Unlocking Diagnostic Potential with Fusion Techniques: An Investigation of Multimodal Chest X-Ray Diagnosis

Zhe Fan[1]
Tianyu Cao[1]
Chenghao Gong[1]
Yujie Chen[1]
[1]Department of Computer Science, University of Toronto
{zhefan,evelyncao,garyc,yujiech}@cs.toronto.edu

Abstract

Deep learning on uni-modal data has been a deeply researched topic in healthcare with numerous production-level use cases. In recent years, deep learning models have tried incorporating multi-modal representations for medical diagnostic classification, which showed improvement compared to uni-modal counterparts. However, previous fusion studies largely relied on tabular electronic health record data and lacked in-depth experimentation with different fusion methods. In this study, we implemented uni-modal (visual and text) baseline encoders using BERT and DenseNet for Chest X-ray diagnosis, as well as multi-modal classifiers that leveraged different fusion techniques: 1.) Early Fusion, 2.) Late Fusion, and 3.) Intermediate Fusion. We trained all models against a subset of the MIMIC-CXR dataset and evaluated test performance metrics, including accuracy, F1, and AUROC. We hypothesized that fusion models would result in an overall improvement compared to uni-modal classifiers. We observed that fusion models performed well on the classification tasks based on test metrics despite the lower performance of the uni-modal image classifier. We also observed that the late fusion model performed the best relative to other fusion models. In conclusion, we observed that with the MIMIC-CXR dataset, the fusion models performed better than their low-performing uni-modal counterpart (DenseNet image classifier), however, their average performance did not differ significantly from the best-performing uni-modal counterpart (BERT-based text classifier).

1 Introduction

The extensive exploration of unimodal deep learning in medical data has yielded significant applications, exemplified by the Target Real-Time Warning System (TREWS) for sepsis detection and CheXNet for pneumonia diagnosis [Adams et al., 2022, Rajpurkar et al., 2017]. The success of these approaches highlights the strides made in leveraging machine learning for effective medical diagnoses. However, clinical practices inherently involve data from diverse sources. In a clinical setting, healthcare professionals assess not only radiology images but also radiology reports, patient feedback, genetic test results, and various other cues to make diagnostic decisions. Given this, embracing a multimodal framework is imperative for achieving a comprehensive analysis. Despite some exploration of multimodal learning in healthcare, there remains a dearth of studies discussing how to effectively fuse data from multiple modalities. Moreover, with the advent of transformers and large language models, no studies have incorporated these advancements for clinical text processing in the context of multimodal fusion studies. A work of Huang et al. [2020] examined various fusion models using electronic health records (EHR) and CT scans. However, EHR data is well-structured

tabular data that can be fed into a machine-learning model directly. Fusing natural language modality and image modality, both being not readable as raw data, is more challenging.

Addressing these gaps, this study aims to conduct a systematic analysis of multimodal models that incorporate medical images and text inputs. The approach involves integrating BERT as the radiology reports encoder and DenseNet as the X-ray image encoder for addressing the Chest X-ray radiology diagnosis problem [Devlin et al., 2018, Iandola et al., 2014]. This research seeks to contribute insights into the efficacy of fusion techniques based on different fusion stages, shedding light on the potential synergies between advanced language models and image processing algorithms in the realm of medical diagnostics.

2 Related Work

Every year, over 1 million adults are diagnosed with pneumonia and about 50,000 die from the disease annually in the US [Rajpurkar et al., 2017]. Since the onset of the COVID-19 pandemic, whose main death-causing complication is pneumonia, there has been a conservative estimate of close to 5 million recorded cases in Canada, resulting in around 50,000 deaths[Public Health Agency of Canada, 2023]. Given the prevalence of pneumonia and other lung diseases, Chest X-ray is the most common medical image modality in the world [Johnson et al., 2019a].

Historically, diagnosing medical conditions through Chest X-rays presents formidable challenges, including anatomy complexity, overlapping structures, subtle abnormalities, and imbalances in class distribution within the samples. The diagnostic process is further complicated by the dynamic evolution of diseases over time and the substantial workload for radiologists. Despite efforts to improve accuracy and efficiency using artificial intelligence and deep learning, the diagnostic performance of prominent models like CheXNet has limitations, with an F1 score below 0.5, only slightly better than the performance by expert physicians [Rajpurkar et al., 2017]. Recognizing these difficulties, we believe that a multi-modal approach, combining insights from various sources, could provide valuable enhancements to the diagnostic process for chest X-ray studies.

