Multi-modality 3D CNN Transformer for Assisting Clinical Decision in Intracerebral Hemorrhage

Zicheng Xiong[1]*, Kai Zhao[2]*, Like Ji[3], Xujun Shu[4,5], Dazhi Long[6], Shengbo Chen[1]✉, Fuxing Yang[7]✉

[1] School of Computer and Information Engineering, Henan University, Kaifeng, China
[2] Department of Neurosurgery, The First Medical Center, Chinese PLA General Hospital, Beijing, China
[3] School of Software, Henan University, Kaifeng, China
[4] Department of Neurosurgery, Jinling Hospital, Nanjing, China
[5] Department of Neurosurgery, Affiliated Jinling Hospital, Medical School of Nanjing University, Nanjing, China
[6] Department of Urology, Ji'an Third People's Hospital, Ji'an, China
[7] Department of Neurosurgery, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, China

Abstract. Intracerebral hemorrhage (ICH) is a cerebrovascular disease with high mortality and morbidity rates. Early-stage ICH patients often lack clear surgical indications, which is quite challenging for neurosurgeons to make treatment decisions. Currently, early treatment decisions for ICH primarily rely on the clinical experience of neurosurgeons. Although there have been attempts to combine local CT imaging with clinical data for decision-making, these approaches fail to provide deep semantic analysis and do not fully leverage the synergistic effects between different modalities. To address this issue, this paper introduces a novel multi-modality predictive model that combines CT images and clinical data to provide reliable treatment decisions for ICH patients. Specifically, this model employs a combination of 3D CNN and Transformer to analyze patients' brain CT scans, effectively capturing the 3D spatial information of intracranial hematomas and surrounding brain tissue. In addition, it utilizes a contrastive language-image pre-training (CLIP) module to extract demographic features and important clinical data and integrates with CT imaging data through a cross-attention mechanism. Furthermore, a novel CNN-based multilayer perceptron (MLP) layer is designed to enhance the understanding of the 3D spatial features. Extensive experiments conducted on real clinical datasets demonstrate that the proposed method significantly improves the accuracy of treatment decisions compared to existing state-of-the-art methods. Code is available at https://github.com/Henry-Xiong/3DCT-ICH.

Keywords: ICH · Multi-modality · CNN · Transformer · Clinical Decision.

* Equal Contribution, ✉ Corresponding Author.

1 Introduction

Intracerebral hemorrhage (ICH) is a prevalent cerebrovascular disease caused by the non-traumatic rupture of intracranial vessels and carries exceptionally high rates of mortality and disability [8,20]. Recently, the incidence of ICH has been on a sharply rising trend, particularly in China, where the proportion of strokes attributed to ICH ranges from 18.8% to 47.6%, substantially surpassing the international average. In clinical practice, choosing the appropriate treatment approach—whether surgical or conservative—is essential for the prognosis of ICH patients [1,10]. This is especially pertinent for patients who, based on initial cranial CT scans, are deemed to lack surgical indications but, in reality, require operative intervention; selecting an incorrect treatment method can result in grave outcomes, even threatening the patient's life [18,23]. Consequently, there is a pressing need for more objective decision-making tools in clinical practice to aid physicians in making precise treatment decisions. Unfortunately, the current strategies for ICH treatment primarily rely on clinicians' interpretations of medical guidelines combined with their own experience, which introduces a degree of subjectivity and limitations, such as challenges in making effective early-stage predictions [16].

