**ARTICLE**

# MEDICAL IMAGE DESCRIPTOR FOR REPORTING AID

Gomez Jorge, Tlacuilo Jose, Vargas Oscar.

Having an efficient diagnosis in occasions involves having a specialist to check medical exams like X-Ray imaging, MRI, or CT scans making this process long and enclosed. The creation of methods of Deep Learning applied into making this process easier have been emerging into creating even a method that relates images to sentences like OpenAI.com created the CLIP system which was used and modified in this paper. The model created, MeDCLIP, was trained with a curated data set from MedPix® and other sources to guarantee the best results in different metrics used as: ROUGE, BLEU, METEOR, and CIDEr. The program performed accurately in generating captions related to the reference in an image to function as a guide for medical diagnosing.

## INTRODUCTION

Diagnosis based on medical imaging is a powerful tool towards early disease treatment[1]. The process depends mainly in the technologies used, such as MRI, X-Rays, CT, Ultrasound, or others. When the image is obtained, the next step is medical interpretation of the results to give a guide or reference towards a medical specialty. Although in most of the cases there is a correct diagnosis, there are other cases in which incorrect interpretation causes diagnosis-treatment dates shortening[2] and even inadequate prescriptions given to the patient[3]. Usage of new methods, to avoid incorrect diagnosis, are a need and are being developed[4-8]. Recent methods not only include medical interpretation but machine evaluation[6-8]. Success in the new methods may be measured by its accuracy in detecting the illness on time[2] but it might not be the most efficient way to do it. Other methods, based on Deep-Learning[9-13], have achieved a considerable accuracy when helping in the medical diagnosis.

DL (Deep-Learning) techniques have acquired more popularity with new computer resources[14]. Programming in multiple codes with libraries like Pytorch[15], TensorFlow[16] and Keras[17], amplify the possibilities when using any programming language. Data processing may be a complicated task in low-based languages as C, therefore other languages as Python o R were developed to bring a more user-friendly environment. Researchers have preferred Python over other languages due to its high-level easy use syntax, multiple libraries and its capacity to integrate with other languages[18]. Specific DL techniques for this language are being developed more and more, and even combining these methods to develop a more complete one as CLIP[19].

CLIP, Contrastive Language-Image Pre-training, is not a traditional method of image or natural text classification, it instead learns the relationship between a sentence and an image[19]. The CLIP model encodes an image and a text to a vector, then it takes the dot product between them to find the similarities to learn the most relevant characteristics of each part[19]. The specifics of this process may be revised in the original article, but a simple flowchart is shown at Figure 1.[19] The original encoding of text and image in CLIP is made by ResNet and DistilBERT, but other encoders might be used to perform better in specific training datasets. The input to this kind of program may be an image to obtain a caption or a caption to obtain an image making it very different to common DL traditional techniques. Applying this method to medical diagnosis can become an important step to perform better at this task since accuracy can be measured now.

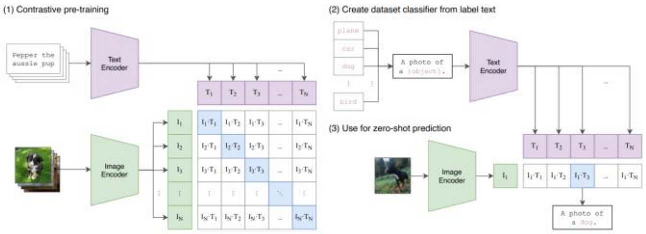

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
$T_1$ $T_2$ $T_3$ … $T_N$
$I_1$ $I_2$ $I_3$ ⋮ $I_N$
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ … $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ … $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ … $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ … $I_N{\cdot}T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
$I_1$
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ … $I_1{\cdot}T_N$
A photo of a dog.