There have been previous studies that investigated the potential for multi-modal deep learning in the healthcare domain. One example is Liu et al. [2022]'s hybrid deep learning model for molecular subtype prediction from breast cancer data. In their study, genetic modality data and image modality data from each patient were transferred into respective feature extraction pipelines. Genetic modality data was presented in tabular format, therefore it was processed directly through a fully connected (FC) multi-layer neural network; on the other hand, a convolutional neural net (CNN) was used to process histopathological images. The study showed a 7.45% improvement over the uni-modal prediction on genetic data only.

Several works have explored multi-modal deep learning on the chest X-ray problem. For example, in Chauhan et al. [2020]'s study, a joint model is employed to improve uni-modal image analysis by training a late fusion model on both clinical reports and image modalities and extracting only the image classification pipeline for inference. It was shown that this approach drastically increased the AUROC and F1 metrics compared to the uni-modal baseline that was not co-trained with clinical text data. This study showed promise for multi-modal fusion of clinical transcript and image for chest X-ray classification but did not go into further analysis of the various fusion techniques that could have been employed, and the model was never used for multi-modal inference. In another work by Huang et al. [2020], multi-modal fusion deep learning was performed on Chest CT scans and tabular electronic health records. In this study, various fusion techniques were explored, including early, intermediate/joint, and late fusion, and late fusion performed the best in classification accuracy and AUROC compared to all other fusion models and the baseline unimodal models.

So far, the study to compare and evaluate different fusion techniques for a multi-modal deep learning model that incorporates both an LLM-based text encoder and an X-ray image encoder have not been done in clinical healthcare analysis. There exist previous models such as the TieNet, which used attention-based text embeddings for text encoding [Wang et al., 2018], as well as models that created an image-text embedding setup for medical image annotation [Moradi et al., 2018]. However, these models all employed a recurrent neural network (RNN) based pipeline. In contrast to previous studies, we will leverage the large language model BERT to create text embeddings that are contextualized, robust representations of clinical transcript data [Devlin et al., 2018], and compare the performance of BERT-based multimodal fusion models, differed by fusion technique, on Chest X-ray diagnosis.

3 Dataset

Our study leveraged the MIMIC-CXR Dataset (Johnson et al. [2019a]) and its derivative MIMIC-CXR-JPG, which comprises compressed JPG format images, a comprehensive collection of multi-directional chest X-ray images paired with free-text radiology reports. This dataset contributes to the medical imaging field, especially in facilitating the development of advanced multi-modal (images and natural language) models for medical diagnostic assistance.

We focused on a subset consisting of 5,601 pairs of chest X-ray images and corresponding radiology reports, representing a relatively diverse patient population across 14 labeled diseases. Each instance may carry multiple labels, allowing for the development of classifiers that can identify various pathologies in a single instance. The subset was split into training, validation, and test sets with a ratio of 7:1:2 to facilitate a thorough training and evaluation cycle. Data pre-processing steps involved downscaling images to 224x224 pixels to reduce computational demands while preserving essential diagnostic detail and tokenizing reports to an average length of 42 tokens, approximately 450 characters, to standardize model inputs.

Although this is a multi-labeled dataset, most of the instances are unique-labeled as shown in Figure 1, which might lead to reduced model ability to recognize complex cases. Additionally, there is an obvious class imbalance, with some diseases being markedly more common than others, with 'No Finding' as the most frequent label while labels like 'Fracture' and 'Pleural Other' are significantly underrepresented. as shown in Figure 2, which might lead to biased models with less sensitivity to those less frequent but clinically significant findings.