However, to address the limitations of experience-based decision-making methods, researchers have explored evidence-based and individualized predictive models [3]. While most existing studies rely solely on data from a single modality, such as imaging data, in actual medical scenarios, neurosurgeons need to integrate a wide range of information, including the patient's radiographic features, clinical presentation, and laboratory test results [7,11,17]. In this process, the interpretation of imaging data is significantly influenced by clinical information. For example, neurosurgeons need to consider a patient's overall health and symptoms when evaluating a brain scan. While some studies integrate multi-modality features, they often directly input numerical values of laboratory results or scoring scales [15,22,25]. This approach makes it challenging for the model to fully explore the potential correlations between imaging and clinical data [24]. More critically, although some research has adopted multi-modality approaches, they have only collected data from specific brain regions without using the patient's whole-brain imaging data as input, hindering the model's understanding of intracerebral hemorrhage and surrounding brain tissue [2,15,22], as well as limiting the accuracy and comprehensiveness of predictions. Crucially, the hemorrhage's size, location, and impact on surrounding brain tissues are critical factors in determining the most appropriate treatment. Existing methods fall short by only analyzing specific brain regions, which limits their ability to provide precise and personalized decision support.

Therefore, this study designs a model that integrates initial CT scans, demographic data, and clinical tests as structured data to assist neurosurgeons in making reliable treatment decisions at an early stage. Recently, Transformer have set new benchmarks in computer vision tasks [4,12]. Specifically, the Vision Transformer (ViT) [5] processes images in small segments, or patches, to model long-range dependencies through an attention mechanism, providing a comprehensive view. However, segmenting 3D CT data directly into patches can miss important spatial details, which 3D CNN can capture with their convolutional operations [4,24]. Motivated by this finding, our model combines 3D CNN and Transformers to grasp both local and global features effectively. It starts with a 3D CNN extracting spatial features and downsampling the 3D images to compact feature maps, which captures local context efficiently and reduces computational load. These maps are then transformed into tokens for the ViT to model global features. Additionally, to enhance understanding of clinical data, we employ a pre-trained text encoder from the contrastive language-image pre-training (CLIP) [21] model to transform demographic and clinical data into descriptive vectors. This step ensures a more profound comprehension of medical knowledge and integrates it with CT imaging features using a cross-attention module [14]. Lastly, recognizing the traditional multilayer perceptron (MLP) layer's limitations in ViT for capturing deep semantic features, we innovate with a CNN-based MLP layer (CMLP), enhancing the model's ability to understand global spatial information and improve prediction accuracy.

The key contributions of this work are as follows.

1. Introducing a multi-modality prediction model that combines 3D CNN with Transformers, utilizing CT images and clinical data at admission to effectively predict treatment approaches for ICH.
2. Utilizing a pre-trained CLIP model, we enhance multi-modality information integration and deepen the model's understanding of specialized medical knowledge. Additionally, we introduce a CNN-based MLP layer to improve global spatial feature detection.

2 Methods

2.1 Architecture of Multi-modality Model

Our multi-modality model architecture is presented in Fig. 1. Concretely, it processes a 3D CT scan (dimensions $H \times W \times D$ for height, width and depth) using 2D and 3D CNNs to capture planar and spatial information. An improved Transformer effectively integrates data from both modalities to generate predictions. The following sections will describe the components of the model in detail.

2.2 Feature Extraction

Given the high computational costs associated with segmenting 3D medical imaging data into small 3D patches and flattening them for Transformer processing, we utilize both 2D and 3D convolutions to efficiently extract features from CT images. Considering the unevenness in depth compared to width and height in CT images ($512 \times 512 \times 128$), we first apply 2D convolutions of size 3×3 to extract features from each axial slice, converting the CT image dimensions to a uniform $128 \times 128 \times 128$ cube for subsequent processing. To further refine features, we apply 3D convolutions of size $3 \times 3 \times 3$ to construct residual