**Figure 1.** Diagram of CLIP model from the CLIP article[19].

Accuracy methods focused on AI caption generation compare the output of the model to a reference taking different characteristics[20]. The methods used in this article, and some of the most popular, are BLEU and ROUGE with their variants, METEOR, and CIDEr. The variants of the first two take the relationship between the words in a sentence and their order[20] while METEOR finds the coincidence 1:1 of the words in a sentence with the possibilities of synonyms and stemming. CIDEr, by the other way, learns how a human would write a sentence to then use that information to analyze the sentence as METEOR does[20].

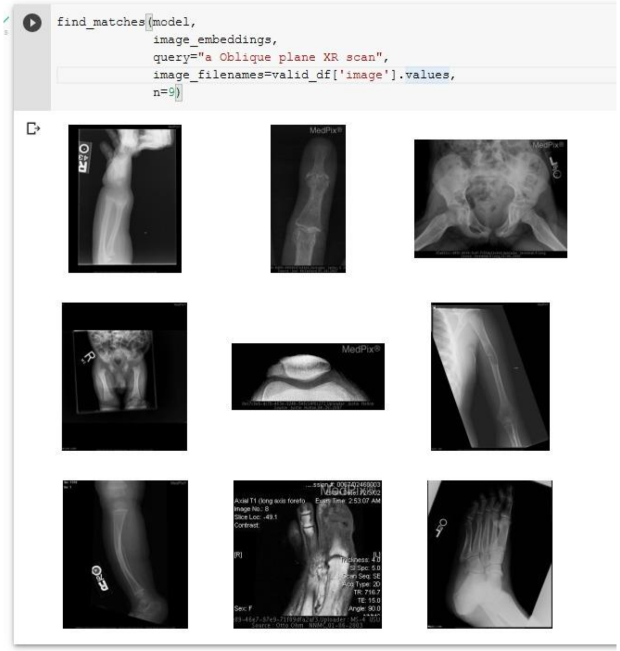
The code used in this article is based in the CLIP model, using Bio_ClinicalBERT[21] as our text encoder and ResNet50 for the image encoder. The dataset used was from MedPix[22] which has a different set of medical images with their corresponding diagnosis captions. Implementation of image and text generation were added to the base code of CLIP to then measure by ROUGE, BLEU, METEOR and CIDEr, the results obtained.

## RESULTS

The results section is divided into 3 parts depending on the output obtained.

### Image generation

The generation of images was the first milestone. The model received a text with the same structure as most of the captions obtained from the dataset. The image obtained from the input "a Oblique plane XR scan" is in the Figure 2. where the output images correspond exactly to what the input was referring to. An oblique plane is a plane which is not totally horizontal or totally vertical.



```
find_matches(model,
             image_embeddings,
             query="a Oblique plane XR scan",
             image_filenames=valid_df['image'].values,
             n=9)
```

**Figure 2.** Image generator with input "a Oblique plan XR scan"

The next query used as input was "a frontal XR scan of spine trauma" which in this case we made a comparison between DistilBERT and ClinicalBERT to identify the accuracy between one encoder and the other. In Figure 3. there is a comparison between the two results where the left output is DistilBERT and the right output is ClinicalBERT.

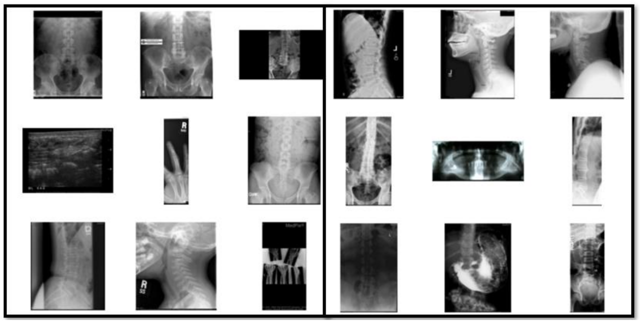
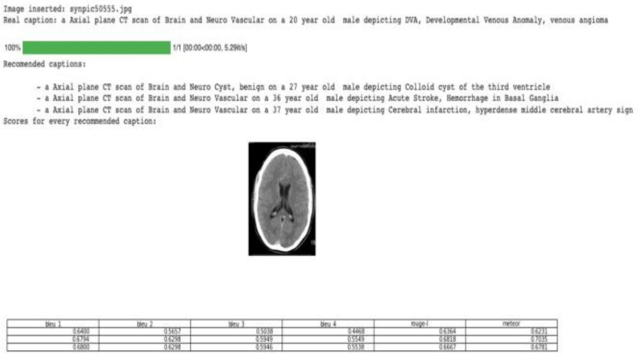