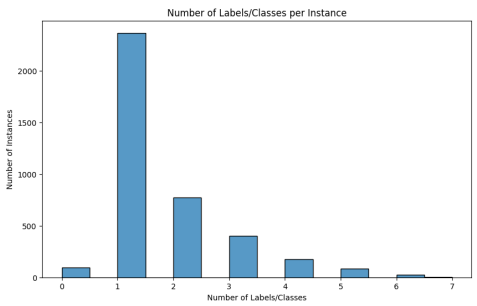

Figure 1: Number of Labels per Instance

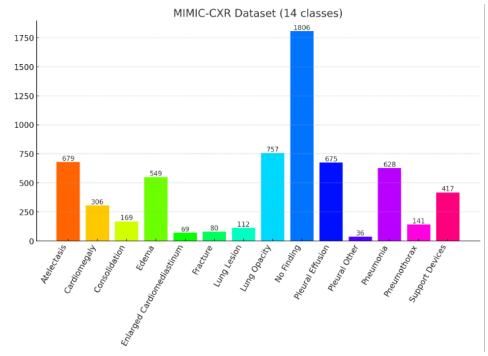

Figure 2: Class Distribution

Apart from an imbalance in individual class instances, there is also an obvious imbalance in class co-occurrence, as shown by the label co-occurrence matrix in Figure 3, which might be partially explained by disease pathogenesis.

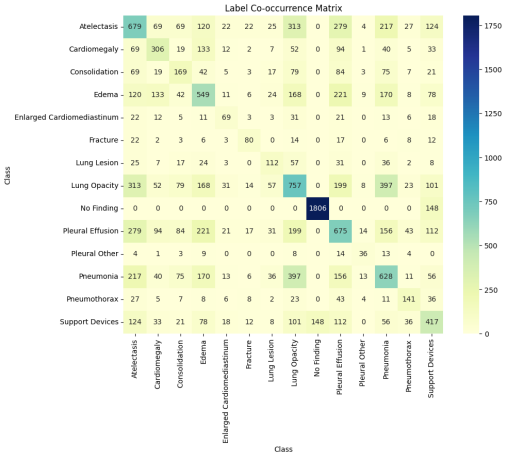

Figure 3: Class Co-occurrence Matrix

The inherent class and label co-occurrence imbalance, common in medical datasets where some conditions are naturally rarer than others, raised the need for considerate model design and evaluation metric selection in order to create multi-label classifiers that are fair and resistant to biases. In our study, evaluation metrics such as class-wise and weighted F1 scores were used to minimize the effect of potential model biases.

4 Methods

4.1 Unimodal Classification

To observe the improvement and differences of using different multimodal fusion techniques, single modality classification models are used as baselines for comparison.

4.1.1 Image Only Model

The image data is sourced from MIMIC-CXR-JPG [Johnson et al., 2019b]. This dataset contains standard JPG images derived from the original Digital Imaging and Communications in Medicine (DICOM) format in MIMIC-CXR. To expedite training, images were down-scaled to 224×224 pixels. Image augmentation techniques, including random crop and flips, were employed to avoid over-fitting in the training set.

We decided to implement the image modal classification using DenseNet-121, as it generates the overall best performance across 14 different labels. DenseNet-121 consists of 121 layers. As shown in Figure 4, the architecture is characterized by three main layers: dense blocks, transition layers, and a classification layer. They have a unique connectivity pattern where each layer is connected to every other layer in a feed-forward fashion. This dense connectivity results in feature reuse and facilitates gradient flow throughout the network, which can lead to improved training efficiency and model performance. The output of the dense net is eventually fed into a sigmoid non-linearity to calculate the probabilities for each label. The threshold that turns the projected probabilities into a class label 1 and 0 is set to 0.5 by default. During the training phase, we used the following unweighted binary cross entropy loss sum and $\hat{y}_i = 1$ indicates certain disease label exists and 0 otherwise: $L(X, \hat{y}) = \sum_{i=1}^{14} -y_i \cdot \log p(\hat{y}_i = 1|X) - (1 - y_i) \log p(\hat{y}_i = 0|X)$.

