Multimodal approaches for pleural effusion diagnosis by leveraging X-ray images and clinical reports using deep neural networks and transformer-based architectures

Tony Ding
Massachusetts Institute of Technology
tding@mit.edu

Abstract

Chest X-rays are among the most commonly ordered imaging tests. Applying deep learning techniques to X-ray images is a typical application of computer vision in healthcare. Nevertheless, using X-ray images alone does not always lead to decent and generalizable model performances. Combining patients' clinical reports, which contain rich and important patient diagnostic information, with X-ray images could give the model more information for prediction. As a result, our work will focus on deriving and examining the best fusion strategies for implementing a multimodal approach with regard to pleural effusion prediction. Using X-ray images and clinical text reports, we mainly combine VGG16 with DistilBERT to better predict the presence of pleural effusion. We propose two sets of fusion strategies, namely early fusion, where we concatenate the learned vector representations of images and texts before classification, and late fusion, where we leverage the predicted probabilities from the two modalities. Ultimately, we found that the late fusion multimodality model with an elastic net regularized logistic regression model achieved the best overall performance, with an AUC value of 0.9887. On the other hand, the early fusion strategy achieved inferior results, which indicates that the early fusion strategy that we utilized here is not specifically suitable for integrating X-ray images with clinical text data.

1. Introduction

Pleural effusion refers to an abnormal accumulation of fluid in the pleural space (Light, 2002). It is beneficial to diagnose pleural effusion early in time because it can lead to medical complications such as difficulty in breathing, chest pain, and reduced lung and heart function (Light, 2002). In addition, the diagnosis of pleural effusion usually relies on medical imaging technologies such as X-ray imaging (Karkhanis & Joshi, 2012).

Moreover, in recent years, deep learning techniques have been frequently applied to clinical image data to improve the accuracy and efficiency of diagnoses. Convolutional neural networks(CNNs) have been particularly successful in processing clinical imaging data, such as X-rays and CT scans (Tang et al., 2020, Kshatri & Singh., 2023). Past research has also shown that CNNs can be used to accurately classify clinical images and detect abnormalities, such as tumors, fractures, and pneumothorax, with performance that is comparable to human experts (Tang et al., 2020, Rajpurkar et al., 2017). In addition, natural language processing techniques based on transformer architectures, such as the BERT model, have also shown remarkable success in processing unstructured textual data, such as electronic health records (Alsentzer et al., 2019).

The ability to process textual data can provide complementary information to the analysis of medical images. This means a multimodal approach that can leverage both image and clinical textual data may provide a more comprehensive understanding of patient health status and potentially improve diagnostic accuracy. While there have been significant advances in both CNN and transformer-based approaches, there has been very few multimodality model that combines X-ray images with clinical texts or reports, which can be quite messy sometimes due to clinical abbreviations and text modifications due to de-identification purposes. This also means there has been very little work on deciding what's the best fusion strategy for combining these two specific modalities of data.

In this paper, we propose two fusion strategies for building such a multimodality model. The first method, generally categorized as an early fusion approach, leverages the features extracted from the images using pretrained CNN models and the features extracted from the clinical text reports from pretrained transformer-based models by concatenating them before the final classification layers. This can potentially allow for a better modeling of the interactions between the features learned from different modalities and a more comprehensive and holistic representation of the input data. Another approach, generally categorized as the late fusion approach, uses the final predicted probabilities from the two models, each trained on a different modality, and then further trains these output probabilities using an additional generalized linear model to obtain our final classification results. This approach can potentially preserve the unique characteristics and nuances of each modality, which may help the model capture fine-grained information specific to each modality, thus leading to a more comprehensive

understanding of the input data. We will compare the performances of these two approaches to determine the better fusion strategy for these specific modalities.

