Application of Vision-Series Transformer in screening for coronary heart diseases using coronary CT angiography

Kunlun Wang
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
wangkl20@mails.tsinghua.edu.cn

Hanyang Meng
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
mhy@mails.tsinghua.edu.cn

Xingjun Wang*
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
wangxingjun@mails.tsinghua.edu.cn

ABSTRACT

With the development of computer vision, more and more researches support the utilization of machine learning (ML) in the diagnosis of coronary heart disease (CHD) using CT angiography. Artificial intelligence shows strength in assisting researchers and clinicians in CT scan information reading. CTA data provides a series of scans of images from patients and thus could be reconstructed as three dimensions (3-D) images, also known as volume rendering images. We use the projection of those 3-D images as a dataset and developed a deep neural network with Transformer to diagnose coronary heart diseases automatically. Compared with several traditional and classic deep learning methods, i.e. simple CNN, VGGNet, and ResNet which show accuracy of 52.5%, 57.82%, and 76.3%, respectively, vision transformer(ViT) improves it to 81.5%. Besides, we take great advantage of the vision transformer and series of CTA. Based on that, we developed a new deep learning model Vision-Series Transformer(ViST). It has better performance in accuracy(83.78%). 5-fold cross-validation was used to determine whether it performs better in other statistical indicators. It shows the same high sensitivity and specificity as ViT, better than that of other models. The results show that both ViT and ViST could be used in screening for CHD, while ViST has better statistical indicators.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision tasks**; • **Biometrics**;

KEYWORDS

machine learning, transformer, coronary heart disease, coronary CT angiography

ACM Reference Format:
Kunlun Wang, Hanyang Meng, and Xingjun Wang. 2023. Application of Vision-Series Transformer in screening for coronary heart diseases using coronary CT angiography. In *2023 4th International Conference on Computing, Networks and Internet of Things (CNIOT '23), May 26–28, 2023, Xiamen, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3603781.3603858

*Corresponding author.

CNIOT '23, May 26–28, 2023, Xiamen, China

ACM ISBN 979-8-4007-0070-5/23/05.
https://doi.org/10.1145/3603781.3603858

1 INTRODUCTION

Nowadays, malignant tumors, cerebral apoplexy, and coronary heart diseases are the highest lethal disease in the world [1]. According to the statistics of the World Health Organization (WHO), 17,000,000 deaths are caused by cardiovascular diseases, accounting for 31% of all deaths worldwide. It was estimated that 7,400,00 deaths caused by CHD in the past 30 years. Largely due to CHD, an estimated annual total of 4,000,00 deaths happened in Europe and 1,900,000 deaths in the European Union [2]. According to the data in the China Health Statistics Yearbook in 2019, In China, the death rate of CHD in cities is 120.18/100,000, while that in the countryside is 128.24/100,000.

Coronary heart disease has a large number of patients and is difficult to cure. How to efficiently diagnose, manage and even predict CHD and find out the high-risk groups of CHD has become a public medical problem that needs to be solved urgently. The traditional diagnosis of CHD relies entirely on doctors, and pathological diagnosis is generally used as the gold standard.

However, due to the long and difficult training period for clinicians, there is a large shortage of doctors, especially in grassroots hospitals, which leads to the low diagnostic level of grassroots hospitals, thus misdiagnosis and missed diagnosis happens from time to time.

In clinical practice, there are mainly two methods for diagnosing whether a patient has CHD, namely coronary angiography and coronary computed tomography angiography. Coronary angiography is a relatively broad detection and diagnosis technology that is traditionally used. This technology has high diagnostic accuracy, but this diagnostic technology requires relatively high technical requirements for operators. It is an invasive operation, and patients can be diagnosed in the operating room. It requires high surgical expenses, bears certain surgical risks, and bears certain risks of other postoperative complications, so it is a relatively expensive inspection method; The examination and diagnosis mode of CTA image detection is a relatively new diagnosis mode. It appeared with the improvement of computer-aided diagnosis technology. With the improvement of arterial CT angiography technology, CTA image detection and diagnosis also have high accuracy. The advantage of this kind of detection is that the procedure and operation are simple, and it costs less time and money. So that it is easier to be accepted by patients. The disadvantage is that the accuracy rate is lower than that of coronary angiography.