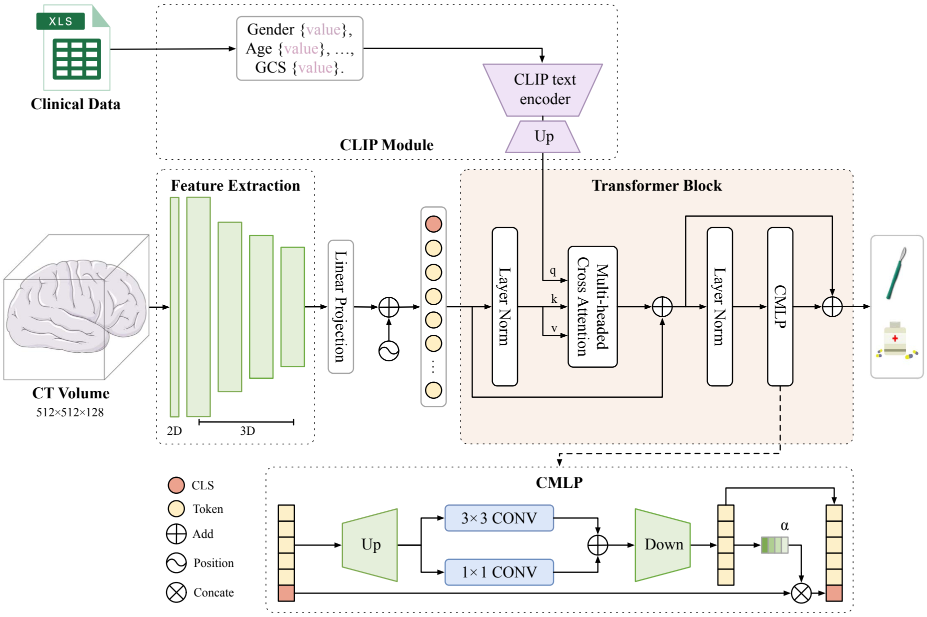

Fig. 1. Architecture of our multi-modality model.

connection blocks, reducing the high-dimensional 3D CT images through four downsampling steps to feature maps ($16 \times 16 \times 16$) with rich, high-level representations. Finally, we resize these feature maps and add positional encoding to form the final input feature maps, which are then fed into subsequent modules for a deeper exploration of the global receptive field.

2.3 CLIP Module

Leveraging CLIP's superior textual feature comprehension, we initially merge numerical and textual data into sentences depicting patient conditions. CLIP's text encoder, a Transformer-based module, extracts crucial information for deep semantic representation. To reconcile textual and image data dimension disparities, an upsampling layer adjusts text feature sizes for image-text modality alignment. Feature fusion employs a cross-attention mechanism by:

$$\text{Attention}(Q^{(t)}, K^{(i)}, V^{(i)}) = \text{softmax}\left(\frac{Q^{(t)}(K^{(i)})^T}{\sqrt{d_k}}\right)V^{(i)}, \tag{1}$$

where $Q^{(t)}$ represents the query vectors derived from the textual data, while $K^{(i)}$ and $V^{(i)}$ correspond to the key and value vectors generated from the image data, respectively. d_k denotes the dimensionality of the key vectors, serving as a scaling factor. This processing enhances CT and clinical data integration, facilitating a more comprehensive analysis. This not only deepens clinical data understanding but also mimics the clinical diagnostic approach.

2.4 CMLP

In ViT, the features extracted by the attention mechanism typically rely on the processing by MLP. Building on this, we redesign the MLP layer using CNN to enhance spatial information capture. Initially, we separate the CLS token from feature vectors. Then, we perform an upward projection of the remaining features, allowing them to pass through a 3 × 3 convolution layer followed by a 1 × 1 convolution layer. The features produced by these convolution layers are merged and resized to their original dimensions through a downward projection. Subsequently, these features are transformed into a single weight α, via average pooling, multiplied by the CLS token. Finally, this processed CLS token is concatenated with the CNN-processed features, aiming to improve the model's ability to capture adjacent features and enhance spatial information perception and utilization.