Lastly, to achieve interpretability in a multimodal network is also important. This can enable healthcare professionals to comprehend the rationale behind the model's discernment of a specific diagnosis. Given that this is a computer vision course, we will mainly be focusing on generating visual interpretability for our X-ray imaging data using techniques such as saliency map and Gradient-weighted Class Activation Mapping (GradCAM) (Simonyan et al., 2013, Selvaraju et al., 2017).

2. Related Work

There have been quite a few attempts of using computer vision based approaches and natural language processing techniques for obtaining classifications of clinical images. Rajpurkar et al. (2017) compared the performance of their CheXNeXt algorithm to practicing radiologists on the ChestX-ray14 dataset and found that the algorithm outperformed the radiologists on certain diseases. In addition, natural language processing techniques, particularly transformer-based approaches, have also shown promising results in clinical text processing. Alsentzer et al. (2019) introduced publicly available clinical BERT embeddings, which has achieved great results on baseline text data and can be used for various clinical natural language processing tasks. Wang et al. (2020) incorporated domain knowledge into a clinical transformer for clinical named entity recognition and achieved state-of-the-art performance on benchmark datasets, achieving a micro F1-score of 90.8%. Lastly, Huang et al. (2020) used multimodal fusion by leveraging both CT imaging data and electronic health records data to predict pulmonary embolism. Their best performing multimodality model, which is a late fusion model, achieved an AUC of approximately 0.947 and outperformed both their early fusion models and single modality models. This paper laid the foundation for deploying multimodality models on clinical data of different modalities. In short, these works have all shown promising results with regard to the techniques that we discussed, and they can all help us better understand and perfect the use of both computer vision and natural language processing techniques for analyzing clinical images and texts.

3. Methodology

3.1 Hypothesis and Challenges

Given the rich, multidimensional information embodied within textual data, we hypothesize that an integrative approach, where we combine the analysis of both image and text data, may enhance the overall prediction accuracy or AUC value compared to just using the image data for prediction.

To understand what specific words contribute to the model's predictions, we initiated a comprehensive review of the clinical text reports. These reports, crafted for chest X-rays with 13 descriptive labels, cover a wide range of pulmonary diseases. Pleural effusion, a common condition, features prominently in many reports either as 'pleural effusion' or 'no pleural effusion'. However, due to its frequent occurrence, the mere mention of 'pleural effusion' does not serve as a definitive diagnostic indicator.

To decipher the terms that influence the prediction of pleural effusion, we undertook a thorough analysis of the terms used in its classification. These terms were ranked based on their relative frequency, calculated as the difference between the frequency of words in reports where the predicted label is 1 and those where the predicted label is 0. This analysis would help us discern the words that are pivotal in shaping the model's predictions, thereby informing our approach to processing the textual data.

3.2 Data

Our dataset comes from the MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 and the MIMIC Chest X-ray JPG (MIMIC-CXR-JPG) Database v2.0.0 (Johnson et al., 2019). These data are collected at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The datasets include Chest X-ray images from more than 65,000 patients with over 370,000 chest X-ray images, collected over a period of 10 years (Johnson et al., 2019). The CXRs were de-identified and labeled with 13 different radiographic findings, such as pneumonia, atelectasis, and cardiomegaly. The MIMIC-CXR Database also includes associated clinical text reports and some other clinical metadata such as ViewPosition, which is the orientation in which the chest radiograph was taken. The diversity of the two datasets will contribute to the development of multimodality models for effective detection of pleural effusion.

In addition, the data packages are rather large (around 5.5TB), so we decided to only consider around 10% of the entire database.

3.3 Data Preprocessing

Our ultimate goal of data preprocessing is to derive a dataframe that contains patient_id, study_id, X-ray image (i.e. file path to the image stored in the Google Drive), clinical_report, and the label for pleural effusion. Consequently, we merged the two datasets using dicom_id in order to match and combine the clinical text reports, the corresponding JPG images, and the labels. Furthermore, we only used Anteroposterior(AP) and Posteroanterior(PA) positions of the patients' X-ray images and excluded images that are in lateral positions since images in AP and PA positions are the ones that are

relatively more commonly used for pleural effusion diagnosis (Na, 2014).