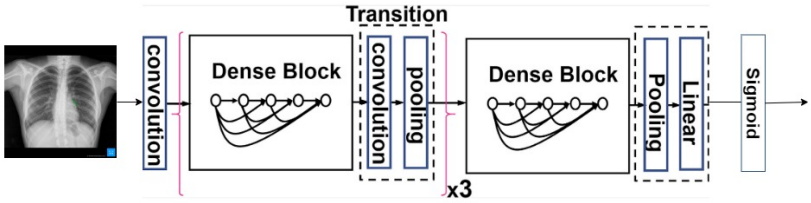

Figure 4: Image classification model using DenseNet-121

4.1.2 Text Only model

The second modality comprises textual data extracted from radiology reports in MIMIC-CXR. Specifically, we focused on the "Impression" and "Findings" sections, which contain radiologists' descriptions of X-ray films and primary diagnoses. Standard natural language processing procedures, including tokenization, stop-word removal, and stemming, were applied for data preprocessing. To mitigate label leakage, sentences were filtered to exclude any mention of words associated with the 14 labels. To conduct text classification, we fine-tuned a pre-trained BERT transformer from Devlin et al. [2018], followed by a linear layer and the sigmoid activation function. The complete model is shown in Figure 5. We chose the base-uncased BERT over specialized models like CheXbert to ensure our model generalizes well across diverse writing styles in radiology reports, which can vary greatly between medical professionals. Specialized models might carry a sample bias from their training data, limiting their generalizability. Moreover, BERT's training on a wide range of language data allows it to handle not only current expert-written texts but also more conversational language, such as patient-doctor dialogues, which may open more avenues for future studies.

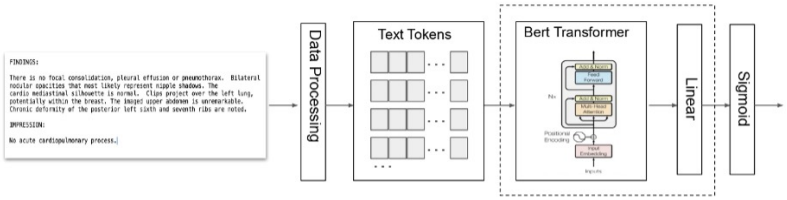

Figure 5: Text classification model using BERT

4.2 Multimodal Classification

One of the essential parts of multimodal learning is the fusion algorithm. There are many types of fusion, and the choice of the fusion method depends on factors like the characteristics of the data, the specific task, and the goals of multimodal learning. Our work aims to perform an empirical comparison among common fusion techniques.

Multimodal fusion refers to the integration of information from multiple modalities. It can be used to leverage diverse sources of information to provide a more comprehensive view of patients' conditions. In general, a fusion model for a multi-label classification task can be formulated as follows:

$$\hat{Y} = f_{class}\left(f_{fusion}\left(Enc_1(X_1), Enc_2(X_2), ..., Enc_m(X_m)\right)\right) \tag{1}$$

- $X = \{X_1, X_2, ..., X_m\}$ are the input data from m number of modalities.
- Enc_i are encoding functions, which respectively transform the input data from each modality into latent representations.
- f_{fusion} is the fusion function, which combines the encoded representations from different modalities into a joint representation.
- f_{class} is the classifier that produces the final predictions. The classifier is typically a multi-layer neural network ending with an activation function (sigmoid in our models).
- $\hat{Y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_k]$ is the output, where $\hat{y}_i$ is the predicted probability for the i-th label. The final prediction is obtained by applying a threshold to these probabilities to determine the presence or absence of each label.

For this study, we implemented three fusion models based on different fusion stages: early, late and intermediate. Early Fusion combines features or feature representations at the input level before feeding them into the model. Late fusion, which is also referred to as decision-level fusion, aggregates the outputs from different single-modality models to arrive at a final prediction. There are various ways to combine the outputs, including averaging, voting, and other more complex strategies. Intermediate fusion offers a balance between early and late fusion. Theoretically, It allows the model to learn a joint representation that captures the synergies between modalities while still benefiting from the separate processing of each modality.

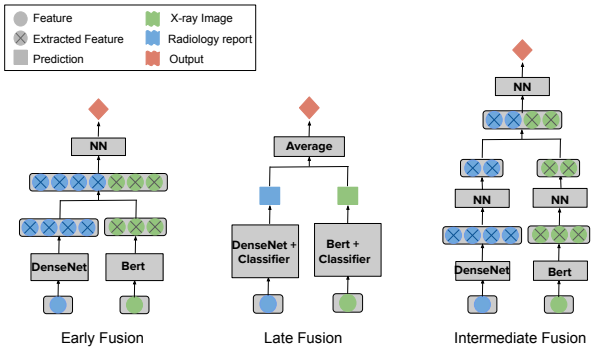

Figure 6: Fusion model architectures. The three main fusion architectures built for this project, including (a) Early Fusion, (b) Late Fusion, and (c) Intermediate Fusion.

