In particular, we observe that *History* does not provide the information of the current report and that *Technique* explains the operational procedure; they are therefore not helpful for our task. On the other hand, the *Findings* section describes all observations made by a radiologist, and the *Impression* section summarizes and interprets the key findings. Thus, we discard *History* and *Technique* in our approach, but make use of *Findings* (denoted by $\mathbf{x}_{\text{fnd}}$) and *Impression* (denoted by $\mathbf{x}_{\text{imp}}$) as the two views for co-training. In other words, the input of a sample can be represented by $\mathbf{x} = (\mathbf{x}_{\text{fnd}}, \mathbf{x}_{\text{imp}})$.

**Supervised Initialization.** Before co-training, we first use the small labeled data $\mathcal{D}_l$ to initialize the *Findings* and *Impression* classifiers, $P_{\text{fnd}}(y|\mathbf{x}_{\text{fnd}}; \boldsymbol{\theta}_{\text{fnd}})$ and $P_{\text{imp}}(y|\mathbf{x}_{\text{imp}}; \boldsymbol{\theta}_{\text{imp}})$. Specifically, we first train them by finetuning DistilBERT (Sanh et al., 2019), a small distilled version of the pretrained language model BERT (Devlin et al., 2019).[2] Take the *Findings* view as an example: we apply linear transformation to the final layer's hidden state associated with [CLS], a token prepended to a sequence for classification. Then, a softmax function predicts a probability distribution[3] by $P(y|\mathbf{x}_{\text{fnd}}; \boldsymbol{\theta}_{\text{fnd}}) = \text{softmax}(\boldsymbol{W}_{\text{fnd}} \boldsymbol{h}_{\text{[CLS]}} + \boldsymbol{b}_{\text{fnd}})$, where $\boldsymbol{h}_{\text{[CLS]}}$ is the [CLS] token's representation at the last hidden layer. $\boldsymbol{\theta}_{\text{fnd}}$ is the entire parameter set, including softmax-layer parameters ($\boldsymbol{W}_{\text{fnd}}$ and $\boldsymbol{b}_{\text{fnd}}$), as well as the parameters of the *Findings*-view model. The training is accomplished by maximum likelihood estimation with labeled data $\mathcal{D}_l$:

$$\boldsymbol{\theta}_{\text{fnd}} = \text{argmax}_{\boldsymbol{\theta}_{\text{fnd}}} \sum_{i=1}^{L} \log P(y^{(i)}|\mathbf{x}_{\text{fnd}}^{(i)}; \boldsymbol{\theta}_{\text{fnd}}) \quad (1)$$

Likewise, we train a classifier $P_{\text{imp}}(y|\mathbf{x}_{\text{imp}}; \boldsymbol{\theta}_{\text{imp}})$ for the *Impression* view. These supervised classifiers serve as a good starting point for our co-training procedure.

**Co-Training.** The overview of our approach is presented in Figure 1b. We maintain two classifiers for *Findings* and *Impression*, respectively. Our co-training approach alternately applies each classifier to the unlabeled data $\mathcal{D}_u$, which produces pseudo-labels to co-train the other classifier. This process is repeated for performance improvement.

Consider using the classifier for *Findings* to co-train that for *Impression*. For every unlabeled sample $\mathbf{x}^{(j)}$ in $\mathcal{D}_u$, we apply the *Findings* classifier and obtain its prediction

$$\hat{y}_{\text{fnd}}^{(j)} = \text{argmax}_y P(y|\mathbf{x}_{\text{fnd}}^{(j)}; \boldsymbol{\theta}_{\text{fnd}}) \quad (2)$$

with its predicted probability $P(\hat{y}_{\text{fnd}}^{(j)}|\mathbf{x}_{\text{fnd}}^{(j)}; \boldsymbol{\theta}_{\text{fnd}})$.

[2] We chose the DistilBERT model because regulations require that the research has to be conducted within the server provided by AHS, which has limited memory.

[3] We slightly misuse the notation that $P(y|\cdot)$ refers to a probability distribution, where $y$ is a random variable taking values in $\{0, \cdots, K-1\}$ for $K$-way classification.