For image pre-processing, we crop the images to smaller sizes of 224x224 to make sure all images have the same size. For clinical reports, all text data is tokenized by the pertained DistilBERT model before being fed into the multimodal network. We did a train-validation-test split using a ratio of 0.7-0.1-0.2.

Lastly, the original data have 13 descriptive labels indicating 13 common lung diseases and conditions, including 'Atelectasis', 'Cardiomegaly', 'Consolidation', 'Edema', 'Enlarged Cardiomediastinum', 'Fracture', 'Lung Lesion', 'Lung Opacity', 'Pleural Effusion', 'Pleural Other', 'Pneumonia', 'Pneumothorax', and "Support Devices'. Subsequently, we filtered for the data that have a clear label(0 or 1) on pleural effusion since the dataset that we obtained contains a lot of NAs for the labels. Ultimately, we ended up with 8183 valid X-ray images and corresponding clinical reports.

3.4 Performance Metrics

During training, the performances for all of our models will be assessed using accuracy(i.e. validation accuracy). The final model's performance on the test set will be evaluated using AUROC values.

3.5 Methods

Our study design included the development of two varieties of fusion strategies for multimodality models, designed to seamlessly integrate textual data from the clinical texts with the corresponding chest X-ray images.

3.5.1 Image Analysis

For the analysis of imaging data, we utilized pre-trained Convolutional Neural Networks (CNNs) to exploit the high-level feature representations derived from models pre-trained on significantly larger datasets. After comprehensive experimentation with various pre-trained models, we selected VGG16 as our convolutional base due to its superior performance for our specific dataset. To enhance our model's ability to generalize, we implemented data augmentation techniques such as horizontal flipping, rotational transformations, and width/height shifts. Subsequent to the pre-trained VGG16 convolutional base, we introduced batch normalization, dense, and dropout layers to form our classification layers. We used the Adam optimizer for model training, alongside binary cross entropy as the loss function and an early stopping mechanism of patience equal to 15 (Kingma et al., 2014). The model with the highest validation accuracy was chosen as our final baseline model for X-ray image analysis.

3.5.2 Text analysis

For processing clinical textual data, we integrated a pre-trained transformer-based model: DistilBERT, which is known for its compact model size and efficient training time (Sanh et al., 2019). We tried to use other larger and less compact transformer-based models; however, limited by the computational resources at our hands, we weren't able to do so in time. The text data was tokenized using the 'distilbert-base-uncased' tokenizer, which was subsequently inputted into the transformer model. We set the maximum token length at 512 to adequately capture the content while maintaining computational efficiency. The Adam optimizer, with a learning rate of 1e-4 and decay rate of 1e-5, was used alongside the binary cross entropy loss function. We selected a batch size of 32 to balance model performance and computational resources. Similar to the image analysis model, the model demonstrating the highest validation accuracy was selected as our final model for the text modality.

3.5.3 Multimodal Analysis of Images and Texts

Our first approach to the multimodal architecture is using early fusion of the learned features, where we use the output of the encoder part of the DistilBERT model and take it as a representation feature and then concatenate with the output of the pretrained VGG16 convolutional base (after the flatten layer). Subsequently, we added several batch normalization layers, dense layers, and drop out layers as our classification layers for the early fusion approach.

Our second fusion strategy, generally recognized as late fusion, uses the final predicted probabilities from the two models for final prediction, where each model is trained on a different modality. After we obtain the predicted probabilities, we came up with three ways to further process or train these probabilities. The first method is a simple averaging of the probabilities produced from the two models of the two modalities. The second method further trains a linear regression model using the predicted probabilities to obtain our final classification results. The third method trains a logistic regression with elastic net penalty using the predicted probabilities for final classifications. We also used GridSearchCV to help us locate the best hyperparameters for our elastic net regularized logistic regression model.