Figure 6 shows the details of the models we experimented with in this work. The very first block of the three models is common. It's used to learn the encoding of the two modalities, text and image data, as they are not readable as raw data. The text encoding is extracted by BERT, while the image encoding is the learned representation from the last layer of DenseNet-121. Our **Early Fusion** model is a fully connected neural network. Image and text data are concatenated at the feature level into a single representation before entering the classifier. **Late (Average) fusion** involves training separate classifiers for image and text encodings and takes an average of their outputs, which are the probabilities for the 14 diseases. **Intermediate fusion** has an extra block of neural network compared to early fusion. Since the image encoding is longer than the text encoding in dimension (1000 to 768), the image encodings may take up a greater proportion of the concatenated features. Therefore, intermediate fusion may leverage the extra layers to work as a dimensionality reduction, matching text and image representations' dimensions (100) before concatenation.

5 Results

All models were trained until convergence using a learning rate of 0.00001. To evaluate the models' performance, we examined multiple metrics, including overall and class-wise averages of Precision and Recall as shown in Table 1 as well as Accuracy, F1, and AUROC score as shown in Table 2. We reported both the unweighted and weighted metrics. Since weighted measures considered label support, they may give a more balanced interpretation. The performance analysis yielded some notable findings.

Model	**Micro**		**Weighted**	
	Precision	**Recall**	**Precision**	**Recall**
BERT + Classifier	0.91	0.82	0.89	0.82
DenseNet + Classifier	0.63	0.26	0.22	0.26
Early fusion	0.89	0.74	0.85	0.74
Late (Average) fusion	0.93	0.78	0.90	0.78
Intermediate fusion	0.89	0.74	0.85	0.74

Table 1: Unimodal and Multimodal Experiment Results - Precision & Recall

Model	**Unweighted**			**Weighted**		
	Accuracy	**F1**	**AUROC**	**Accuracy**	**F1**	**AUROC**
BERT + Classifier	**0.973**	**0.683**	**0.812**	**0.957**	**0.859**	**0.906**
DenseNet + Classifier	0.902	0.050	0.510	0.796	0.237	0.547
Early fusion	0.960	0.460	0.703	0.948	0.774	0.855
Late (Average) fusion	**0.969**	**0.554**	**0.739**	**0.955**	**0.819**	**0.875**
Intermediate fusion	0.962	0.469	0.707	0.952	0.784	0.863

Table 2: Unimodal and Multimodal Experiment Results - Accuracy & F1 & AUROC

The BERT-based text model excelled as the best unimodal predictor, outperforming the DenseNet image model across all evaluation metrics with a weighted F1 score of 0.859. Demonstrating balanced precision and recall across classes, the BERT model proved effective for clinical diagnosis. By contrast, the DenseNet-based image unimodal model performed poorly across classes, struggling to identify any positive cases for several classes. This deficiency highlights the challenges of extracting and learning features from image data alone, underscoring the need for better image representation. The low performance of the DenseNet classifier may be related to constraints in the training set volume as well.

Among multimodal models with different fusion techniques, the Late Average fusion model stood out with the highest overall weighted F1-score of 0.819, slightly shying from the performance of the BERT model. When compared to early and intermediate fusion models, it appeared to harmonize the text and image data efficiently with higher precision and recall. All fusion models outperformed the image-only DenseNet classifier, suggesting that integrating text and image data may enhance predictive accuracy, potentially by leveraging text data to offset the limitations of image-based models.

Hypothetically, the fused representations may be highlighting text representations to counteract the noise in the image encodings, improving the overall performance compared to the unimodal image model.

6 Discussion and Limitation

In our analysis, several limitations may be present. Although we created fused representations between text and image encoding, there isn't a determined way to standardize the representations so that comparison studies could be performed to understand the influence of each uni-modal input's influence on the final classification output. Given that the late fusion model was the best-performing fusion model, a hypothesis is that the initial encoding outputs from both the BERT and DenseNet encoding pipelines may have been uncorrelated, thus requiring the downstream classification layers to "standardize" the encodings into the same prediction output. In other words, instead of using pre-trained BERT and DenseNet, it may have been beneficial (especially for early and intermediate fusion models) to fine-tune the encoding pipelines to output similar encoding representations for both text and image. This approach may be similar to the CLIP model by Radford et al. [2021], where both image and text encodings are projected to the same latent space, and their dot product could be interpreted as a similarity score.