Recently, deep learning algorithms have been gradually and frequently applied in medical image analysis to process a wide range of data and automatically classify diseases with high accuracy without human interference [3] [4]. It has better performance and

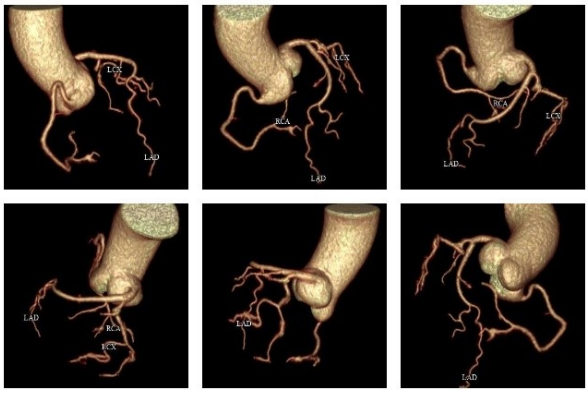

Figure 1: Coronary artery volume rendering images.

is still in a quick speed developing [5]. Li Y et al [6]mentioned that the created model reached the accuracy of 0.750±0.056, while the value of the area under the curve is 0.737.

2 METHODS

2.1 Study population

In this study, patients in Peking University Shenzhen Hospital were taken as the research object. By analyzing and processing their data, we train a machine learning model to predict whether they are CHD patients.

The study selected a total of 421 hypertensive patients who were treated at Peking University Shenzhen Hospital from 2019 to 2020 and obtained more than 400,000 effective computed tomography angiography images. Among the patients, 300 cases were diagnosed with CHD, and 140 cases were not suffering from CHD.

Among them, the computerized tomography blood vessel image data includes the stenosis of the left and right trunk, anterior descending artery, and circumflex artery. The data selected for this experiment have been desensitized and do not contain patient privacy information. The data application has been approved by the Research Ethics Committee of Peking University. The data application conforms to the ethical principles and the application is agreed upon.

To balance the data set, 140 negative patients and 142 positive patients were selected to make up the data set. Coronary artery tree volume rendering (VR) images were used to train the deep learning model. There are 1680 coronary artery tree VR images in total. Due to the different conditions of patients, the number of coronary heart tree VR images varies. In total, 84 patients have 5 coronary artery VR images each, 186 patients with 6 coronary VR images each, 6 patients with 10 coronary VR images, 5 patients with 12 coronary VR images, and 1 patient with 24 coronary VR images. The images are as shown in Fig. 1.

2.2 Model development

We use deep neural networks to predict whether a sample suffering from CHD. We selected AlexNet [7], VGGNet [8], ResNet [9–13] as controlled trials [14] [15]. Though some of them are invalid, ResNet

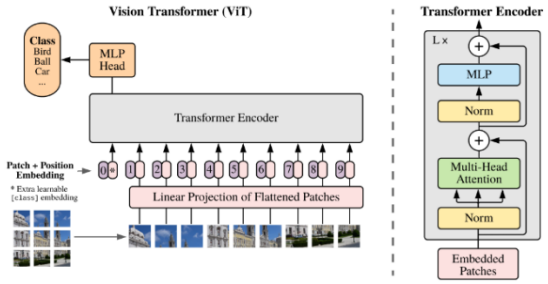

Figure 2: ViT model [17]

Figure 3: Coronary Artery Volume Rendering Image Segmentations. ViT models segmented images into pieces and serialize them then input them into the Transformer Encoder.

shows the ability to classify those samples. Compared with those methods, we found that Vision Transformer (ViT) [16]performs better in screening for CHD using CTA. Based on that, we developed a new deeplearning model named Vision-Series Transformer(ViST). The result shows that ViST performs better in accuracy and other statistics..

- We compare the results of SimpleCNN, VGGNet, and ResNet.
- We develop the ViT model (Fig. 2). It introduces Transformer [17] into a deep neural network. In 2020, the ViT model was developed and became the state-of-the-art in many datasets. Attention mechanism, which was first proposed in natural language processing by Vaswani [18] in 2017, was brought into computer vision and performed better than many other models. The first step of ViT is to segment images and serialize that segmentation.
- We developed Vision-Series Transformer Model. Images in ViT are segmented into small ones (Fig. 3). Those segmentations damaged the integrity of images. Though it could be ignored when the data set is large enough. Our data set is not large. Those slices of segmentations were treated as series and then sent into the deep learning model. Our data set, however, is initially the series. Every patient has a series of coronary artery volume rendering images (Fig. 4). This method avoids information loss during image segmentation.

3 EVALUATION

3.1 Statistics

We use the value of accuracy as the main evaluation indicator. In addition, we also report the experimental results using other commonly used evaluation indicators, including true positive(TP),

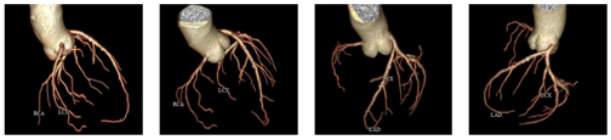

Figure 4: Coronary Coronary Artery Volume Rendering Image Series. ViST models input those series into the Transformer Encoder.