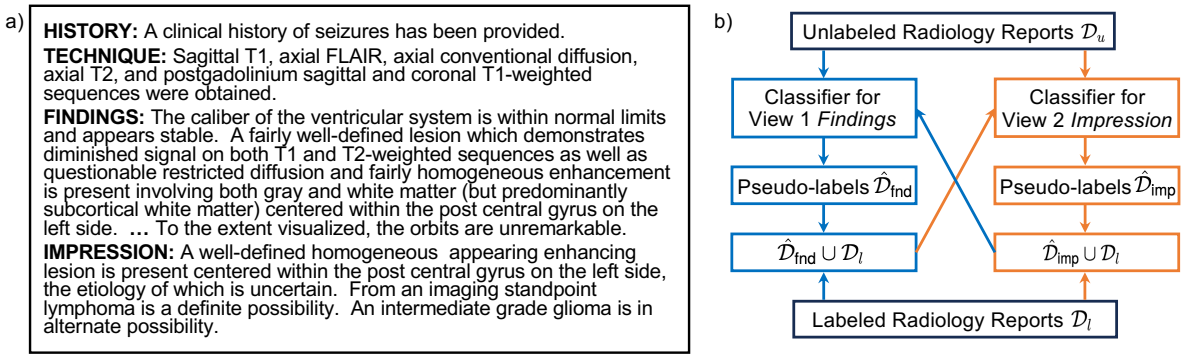


Figure 1: a) A typical radiology report. b) An overview of our co-training approach.

Then, we add high-quality labels to the training set based on two criteria. First, we select samples for which the two classifiers agree, in order to avoid confusion during co-training. That is, we have $\hat{y}^{(j)}_{\text{fnd}} = \hat{y}^{(j)}_{\text{imp}}$, where $\hat{y}^{(j)}_{\text{imp}} = \text{argmax}_y P(y|\mathbf{x}^{(j)}_{\text{imp}}; \boldsymbol{\theta}_{\text{imp}})$. Second, we choose the samples with top-$k\%$ of the *Findings*-predicted probabilities among the agreed labels. This is based on the intuition that labels with higher probabilities are more likely to be correct (Blum and Mitchell, 1998; Yarowsky, 1995), which further ensures the quality of our pseudo-labels. Overall, our pseudo-labeled dataset has the form of $\hat{\mathcal{D}}_{\text{fnd}} = \text{top-}k\%\{(\mathbf{x}^{(j)}, \hat{y}^{(j)}_{\text{fnd}}) : \hat{y}^{(j)}_{\text{fnd}} = \hat{y}^{(j)}_{\text{imp}}, \mathbf{x}^{(j)} \in \mathcal{D}_u\}$, which is merged into the labeled one as $\hat{\mathcal{D}}_{\text{fnd}} \cup \mathcal{D}_l$ to train the *Impression* classifier.

The roles then reverse to re-train the *Findings* classifier $P_{\text{fnd}}(y|\mathbf{x}_{\text{fnd}}; \boldsymbol{\theta}_{\text{fnd}})$ using the *Impression*-predicted pseudo-labels along with the original small labeled data, given by $\hat{\mathcal{D}}_{\text{imp}} \cup \mathcal{D}_l$ . Co-training continues in a such a way until validation performance peaks.

This framework allows for two views of a radiology report: the detailed factual observations in the *Findings* section and the concise synthesized information in the *Impression* section. Together, they can help each other during the co-training process and improve the classification performance.

**Ensemble for Inference.** To perform inference, we combine the co-trained classifiers by an average ensemble (Dietterich, 2000). Given an unseen radiology report $\mathbf{x}^* = (\mathbf{x}^*_{\text{fnd}}, \mathbf{x}^*_{\text{imp}})$, we apply both the *Findings* and *Impression* classifiers to the respective sections and choose the most likely category based on averaged predicted probabilities:

$$\hat{y}^* = \text{argmax}_y \tfrac{1}{2}\left[P(y|\mathbf{x}^*_{\text{fnd}}; \boldsymbol{\theta}_{\text{fnd}}) + P(y|\mathbf{x}^*_{\text{imp}}; \boldsymbol{\theta}_{\text{imp}})\right]$$

The ensemble approach makes use of the two views (*Findings* and *Impression*) by smoothing out the noise of the individual classifiers.

| Task \ Label | 0 | 1 | 2 |
|---|---|---|---|
| BT | 331 | 537 | – |
| Aggressiveness | 331 | 344 | 193 |

Table 1: Label distribution in each task. BT and Aggressiveness labels do not necessarily agree, and the number of 331 is a coincidence.