3 Experiments

3.1 Dataset and Experimental Details

Dataset. We use a dataset of ICH cases obtained from The Second Affiliated Hospital of Fujian Medical University, comprising CT images and clinical data from 442 patients, 196 treated conservatively and 246 surgically. The clinical data includes 20 variables (gender, age, hypertension, diabetes, smoking history, drinking history, time of onset, glasgow coma scale (GCS), national institute of health stroke scale (NIHSS), heart rate, potassium, sodium, white blood cells, platelets, prothrombin time (PT), international normalized ratio (INR), activated partial thromboplastin time (APTT), fibrinogen (FIB), thrombin time (TT), D-dimer) at admission, most of which have been shown to correlate with ICH treatment and prognosis [9]. The ethics committee approved all data collection and usage procedures in this study.

Experimental details. CT images are preprocessed by performing skull stripping using a trained nnUnet to eliminate extraneous interference. Images are normalized and zero-padded to standardize varying slice thicknesses to a depth 128. We set epochs to 50, batch size to 8, with cross-entropy as the loss function, and employ the AdamW [13] optimizer with an initial learning rate of 1e-4. Experiments are conducted using PyTorch on two NVIDIA Quadro RTX 6000 GPUs (each has 24GB memory). Model performance is evaluated through five-fold cross-validation, using Accuracy, AUC, Precision, Recall, and F1-Score metrics.

3.2 Results

To evaluate our model's performance, we compare it against the state-of-the-art predictive models. These models include ResNet [6], a commonly used model in medical diagnostic predictions serving as the baseline IRENR [25], which employs two parallel Transformer blocks for multimodal fusion and prediction

Table 1. Comparison with state-of-the-art methods on dataset.

Method	*AUC*	*Accuracy*	*Precision*	*Recall*	*F1-score*
ResNet [6]	.636	.679	.662	.844	.732
IRENR [25]	.643	.619	.602	.573	.587
DAFT [19]	.812	.750	.743	.802	.780
GCS-ICHNet [22]	.686	.621	.610	**.940**	.730
TOP-GPM [15]	.767	.736	.857	.680	.756
Ours (only CT)	.568	.650	.570	.896	.686
Ours	**.903**	**.846**	**.864**	.818	**.856**

DAFT [19], featuring a CNN module that dynamically rescales and shifts convolutional feature maps for multimodal prediction GCS-ICHNet [22], integrating brain CT slices and GCS scores via attention mechanism for patient prediction through voting and TOP-GPM [15], using a variational distribution combination module to merge CT images, clinical data, and treatment assignment for generating patient prognosis scores. Apart from ResNet, other models are applied based on our multimodal data for a fair comparison.

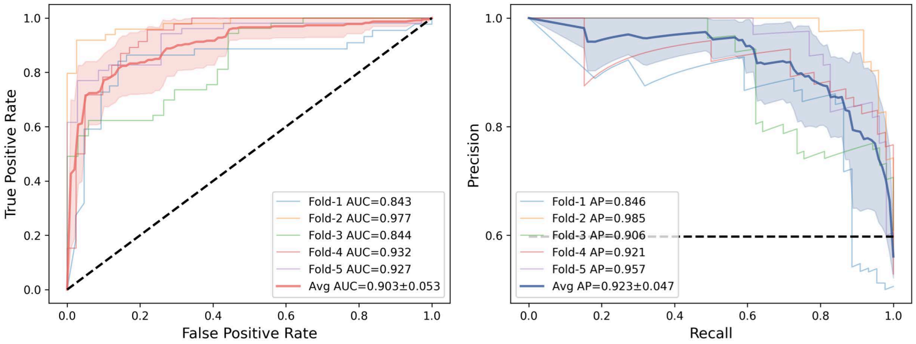

Fig. 2. Five-fold cross-validation ROC and PRC.