**Figure 3.** Comparison con image generator using DistilBERT (left) and ClinicalBERT (right)

Comparing the output between a model trained with a medical text encoder and a general text encoder there is a difference in accuracy. Using DistilBERT there is a 66% of accuracy since only 6 out of 9 images are correct. On the other side, the encoder ClinicalBERT had 100% accuracy because all the images correspond to the input introduced.

### Text Generation

Implementation of the zero-shot part was the trickier task since the output was not always the correct diagnosis even though the metrics marked a correct description. After modifying the code to obtain the 3 best captions based more on Rouge-L metric we obtained an output as in the Figure 4. where the original caption of the image is compared to the generated captions.



```
Image inserted: synpic50555.jpg
Real caption: a Axial plane CT scan of Brain and Neuro Vascular on a 20 year old  male depicting DVA, Developmental Venous Anomaly, venous angioma

100% 1/1 [00:00<00:00, 5.29it/s]
Recomended captions:

        - a Axial plane CT scan of Brain and Neuro Cyst, benign on a 27 year old  male depicting Colloid cyst of the third ventricle
        - a Axial plane CT scan of Brain and Neuro Vascular on a 36 year old  male depicting Acute Stroke, Hemorrhage in Basal Ganglia
        - a Axial plane CT scan of Brain and Neuro Vascular on a 37 year old  male depicting Cerebral infarction, hyperdense middle cerebral artery sign
Scores for every recommended caption:
```

| bleu 1 | bleu 2 | bleu 3 | bleu 4 | rouge-l | meteor |
|---|---|---|---|---|---|
| 0.6400 | 0.5657 | 0.5038 | 0.4468 | 0.6364 | 0.6231 |
| 0.6794 | 0.6298 | 0.5949 | 0.5549 | 0.6818 | 0.7035 |
| 0.6800 | 0.6298 | 0.5946 | 0.5538 | 0.6667 | 0.6781 |

**Figure 4.** Text generator output from an image with reference caption: a Axial plane CT scan of Brain and Neuro Vascular on a 20 year old male depicting DVA, Developmental Venous Anomaly, venous angioma.

It is interesting how the generated text creates a "diagnosis" which could be given by any physician with no more information but the medical image. It is important to remember that the program is generating a medical guide so any details may be considered when creating a diagnosis. The output generated gives the possibilities of what could be what it is shown in the image with no more information nor other medical tests.

### Comparison between metrics

To evaluate the accuracy of our model, 40 samples were made and the mean of each of the metrics can be seen in Table I.

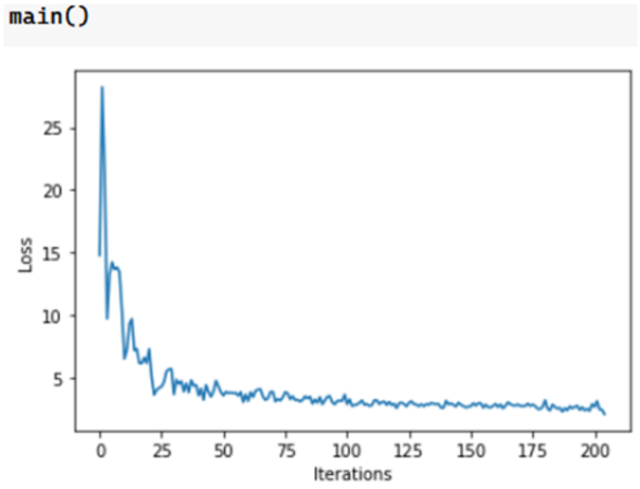
**Table I.** Mean metrics of 40 samples

| Bleu 1 | Bleu 2 | Bleu 3 | Bleu 4 | METEOR | ROUG-L | CIDEr |
|---|---|---|---|---|---|---|
| 0.5667 | 0.4779 | 0.4016 | 0.3474 | 0.2674 | 0.5824 | 0.4278 |