## 4. Experiments

### 4.1. Setup

We evaluated our approach for a project collaborated with Alberta Health Services, where the goal is to improve surveillance for brain tumors with historical textual radiology reports, including both CT and MRI scans. To reach this goal, the project focuses on two important labels:

• **Brain Tumor (BT):** The classification goal is $y \in \{0, 1\}$ indicating whether the radiology report suggests there is one or more brain tumors observed in the scan.

• **Aggressiveness:** Here, the classification goal is $y \in \{0, 1, 2\}$ referring to non-aggressive, aggressive, or possibly aggressive, respectively. Notice that the aggressiveness label provides different information from BT, which can be either aggressive or non-aggressive; on the other hand, an aggressive label can also be a cancer metastasis (cancer spread) that has no tumor, e.g., leukemia spread into the brain (Nguyen et al., 2023).

The dataset contains 868 radiology reports, manually annotated with the above two labels of interest, as well as 10K unlabeled radiology reports. Each report has a *Findings* section and *Impression* section, containing on average 219 and 55 tokens, respectively. The label distribution of each task is shown in Table 1. In this paper, we treat the two labels as independent tasks for the evaluation of our approach.

**Implementation Details.** Due to the small

| Setting | Row | Model | BT | Aggressiveness |
|---|---|---|---|---|
| Supervised (520 labeled samples) | 1 | Concat | 0.8837 (0.8832±0.009) | 0.8330 (0.8270±0.019) |
| | 2 | Findings | 0.8825 (0.8833±0.012) | 0.7546 (0.7768±0.018) |
| | 3 | Impression | 0.9102 (0.8915±0.013) | 0.8445 (0.8549±0.007) |
| | 4 | Ensemble (2 + 3) | 0.9148 (0.9122±0.007) | 0.8676 (0.8646±0.012) |
| Semi-Supervised (+10K unlabeled samples) | 5 | Concat (self-train) | 0.8952 | 0.8563 |
| | 6 | Findings (self-train) | 0.8906 | 0.7845 |
| | 7 | Impression (self-train) | 0.8860 | 0.8560 |
| | 8 | Ensemble (6 + 7) | 0.9160 | 0.8802 |
| | 9 | Findings (co-train) | 0.8906 | 0.8399 |
| | 10 | Impression (co-train) | 0.9044 | 0.8621 |
| | 11 | Ensemble (9 + 10) | **0.9286** | **0.8848** |

Table 2: Accuracy on the Brain Tumor (BT) and Aggressiveness classification tasks. "Concat" refers to training a single model with concatenated *Findings* and *Impression* sections (without the section titles) as the input. For the supervised setting, we trained the model with 5 different seeds, and report the median and the (mean±standard deviation). Due to limited computing resources, self-training and co-training settings were run once initialized from the median run of the supervised setting.

dataset, we performed 5-fold cross-validation for robust evaluation. We split the entire dataset into five folds, and for the test of each fold (174 samples), we used three folds for training (520 samples) and the other fold for validation (174 samples).

We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 16 and a standard learning rate of 5e-5 (Devlin et al., 2019). Early stopping was implemented based on validation performance in each co-training round; we set the number of maximum co-training rounds to be 5, also early stopped by validation. For the BT task, the top-50% pseudo-labels were added at each co-training step, whereas for the Aggressiveness task, the top-25% were added. We will analyze the choice of top-$k$% in §4.2.

## 4.2. Results

Table 2 shows the main results on the two tasks, BT and Aggressiveness. Here, the performance is measured by accuracy, since our dataset is relatively balanced as shown in Table 1.

We first analyze the use of *Findings* and *Impression* sections in the supervised setting only (Rows 1–4). As seen, *Impression* (Row 3) yields higher performance than *Findings* (Row 2), especially for the Aggressiveness task. This is understandable because *Impression* is a synthesized summary and better aligns with the actual label. Concatenation of the two sections, without section titles (Row 1) does not outperform *Impression* only (Row 3), as the *Findings* section may be of considerable length and confuse the model. Our ensemble approach, even without co-training (Row 4), achieves consistent improvement on both tasks, justifying our dual views that make use of the internal structure of a radiology report.

Then, we performed semi-supervised learning using 10K unlabeled samples. In addition to our co-training method, we experimented with a self-training baseline, where a model uses its own predictions to boost its performance. In general, semi-supervised learning (Rows 5–11) outperforms small-scale supervised learning (Rows 1–4), except for a minor inconsistency of *Impression* on the BT task. The results verify that massive unlabeled data can alleviate the data sparsity problem in the medical domain.