4. Experimental Results and Discussion

4.1 Experimental Results

For our baseline imaging model, the VGG16-based CNN with data augmentation, it ultimately achieved a test accuracy of 84.92% and an AUC of 0.9099. Nevertheless, this model consistently overfitted on the training data, regardless of the application of data augmentation techniques, as shown in Figure 1a. Consequently, this indicates that this model lacks a good generalizability to unseen data beyond the overfitting point.

Furthermore, for our early fusion approach towards our multimodality model, the validation accuracy appeared to

plateau at around 81%, even under elongated training time, as shown in Figure 1b. This suggests that the early fusion strategy may not be fully capitalizing on the potential synergies between the features of the X-ray images and the clinical texts to maximize the predictive accuracy. Ultimately, this early fusion strategy achieved an AUC of 0.8791, which is slightly worse than that of our baseline model.

Our late fusion strategy achieved much more success compared to the aforementioned baseline approach and the early fusion multimodality model. To elaborate, a simple averaging of the predicted probability values from both models resulted in an AUC of 0.9804. For the regression-based late fusion approaches, by implementing a linear regression using the predicted probabilities, we were able to achieve an AUC of 0.9817. Remarkably, when we applied a logistic regression model with elastic net regularization, the AUC further increased to 0.9887, which is also our highest AUC value among all approaches and models. These results underline the effectiveness of late fusion multimodal learning in harnessing the predictive potential of both X-ray images and clinical reports for superior model performances.

All of our results described above are also listed in Table 1.

Methodology	Model AUC
Pretrained VGG-16 w/o data augmentations	0.8801
Pretrained VGG-16 with data augmentations	0.9099
Early Fusion – VGG-16+DistilBERT	0.8791
Late Fusion – simple average	0.9804
Late Fusion – linear regression	0.9817
Late Fusion – logistic regression with elastic net	**0.9887**

Table 1. Different methodologies and their corresponding AUC scores.

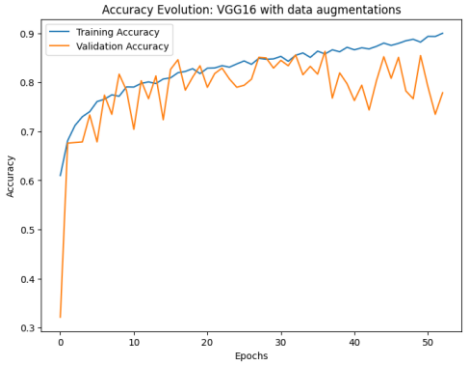

Figure 1a. Model accuracy plot for VGG16 with data augmentations.

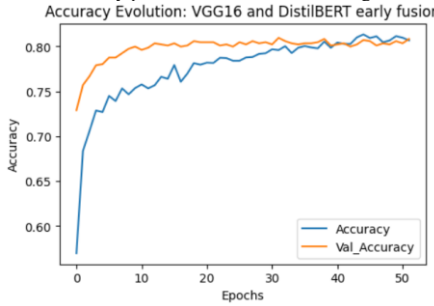

Figure 1b. Accuracy history plot for multimodality model using early fusion strategy by leveraging VGG16 and DistilBERT. VGG16(Figure 1a) overfitted and early fusion model(Figure 1b) validation accuracy plateaued.

4.2 Interpretability Results

4.2.1 Images

To interpret our vision network, we employed two gradient-based methods: saliency maps and GradCAM, shown in Figure 2. Both techniques utilize gradients to identify important regions within an image. Saliency maps calculate the gradient of the network's loss with respect to the input image and visualize these gradients as a heatmap (Simonyan et al., 2013). GradCAM, on the other hand, calculates gradients only for the last convolutional layer before the global pooling operation, resulting in a more abstract representation of input-output relationships and providing a different perspective on the network's underlying mechanisms (Selvaraju et al., 2017).