Tab. 1 summarizes the performance comparison of different approaches for predicting ICH treatment methods using these five metrics. Additionally, a uni-modal prediction using only CT images is performed with our model. Compared to the baseline, our multi-modality models generally outperform the uni-modal one due to the access to more comprehensive patient data. IRENR and GCS-ICHNet, which rely only on information from individual CT slices, fail to represent the complete patient profile. DAFT and TOP-GPM only utilize data from specific brain regions, lacking a global understanding of the imaging. Our model surpasses others by analyzing the entire brain, addressing these shortcomings, and achieving more accurate predictive performance, as detailed in the five-fold cross-validation results shown in Fig. 2. In the original study, GCS-ICHNet achieves an AUC and accuracy of 0.860 and 0.857 respectively, while TOP-GPM has average AUC and accuracy of 0.787 and 0.811 across two categories. However, compared to the original experimental results, our results show a certain degree of decline. This decrease is primarily attributed to differences in data distribution and the fact that the tasks our research addresses differ from those targeted by these two models designed for ICH.

3.3 Ablation Study

We conduct ablation experiments as shown in Tab. 2 to validate the effectiveness of combining the CLIP module with the CMLP in enhancing multi-modality data processing capabilities. We first evaluate the model's performance without the CLIP or CMLP module. Then, we assess its performance with each module integrated independently. With only the CLIP module deployed, the model demonstrates its capability to fuse deep semantic features between text and images. However, due to the lack of sufficient spatial information processing, it fails to fully leverage clinical data for comprehensively understanding complex 3D CT images, which limits the enhancement of performance. Similarly, while the CMLP aims to boost spatial information capture, the absence of CLIP module support prevents the effective integration of deep semantic links between modalities. Despite improvements in spatial feature handling by the CLIP module, the lack of optimized multi-modal data fusion still hinders performance enhancement.

Table 2. Ablation study

CT	Texts	CLIP	CMLP	*AUC*	*Acc*
✓	✓	✗	✗	.608 ± .064	.563 ± .033
✓	✓	✓	✗	.604 ± .028	.563 ± .036
✓	✓	✗	✓	.603 ± .030	.565 ± .035
✓	✓	✓	✓	**.903 ± .053**	**.846 ± .064**

It is evident that we designed the CLIP module and CMLP to exhibit clear complementary functionality at the operational level. When combined, their combined effect not only optimizes the fusion process between modalities but also significantly enhances spatial feature capture, thereby substantially improving model performance.

4 Conclusion

This paper introduces a novel multi-modality model that combines CT images and clinical data to provide reliable treatment decisions for patients with early-stage ICH. The model leverages 3D CNN with Transformers to analyze comprehensive brain information and employs a CLIP module with cross-attention for multi-modality data fusion. Additionally, a redesigned CNN-based MLP layer enhances the model's ability to capture data features. The model's effectiveness is validated through five-fold cross-validation.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62102133, the High-level and Urgently Needed Overseas Talent Programs of Jiangxi Province under Grant No. 20232BCJ25026, the Natural Science Foundation of Henan Province under Grant No. 242300421402, the Kaifeng Major Science and Technology Project under Grant No. 21ZD011, the Ji'an Finance and Science Foundation under Grants No. 20211-085454, 20222-151746, 20222-151704, Ji'an key core common technology "reveal the list" Project under Grant No. 2022-1, Science and Technology Fund Project of Guizhou Provincial Health Commission No. gzwkj2021-198, Quanzhou Science and Technology Plan Project No. 2024NY040, the Seed Fund Project of the Second Affiliated Hospital of Fujian Medical University No. 2021MP01 and the Henan University Graduate Student Excellence Program No. SYLYC2023135.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adeoye, O., Broderick, J.P.: Advances in the management of intracerebral hemorrhage. Nature Reviews Neurology **6**(11), 593–601 (2010) 2
2. Borsos, B., Allaart, C.G., van Halteren, A.: Predicting stroke outcome: A case for multimodal deep learning methods with tabular and ct perfusion data. Artificial Intelligence in Medicine **147**, 102719 (2024) 2
3. Boutet, A., Madhavan, R., Elias, G.J., Joel, S.E., Gramer, R., Ranjan, M., Paramanandam, V., Xu, D., Germann, J., Loh, A., et al.: Predicting optimal deep brain stimulation parameters for parkinson's disease using functional mri and machine learning. Nature communications **12**(1), 3043 (2021) 2
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 2, 3
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5, 6
7. He, S., Grant, P.E., Ou, Y.: Global-local transformer for brain age estimation. IEEE transactions on medical imaging **41**(1), 213–224 (2021) 2
8. Heit, J.J., Iv, M., Wintermark, M.: Imaging of intracranial hemorrhage. Journal of stroke **19**(1), 11 (2017) 2