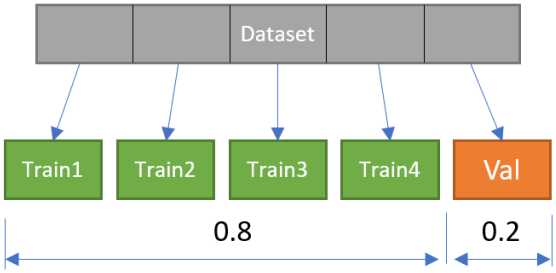

Figure 5: 5-fold cross-validation. We split dataset into 5 parts, 4 of them were used as train set and the remain one is used as validation set.

false positive(FP), true negative(TN), false negative(FN), specificity (SP), and Sensitivity(SN). And draw the corresponding AUC curve of the model.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity\ (SN) = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity\ (SP) = \frac{TN}{TN + FP} \tag{3}$$

3.2 5-fold cross-validation

Cross-validation is often used when the amount of data used for training is medium or small. Here, we use a 5-fold crossvalidation method to optimize the modeling and evaluate its performance on undiscovered data (also called generalization ability). Under this framework, optimize the model (i.e. adjust the super-parameters) using 5-fold cross-validation. Basic steps: first, randomly divide the dataset into five disjoint subsets of the same size. In this experiment, 282 patients were divided into five parts: 57, 57, 56, 56, and 56(Fig. 5). In each training, select four of them as the training set and the remaining one as the validation set, so that all data will participate in the training and prediction, effectively avoiding over-fitting.

Train procedures and 5-fold cross-validation are shown in Fig. 6.

4 RESULT

Zreik et al. [19] proposed an ML-based, convolutional neural network approach for detecting and classifying coronary plaques and stenosis. The authors used coronary CT angiography scans from 81 patients for network training, and 17 for validation, and found an accuracy of 0.77 for plaque analysis and of 0.80 for stenosis analysis. The accuracy of AI-based analysis was not significantly different from the accuracy for human observers, which were 0.80 and 0.83, respectively.

Table 1: Accuracy of AlexNet, VGGNet, ResNet and ViT

	AlexNet	VGGNet	ResNet	ViT
TN	109	119	161	174
TP	120	139	168	179
FN	110	91	62	51
FP	99	89	47	34
Acc.	52.28%	58.89%	**75.11%**	**80.59%**

As every patient has several coronary artery VR images, we split the data set into train and validation data as the ratio of 3:1. We have 208 patients with 1242 images in the train set and 74 patients with 438 images in the validation data. First, we use a single image as model input to train AlexNet, VGGNet, ResNet ViT, and ViST. Accuracy is shown in (TABLE I)

After that, we choose 3 models (ResNet, ViT, ViST) which perform better to evaluate to verify the effectiveness of our ViST. And during the evaluation, we added multiple indicators. Since the ViST model is to directly diagnose a certain patient, during training, we combine multiple coronary artery VR images of the patient into a sequence and input them into the neural network according to the Transformer mode. Therefore, the prediction we get is whether a certain patient is sick or not. In order to be able to compare with this result, we first use a way to evaluate the images predicted by the ResNet and ViT models. If the number of VR images which are predicted positive is greater than that predicted negative, we assume that the patient is positive. If that predicted positive is smaller than that predicted negative, we assume that the patient is negative. Otherwise, we assume that prediction wrong. For the ViST model, we can directly get the predicted results, so we compare and analyze the statistical data of these three models. Detailed statistics is shown in (TABLE II)

We also compute the TP, FP, TN, and FN of each fold and the mean of them of each model. From there we plot the ROC curve (Fig. 7) and PR curve of 5 folds belonging to one model in one figure (Fig. 8). To compare with other models in PR curve, we plot that curve of mean in one figure (Fig. 9).

4.1 Discussion

From the above experimental results, we can see that using the ViST can improve the efficiency of screening CHD patients using CTA images. ViST can improve the generalization ability. In addition, the artificial intelligence model can be very helpful for cardiologists to use CTA images conveniently. They do not need to have rich experience in pathology, because our artificial intelligence model is very convenient to use. Just input the coronary artery image into the model, and the model will automatically output a score, which guides the probability of CHD. Even those who are not familiar with cardiology can operate it easily.