The metric which got the best score was ROUGE-L due to the high similarities found in dot product analyzing the whole sentence and not only the match in words. In second place the Bleu 1 method had a similar score. In third place the Bleu 2 had a unit value lower. In fourth place is CIDEr and in fifth is Bleu 3. The lowest metric values were for Bleu 4 and METEOR. The important values of metrics for the objective of this project are the metrics which measure having the same words, or synonyms, in the sentence with not necessarily the same order as in the reference to hace a better understanding. The metrics which best fit this description are Bleu 1 or 2, METEOR, ROUGE-L and CIDEr. Further analysis of the best metrics is done in the next section.

### Loss function

The perfect relationship between encoded images and captions is described by their encoded representations being the same. This similarity can easily be measured by looking at the softmax between the dot product of the encoded inputs; a perfect encoding will yield the identity matrix. The loss on each iteration is calculated using cross entropy on the dot product between the encodings and this loss may be graphed per epoch as presented in Figure 5.



```
main()
```

**Figure 5.** Graph of the loss function for 200 iterations.

## DISCUSSION

The results obtained from the image generator was an easier task to do since the method gives the best 9 results using specific encoders for medical purposes as ClinicalBERT[21]. The more complicated the input query the less accuracy the model may have since the match to captions of images would be less similar with more descriptions. Therefore, as mentioned before, this section might work as a guide for medical purposes only to narrow the possible images of how the pathology might be represented in an image.

Zero-shot method based in similarities between dot product and gave great results compared to the first method used which only was to pick to most similar images and give the best words that fit most of the captions in these images, such as BLEU or ROUGE 1 method might had been. ROUGE-L, METEOR and CIDEr have a more centered in meaning process to not only search for restricted order captions but to also include matches with synonyms and whole sentence sense. This model got ROUGE-L higher scores which is a good sign that generated captions and reference are very similar. Therefore, the best metrics to describe the real accuracy of the program may be, as mentioned before: ROUGE-L, BLEU 1, METEOR and CIDEr. BLEU 1 describes 1:1 word in the reference and the output not necessarily in the same order. METEOR on the other hand is analyzing the same as BLEU 1 but with the possibility of stemming and synonyms. The reason why this metric got so low score might have been due to the narrow medical language for each image that is used in MedPix and the lack of other references to "learn" more words. CIDEr method does the same as METEOR but in this case is a more accurate measurement since it considers how sentences are constructed. If a final accuracy score might me given to the model it would be the score of ROUGE-L and CIDEr. The final caption outputs give a great sense of what the diagnosis might be based on the image.

This method was never meant to be a replacement of medical diagnosis but a guideline that considers little details that could be ignored when giving a final diagnose. This program fulfills the main objective of this project by a considerable accuracy compared to other methods such as State of the art article[22].

## METHODS

### Dataset

For purposes of this project, the MedPix[23] dataset was used. MedPix consists of nearly 59,000 commentated medical images, categorized by disease, with quantities as in Figure 6. and with a rich text caption accompanying each image. The dataset was narrowed down to include only X-Ray, Magnetic Resonance and MRI images as seen in Figure 7; additionally, only the Locations with more than 354 image samples were selected. Finally, outliers and irrelevant images were removed to total the image count to 26384.