Another drawback of the data processing pipeline is the masking of label words to prevent label leakage. For each of the 14 disease labels, we performed targeted keyword removal in the clinical text data. This may have reduced the readability of the clinical text, which may have weakened the performance of the text-processing pipeline. Another key aspect of this limitation is that by masking the target labels, potential information about other labels secondary to the primary classification may have been removed. Given that the model is trying to perform multi-label (instead of multi-class) prediction, this may have negatively impacted the signal of text encodings from the BERT pipeline, which may have in turn lowered the performance metric for the text unimodal model as well as all fusion models. To resolve this issue, the meaning of the clinical text should be maximally preserved while removing the labels. This may be accomplished using a report-generating large language model (e.g. LLAMA, GPT-4) to create summarized clinical reports which would not only remove occurrences of target labels but also preserve the meaning and congruence of the text by rephrasing sentences that contained the target labels [Touvron et al., 2023, OpenAI, 2023]. Extensive prompt tuning may be needed to achieve this effect.

Finally, although close, we notice that the uni-modal text model outperformed all three fusion models. We suspect that this had to do with the low performance of the DenseNet model. In a previous study, DenseNet-based models have been used for multi-label Chest X-ray analysis and have shown strong performance [Bhusal and Panday, 2023]; however, the study employed a subset of over 90,000 examples. Given the timeline and computing resources provisioned for this study, our models were trained using a much smaller dataset (3921) which may have resulted in high model bias, resulting in poor classification performance. In this respect, the model could be improved by augmenting the training set. However, considering that the models were only trained on the MIMIC-CXR dataset, whose records were all compiled from Beth Israel Deaconess Medical Center in Boston, MA [Johnson et al., 2019a], the generalizability of the study is not yet established as the models have been exclusively trained on radiology images and clinical texts that are standardized under only one institution. To improve the generalizability of the study, it would be beneficial to augment the dataset by incorporating data from multiple sources or perform a meta-analysis of fusion model performance over several datasets in a follow-up study.

7 Conclusion

Our findings underscore the BERT-based text model's robustness in disease classification using clinical notes, significantly outperforming the DenseNet-based image-only model. Moreover, fusion models that combine text and image data, especially the Late Average fusion, have considerable potential for leveraging textual information to bolster the reliability of image-based diagnosis. Future improvement may focus on two aspects: understanding the image model's poor performance, which may be attributed to a limited sample size or low image resolution from downscaling; and improving the image model's feature extraction capabilities. Furthermore, we suspect that the better performance of the late fusion model than other fusion models may be due to the uncorrelated nature of the text

and image encodings, suggesting a need for standardization before combining them: a hypothesis that requires confirmation. We aim to continue probing into how different fusion methods work, with a particular focus on unraveling why the Late Average fusion model is effective. Understanding this will help us refine our fusion techniques even further and enhance their applications in Chest X-ray classification and related medical diagnosis problems.

8 Team Contributions

The project has seen a well-distributed effort from all team members, ensuring a balanced contribution. Zhe and Chenghao dedicated their primary focus to training and fine-tuning the models, while Tianyu and Yujie mainly worked on in-depth data analysis and thorough performance evaluation. Our collaborative approach included weekly meetings, which allowed us to maintain timely progress and foster joint discussions regarding our results. Additionally, everyone contributed to crafting the presentation and composing the research report.

References

R. Adams, K. E. Henry, A. Sridharan, H. Soleimani, A. Zhan, N. Rawat, L. Johnson, D. N. Hager, S. E. Cosgrove, A. Markowski, et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature medicine*, 28(7):1455–1460, 2022.

D. Bhusal and S. P. Panday. Multi-label classification of thoracic diseases using dense convolutional network on chest radiographs, 2023.

G. Chauhan, R. Liao, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, and P. Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 529–539. Springer, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147, 2020.

F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019a.

A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019b.

T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *Irbm*, 43(1):62–74, 2022.

M. Moradi, A. Madani, Y. Gur, Y. Guo, and T. Syeda-Mahmood. Bimodal network architectures for automatic generation of image annotation from text. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 449–456. Springer, 2018.

OpenAI. Gpt-4 technical report, 2023.

Public Health Agency of Canada. Covid-19 epidemiology update: Summary, May 2023. URL `https://health-infobase.canada.ca/covid-19/`.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.