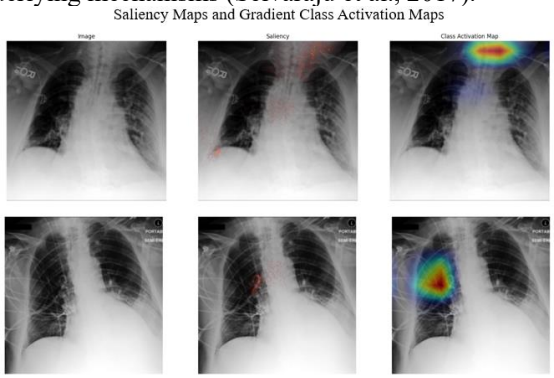

Figure 2. **Left:** Original images. **Middle:** Saliency maps overlaid on the original images. **Right:** GradCAMs overlaid on the original images. The model sometimes captures the area of the chest and sometimes captures other places like medical devices.

Both methods above highlighted similar regions. In addition, aside from the lung areas, the model sometimes pays attention to other regions as well as medical devices. While this could be advantageous for diagnosis, it is not ideal for model interpretability because the model's focus is not solely on disease locations and related pathology.

4.2.2 Texts

In an attempt to get a better idea of the contributions of specific words to our model's predictions, we also computed the relative frequency of words within the clinical reports, shown in Table 2. This measure is derived by subtracting the word frequency in reports where the label is 'Pleural Effusion' from the word frequency in reports where the label is 'Normal', and vice versa.

Word relative frequency in clinical notes

	Prediction = 1 (pleural effusion)		Prediction = 0 (no pleural effusion)	
Rank	Word	Relative frequency	Word	Relative frequency
1	right	3059	or	337
2	left	3021	appreciable	131
3	report	2923	borderline	100
4	final	2923	cardiopulmonary	89
5	__	2914	cough	89
6	is	2855	relevant	60
7	the	2840	tortuosity	50
8	and	2824	dictation	41
9	chest	2803	elongation	29
10	small	2780	hyperexpansion	25

Table 2. Word relative frequency table. Calculated using the following rule: the top 10 words that appear the most in reports that are classified as pleural effusion compared to reports that are classified as normal, and vice versa.

The above results indicate that, in our early fusion multimodality model, the clinical text data indeed aided in the classification of pleural effusion. This is because the terms 'left' and 'right', which are the top two most frequent words for a 'Pleural Effusion' prediction, are often seen in phrases like 'left pleural effusion' or 'right pleural effusion'. In contrast, when the early fusion model predicts 'Normal', some of the most frequent words, such as 'tortuosity' and 'elongation', are mainly describing other medical conditions, possibly those that do not co-occur with pleural effusion (Bharadwaj, 2022).

4.3 Discussion

Our best multimodality model used a late fusion strategy where the probability outputs of our pretrained VGG16 model and our pretrained DistilBERT model are further trained by a elastic net regularized logistic regression model. All the late fusion models achieved higher AUC values than those of the image-only models, which means the modality of clinical text aided in prediction and that the interactions between text and image outputs can improve model performances. Nevertheless, the AUC value of our multimodality model with an early fusion strategy was the lowest out of all the architectures. Consequently, given that we applied the exact same classification layer architectures for both our multimodal models and image-only models, we can reasonably infer that the early fusion strategy that we utilized is not an ideal approach for integrating these two specific modalities: X-ray images and clinical text data. On the other hand, our finding is consistent with Huang et al. (2020), where they found that, when dealing with clinical images and EHR data, a multimodality model using late fusion strategy generally yields better overall results than the models that used early fusion strategies do.