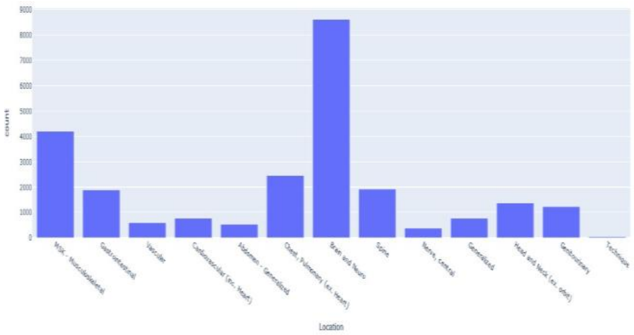


**Figure 6.** Disease distribution of MedPix dataset.

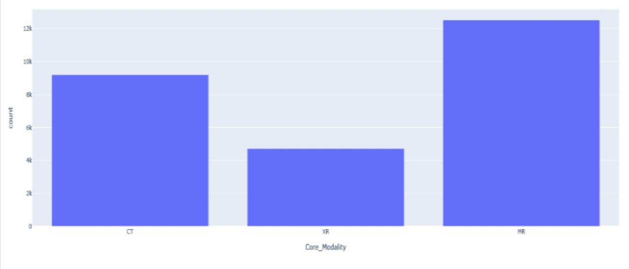


**Figure 7.** Medical imaging quantities in MedPix dataset

## Encoders

As described in Radford et al.[24] an approach of jointly training an image encoder and a text encoder to predict the correct pairing of a batch of image and captions was implemented. The Image encoder used a pretrained ResNet 50 model to encode each image to a 2048 size vector. Additionally, the text encoder was done using a BERT model pre trained on the database MIMIC III of around 880 million words, that contains health records for ICU patients at the Beth Israel Hospital in Boston, MA[25]. As with the image encoder, the text encoder returned tokens of fixed size vectors, in this case of size 768. As explained in Radford[24] a pairing of image, caption was aimed for. Thus, the unequal dimensions for images and text encoding were projected into 256 dimensional vectors, so that they could be compared.

## Zero-shot

Parting from the assumption that the dot product of two vector encoded entities reflects their similarity, a simple "hashing" system can be devised for an input, that when encoded and matched to a known set of possible matches, results in the most similar one being obtained. This process can be utilized in either direction to caption an image with all the known encodings, or search for an image given a caption.

## Metrics used

Metrics are the best way of measuring the accuracy of caption generation models. Every kind of metrics has a different mathematical method to demonstrate the similarities between a generated caption and a reference. For instance, BLEU searches for the similarities in the two sentences by 1 word, 2, 3 or 4 consecutive words called n-grams[20]. This metric does not check the meaning of the sentence but only coincidences in words and consecutive n-grams. By the other way, ROUGE-L has a more trustworthy method because it finds matches in words as in n-grams but there can be other words in between[20].

Other metrics such as METEOR and CIDEr are more centered into language interpretation because of the stemming and synonyms that they can use. METEOR learns these two characteristics of the vocabulary and therefore gives a better result when trained with the sufficient vocabulary. On the other hand, CIDEr does the same as METEOR but learns how human sentences are written and learns this algorithm to give a better score when not using a specific order, or same words.

The best metrics for this model were BLEU 1 because of the matches 1:1 of each word with the reference; ROUGE-L because it checks the order and coincidence of the words with other vocabulary in the middle; CIDEr because it considers the structure of human sentences and gives a better perspective of "meaning" rather than "coincidence". The other methods such as BLEU 2-4 will have lower scores because to always the diagnoses are written in a specific order with no words in between; METEOR will not be the best because we would need a great set of medical synonyms or vocabulary to have a better score.

## DATA AVAILABILITY

MedPix[23] is an open database that anyone can access and download the desired images with any method desired. In this article we base mainly in X Ray, CT, and MRI. The dataset was then cleaned, and the captions depicted to give a perspective of the medical image, and one of the possible diagnose.