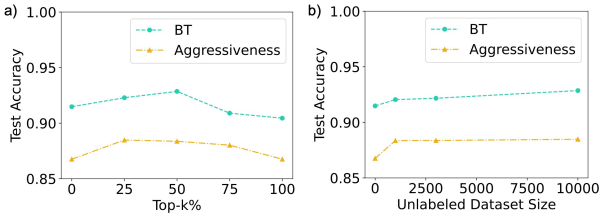


Figure 2: a) Effect of top-k% on co-training performance, when fixing the unlabeled dataset size at 10,000. b) Effect of unlabeled dataset size, using $k = 50$ for BT and $k = 25$ for Aggressiveness.

Among semi-supervised learning methods, we observe that co-training always excels when compared with self-training. For example, our co-trained *Findings* classifier (Row 9) improves the accuracy by 8 points compared with supervised learning (Row 2) on Aggressiveness, whereas the self-trained *Findings* classifier (Row 6) is only improved by 3 points. Our ensemble (Row 11) is able to further boost the performance and surpasses the ensemble of supervised counterparts (Row 4). This indicates that exchanging dual-view information by co-training is more powerful than a *post hoc* ensemble.

Moreover, our full method (co-training and ensemble, Row 11) outperforms a naïve application of DistilBERT to radiology reports (Row 1) by 4.49 and 5.18 percentage points for BT and Aggressiveness, respectively. This significant jump in performance further suggests that our approach is effective.

Overall, the experiments convincingly show that

making use of the *Findings* and *Impression* sections benefit radiology report analysis, and that our co-training approach is able to strategically exploit the dual views of the report to gain additional benefits.

**Detailed Analyses.** In our approach, we choose top-$k\%$ confident samples for co-training, and we analyzed the effect of the hyperparameter $k$ in Figure 2a. As seen, a modest $k$ yields highest performance, which is reasonable because a smaller $k$ results in fewer pseudo-labels, whereas a larger $k$ brings in more noise. Based on the analysis, we chose $k = 50$ for BT and $k = 25$ for Aggressiveness.

We also analyzed the role of the unlabeled dataset size in the co-training process. Figure 2b shows that the performance is generally improved with more data, but is not sensitive to the exact number of samples. We chose 10K unlabeled samples for co-training in our experiments.

## 5. Conclusion

In this paper, we propose a co-training approach to radiology report analysis, where we regard the *Findings* and *Impression* sections as dual views of a radiology report. We conducted experiments on two tasks: Brain Tumor (BT) classification and Aggressiveness classification. The experimental results demonstrate that our co-training method is able to make use of the dual views with unlabeled data in a semi-supervised manner, and outperforms different competing methods. We further provide detailed analyses of our proposed co-training method.

## 6. Ethics Statements

Our study involves de-identified patient data and human annotations. We have obtained ethics approvals from our research institute, as well as the partner government agency of public health. This project will provide timely and accurate data that can be used to inform health system resource allocation and improve understanding of the cancer recurrence/progression at the population level.

## 7. Acknowledgments

The research is supported in part by the Canadian Cancer Society, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Amii Fellow Program, the Canada CIFAR AI Chair Program, an Alberta Innovates Project, a UAHJIC Project, a donation from DeepMind, and the Digital Research Alliance of Canada (alliancecan.ca).

## 8. Bibliographical References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*, pages 72–78.

Yoshimi Anzai, Chun-Pin Chang, Kerry Rowe, John Snyder, Vikrant Deshmukh, Michael Newman, Alison Fraser, Ken Smith, Ankita Date, Carlos Galvao, Marcus Monroe, and Mia Hashibe. 2023. Surveillance imaging with PET/CT and CT and/or MRI for head and neck cancer and mortality: A population-based study. *Radiology*, 307(2):e212915.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Conference on Computational Learning Theory*, page 92–100.

Brandon B. Carlson, Stephan N. Salzmann, Toshiyuki Shirahata, Courtney Ortiz Miller, John A. Carrino, Jingyan Yang, Marie–Jacqueline Reisener, Andrew A. Sama, Frank P. Cammisa, Federico Pablo Girardi, and Alexander P. Hughes. 2020. Prevalence of osteoporosis and osteopenia diagnosed using quantitative CT in 296 consecutive lumbar fusion patients. *Neurosurgical Focus*, 49(2):E5.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. 2022. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–624.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter for Association of Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Tommaso Di Noto, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas Hottinger, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. 2021. Diagnostic surveillance of high-grade gliomas: Towards automated change detection using radiology report classification. In *Machine Learning and Principles and Practice*

*of Knowledge Discovery in Databases*, pages 423–436.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15.

Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David J Lowe. 2020. Supervised and unsupervised language modelling in chest X-ray radiological reports. *PLoS One*, 15(3):e0229963.

Rikhiya Ghosh, Oladimeji Farri, Sanjeev Kumar Karn, Manuela Danu, Ramya Vunikili, and Larisa Micu. 2023. RadLing: Towards efficient radiology report understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 640–651.

Robert Grundmeier, Aaron Masino, T. Casper, J. Dean, Jamie Bell, Rene Enriquez, Sara Deakyne, James Chamberlain, and Elizabeth Alpern. 2016. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. *Applied Clinical Informatics*, 7(4):1051–1068.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597.

Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *Proceedings of the Biomedical Natural Language Processing Workshop*, pages 328–332.

Sameer Khanna, Adam Dejl, Kibo Yoon, Quoc Hung Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. 2023. RadGraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine Learning for Healthcare*, pages 1–28.

Hunter Lang, Monica Agrawal, Yoon Kim, and David A. Sontag. 2022. Co-training improves prompt-based learning for large language models. In *International Conference on Machine Learning*.

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

Akihiro Machitori, Tomoyuki Noguchi, Yusuke Kawata, Nobuhiko Horioka, Akihiro Nishie, Daisuke Kakihara, Kousei Ishigami, Shigeki Aoki, and Yutaka Imai. 2020. Computed tomography surveillance helps tracking COVID-19 outbreak. *Japanese Journal of Radiology*, 38(12):1169–1176.

Nickil Maveli and Shay Cohen. 2022. Co-training an unsupervised constituency parser with weak supervision. In *Findings of the Association for Computational Linguistics: ACL*, pages 1274–1291.

Andrew Nguyen, Alexander Nguyen, Oluwaferanmi T Dada, Persis D Desai, Jacob C Ricci, Nikhil B Godbole, Kevin Pierre, and Brandon Lucke-Wold. 2023. Leptomeningeal metastasis: A review of the pathophysiology, diagnostic methodology, and therapeutic landscape. *Current Oncology*, 30(6):5906–5931.

Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An overview of deep semi-supervised learning. *arXiv preprint*, arXiv:2006.05278.

Yifan Peng, Sungwon Lee, Daniel C. Elton, Thomas Shen, Yu-xing Tang, Qingyu Chen, Shuai Wang, Yingying Zhu, Ronald Summers, and Zhiyong Lu. 2020. Automatic recognition of abdominal lymph nodes from clinical text. In *Proceedings of the Clinical Natural Language Processing Workshop*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Arijit Sehanobish, Nathaniel Brown, Ishita Daga, Jayashri Pawar, Danielle Torres, Anasuya Das, Murray Becker, Richard Herzog, Benjamin Odry, and Ron Vianu. 2023. Efficient extraction of pathologies from C-spine radiology reports using multi-task learning. In *International Workshop on Health Intelligence*, pages 335–346.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Zhanyu Wang, Xiao Zhang, Hyokun Yun, Choon Hui Teo, and Trishul Chilimbi. 2022. MICO: Selective search with mutual information co-training. In *Proceedings of International Conference on Computational Linguistics*, pages 1179–1192.

David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth J. Barker, Sébastien Ourselin, James H. Cole, and Thomas C. Booth. 2020. Automated labelling using an attention model for radiology reports of MRI scans (ALARM). In *Proceedings of the Conference on Medical Imaging with Deep Learning*.

Kabir Yadav, Efsun Sarioglu, Meaghan Smith, and Hyeong-Ah Choi. 2013. Automated outcome classification of emergency department computed tomography imaging reports. *Academic Emergency Medicine*, 20(8):848–854.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

Yan Yuan, Sana Amjad, Angela Eckstrand, Rob Sevick, James Scott, Shebnum Devji, Christine Bertrand, Mary Jane King, Victor Brunka, Emily Maplethorpe, et al. 2018. On capturing radiological diagnoses of brain tumors to provide complete population data in cancer registries in Canada. *Journal of Registry Management*, 45(4):167–172.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.