9. Ji, R., Shen, H., Pan, Y., Wang, P., Liu, G., Wang, Y., Li, H., Zhao, X., Wang, Y.: A novel risk score to predict 1-year functional outcome after intracerebral hemorrhage and comparison with existing scores. Critical Care **17**, 1–10 (2013) 5
10. Keep, R.F., Hua, Y., Xi, G.: Intracerebral haemorrhage: mechanisms of injury and therapeutic targets. The Lancet Neurology **11**(8), 720–731 (2012) 2
11. Li, L., Poon, M.T., Samarasekera, N.E., Perry, L.A., Moullaali, T.J., Rodrigues, M.A., Loan, J.J., Stephen, J., Lerpiniere, C., Tuna, M.A., et al.: Risks of recurrent stroke and all serious vascular events after spontaneous intracerebral haemorrhage: pooled analyses of two population-based studies. The Lancet Neurology **20**(6), 437–447 (2021) 2
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 2
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 5
14. Lu, Z., Xie, H., Liu, C., Zhang, Y.: Bridging the gap between vision transformers and convolutional neural networks on small datasets. Advances in Neural Information Processing Systems **35**, 14663–14677 (2022) 3
15. Ma, W., Chen, C., Abrigo, J., Mak, C.H.K., Gong, Y., Chan, N.Y., Han, C., Liu, Z., Dou, Q.: Treatment outcome prediction for intracerebral hemorrhage via generative prognostic model with imaging and tabular data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 715–725. Springer (2023) 2, 6
16. Magid-Bernstein, J., Girard, R., Polster, S., Srinath, A., Romanos, S., Awad, I.A., Sansing, L.H.: Cerebral hemorrhage: pathophysiology, treatment, and future directions. Circulation research **130**(8), 1204–1229 (2022) 2
17. Nguyen, H.H., Saarakkala, S., Blaschko, M.B., Tiulpin, A.: Climat: Clinically-inspired multi-agent transformers for knee osteoarthritis trajectory forecasting. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022) 2
18. de Oliveira Manoel, A.L.: Surgery for spontaneous intracerebral hemorrhage. Critical Care **24**(1), 45 (2020) 2
19. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3d image and tabular data via the dynamic affine feature map transform. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 688–698. Springer (2021) 6
20. Puy, L., Parry-Jones, A.R., Sandset, E.C., Dowlatshahi, D., Ziai, W., Cordonnier, C.: Intracerebral haemorrhage. Nature Reviews Disease Primers **9**(1), 14 (2023) 2
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3
22. Shan, X., Li, X., Ge, R., Wu, S., Elazab, A., Zhu, J., Zhang, L., Jia, G., Xiao, Q., Wan, X., et al.: Gcs-ichnet: Assessment of intracerebral hemorrhage prognosis using self-attention with domain knowledge integration. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2217–2222. IEEE (2023) 2, 6

23. Wang, K., Liu, Q., Mo, S., Zheng, K., Li, X., Li, J., Chen, S., Tong, X., Cao, Y., Li, Z., et al.: A decision tree model to help treatment decision-making for severe spontaneous intracerebral hemorrhage. International Journal of Surgery pp. 10–1097 (2023) 2
24. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 109–119. Springer (2021) 2, 3
25. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering pp. 1–13 (2023) 2, 5, 6