## CODE AVAILABILITY

The code can be accessed in this GitHub link: https://github.com/Mauville/MedCLIP/

## REFERENCES

1. Xiaofeng Z. Nadine S.Andrew W. Biomedical Information Technology Ch. 1 2008 , Pages 3-27 (Academic Press)
2. Turkington PM, Kennan N, Greenstone MAMisinterpretation of the chest x ray as a factor in the delayed diagnosis of lung cancer. Postgraduate Medical Journal 2002;78:158-160.
3. Vikas K. et al. A Common Medical Error: Lung Cancer Misdiagnosed as Sputum Negative Tuberculosis. Asian Pacific Journal of Cancer Prevention, Vol 10, 2009
4. Westermark P, Stenkvist B. A New Method for the Diagnosis of Systemic Amyloidosis. Arch Intern Med. 1973;132(4):522–523. doi:10.1001/archinte.1973.03650100040007

5. He, Q.P., Qin, S.J. and Wang, J. (2005), A new fault diagnosis method using fault directions in Fisher discriminant analysis. AIChE J., 51: 555-571. https://doi.org/10.1002/aic.10325
6. L. Wen, X. Li, L. Gao and Y. Zhang, "A New Convolutional Neural Network-Based Data-Driven Fault Diagnosis Method," in IEEE Transactions on Industrial Electronics, vol. 65, no. 7, pp. 5990-5998, July 2018, doi: 10.1109/TIE.2017.2774777.
7. L. Guo, Y. Lei, S. Xing, T. Yan and N. Li, "Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines With Unlabeled Data," in IEEE Transactions on Industrial Electronics, vol. 66, no. 9, pp. 7316-7325, Sept. 2019, doi: 10.1109/TIE.2018.2877090.
8. Zargari Khuzani, A., Heidari, M. & Shariati, S.A. COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. Sci Rep 11, 9887 (2021). https://doi.org/10.1038/s41598-021-88807-2
9. Feng, J., Lee, J., Vesoulis, Z.A. et al. Predicting mortality risk for preterm infants using deep learning models with time-series vital sign data. npj Digit. Med. 4, 108 (2021). https://doi.org/10.1038/s41746-021-00479-4
10. S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), 2014, pp. 1015-1018, doi: 10.1109/ISBI.2014.6868045.
11. M. He and D. He, "Deep Learning Based Approach for Bearing Fault Diagnosis," in IEEE Transactions on Industry Applications, vol. 53, no. 3, pp. 3057-3065, May-June 2017, doi: 10.1109/TIA.2017.2661250.
12. Wenqing Sun, Bin Zheng, Wei Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis, 97850Z (24 March 2016); https://doi.org/10.1117/12.2216307
13. Liu, Y., Jain, A., Eng, C. et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 26, 900–908 (2020). https://doi.org/10.1038/s41591-020-0842-3
14. Z. Soferman, D. Blythe and N. W. John, "Advanced graphics behind medical virtual reality: evolution of algorithms, hardware, and software interfaces," in Proceedings of the IEEE, vol. 86, no. 3, pp. 531-554, March 1998, doi: 10.1109/5.662878.
15. Stevens E. Antiga L. Viehmann T. Deep Learning with PyTorch. (Manning Publications 2020)
16. TensorFlow. Libraries and Extensions. 2021 URL:https://www.tensorflow.org/resources/libraries-extensions
17. Keras. Developer guides. 2021 URL: https://keras.io/guides/
18. Kapish K. WHY PYTHON ROCKS FOR RESEARCH....??? International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 09 | Sep 2019
19. RADFORD, Alec, et al. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.
20. Xinlei C. et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325v2 3 Apr 2015
21. Alsentzer E. ClinicalBERT. 2020. URL: https://github.com/EmilyAlsentzer/clinicalBERT
22. Hareem A. et al. Automatic medical image interpretation: State of the art and future directions. Elsevier Pattern Recognition 114 (2021) 107856
23. Lister Hill National Center for Biomedical Communications. MedPix. URL: https://medpix.nlm.nih.gov/home (2021)
24. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. CoRR, abs/2103.00020. Opgehaal van https://arxiv.org/abs/2103.00020
25. Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019, Junie). Publicly Available Clinical BERT Embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72–78. doi:10.18653/v1/W19-1909.

## Author Contributions

T.J. and V.O. made the main development of the code and structure modification, and optimization. G.J. interpreted the results and metrics to analyze the accuracy of the medical model.