Furthermore, evidently, the early fusion strategy that we applied here appears to be the biggest limitation to our model's performance. As a result, we propose that in the future, we can adopt more complex and robust architectures that can better combine the learned features from clinical text and medical images, such as the architectures used in LXMERT. To be more specific, LXMERT uses cross-attention mechanisms that can allow for interaction and mutual influence between textual and visual modalities (Tan & Bansal, 2019). The LXMERT architecture involves three parts, namely an object relationship encoder that is a faster R-CNN, a language encoder like the pre-trained DistilBERT, and a cross-modality encoder, which is a transformer architecture that allows bi-directional attention flow to integrate text and image input (Tan & Bansal, 2019). This architecture could potentially help us better model the interactions between the features extracted from X-ray images and clinical texts.

Lastly, as mentioned in the results interpretability section for texts, terms such as 'right' and 'left' could have strong diagnostic values. Nevertheless, it is worth noting that these words may not demonstrate the same classification power in the context of multi-label pulmonary disease prediction task, such as predicting pneumonia and pneumothorax, where a mere mentioning of 'left' or 'right' will not be enough to instruct the model to make a specific disease prediction.

5. Conclusion

Overall, our study demonstrated two multimodal fusion strategies that can integrate clinical text data with X-ray image data for pleural effusion prediction. We compared their performances with our baseline VGG16 model and found that the late fusion multimodality model with an elastic net regularized logistic regression model has the best overall performance, achieving an AUC score of 0.9887. On the other hand, the early fusion strategy that concatenates learned X-ray image features and clinical text features only achieved slightly worse results than our baseline model did. Therefore, we can infer that the early fusion strategy that we utilized here is not suitable for integrating X-ray images with clinical text data. Consequently, we have shown that combining textual data with imaging data in clinical use could aid disease prediction, but to achieve better results using an early fusion architecture, a more complex and robust architecture will be needed. Future study can work on improving early fusion architectures by identifying the most effective representations of text data and how the clinical text information can be integrated with X-ray images through the interaction and consensus between them. Furthermore, the generalizability of such models is unsure and remains an open problem since we have only shown that clinical text can be supportive to the analysis of imaging data that are from the same database.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv*:1904.03323.

Bharadwaj, S., Chan, C., Choo Tze Liang, J., Sanamandra, S. K., Fortier, M. V., Koh, A. L., & Sundararaghavan, S. (2022). Neonatal Arterial Tortuosity and Adult Aortic Aneurysm—Is There a Missing Link?—A Case Report. *Frontiers in Pediatrics*, 9, 1702.

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

Huang, S. C., Pareek, A., Zamanian, R., Banerjee, I., & Lungren, M. P. (2020). 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), 1-9.

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv*:2004.00849.

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., & Horng, S. (2019). MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). *PhysioNet*. https://doi.org/10.13026/8360-t248.

Karkhanis, V. S., & Joshi, J. M. (2012). Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, 4, 31.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv*:1412.6980.

Kshatri, S. S., & Singh, D. (2023). Convolutional Neural Network in Medical Image Analysis: A Review. *Archives of Computational Methods in Engineering*, 1-18.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., ... & Lu, H. (2018). Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*, 296(2), E65-E71.

Light, R. W. (2002). Pleural effusion. *New England Journal of Medicine*, 346(25), 1971-1977.

Na, M. J. (2014). Diagnostic tools of pleural effusion. *Tuberculosis and respiratory diseases*, 76(5), 199-210.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv*:1711.05225.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv*:1910.01108.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv*:1312.6034.

Tabik, S., Gómez-Ríos, A., Martín-Rodríguez, J. L., Sevillano-García, I., Rey-Area, M., Charte, D., ... & Herrera, F. (2020). COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. *IEEE journal of biomedical and health informatics*, 24(12), 3595-3605.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv*:1908.07490.

Tang, YX., Tang, YB., Peng, Y. et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Med*. 3, 70 (2020). https://doi.org/10.1038/s41746-020-0273-z

Wang, Y., Huang, C., Peng, Y., Yang, Y., & Huang, S. (2020). Incorporating domain knowledge into clinical transformer for clinical named entity recognition. *Journal of biomedical informatics*, 107, 103474.