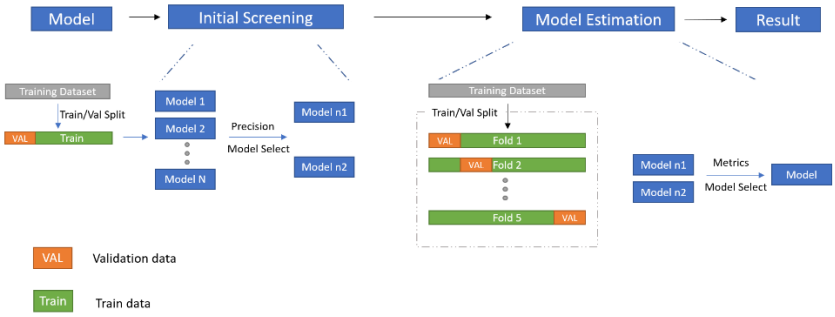

Figure 6: Train procedures and 5-fold cross-validation

Table 2: 5-fold cross-validation Acc. of ResNet, ViT, ViST

Fold	Model	TP	TN	FP	FN	Acc.
F1	ResNet	23	23	2	9	80.70%
	ViT	22	23	2	10	78.94%
	ViST	24	23	2	8	82.46%
F2	ResNet	22	23	5	8	72.19%
	ViT	22	25	3	7	82.46%
	ViST	25	24	4	4	85.96%
F3	ResNet	20	23	6	8	75.44%
	ViT	20	22	7	8	73.68%
	ViST	22	27	2	6	85.96%
F4	ResNet	22	20	9	5	75.00%
	ViT	25	25	7	2	85.45%
	ViST	22	24	5	5	80.35%
F5	ResNet	25	20	9	1	81.82%
	ViT	26	21	8	0	85.45%
	ViST	25	26	3	1	92.73%
Mean	ResNet	111	109	31	31	78.01%
	ViT	115	113	27	27	80.85%
	ViST	118	124	16	24	**85.81%**

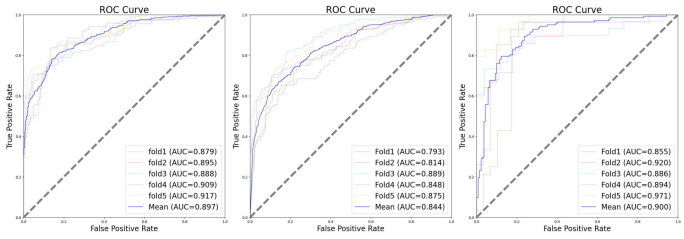

Figure 7: ROC curve of ResNet(left), ViT(middle), and ViST(right)

CT angiography has been proven to be a very powerful and valuable tool for determining whether a patient has CHD, but it requires a high level of clinical experience and knowledge. With the development of artificial intelligence, more and more computer vision technologies are widely used in the medical field. Since

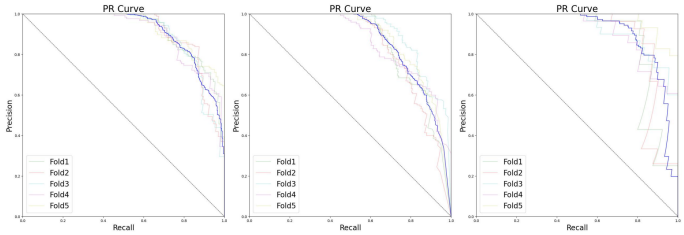

Figure 8: ROC curve of ResNet(left), ViT(middle), and ViST(right)

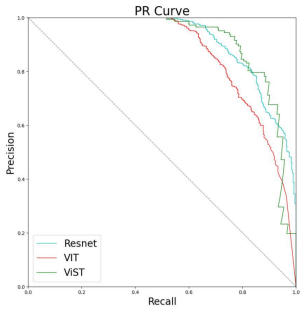

Figure 9: Mean PR curve of ResNet, ViT, ViST

2020, the Transformer has been transferred from the field of natural language processing to the field of computer vision and has achieved excellent results in many image data sets. So we bring the Transformer into CT angiography image diagnosis. We have done experiments on AlexNet, VGGNet, ResNet, and ViT, and compared the results. The results on simple CNN are not good, so we will not continue to analyze them. More statistics were analyzed on ResNet and ViT. On the basis of ViT, the ViST model was established to bring the patient's coronary volume rendering image sequence into consideration. Compared with ResNet, the accuracy of prediction is improved, and its sensitivity and specificity are improved compared with the ViT model.

This is because the image needs to be cut in the model of ViT, and then the cut image is arranged into a one-dimensional sequence, and the image group number of the one-dimensional sequence is sent into the model of Transformer for training. In our ViST model, we

randomly select four VR images of the patient's coronary artery into a sequence and then send the image sequence into the Transformer model with an attention mechanism for training directly. Better results were obtained in clinical data.

Both the ordinary ViT model and our ViST model have achieved better results. Our research results show that the model with Transformer has good accuracy and sensitivity, and can remind doctors of the need to pay attention to the patient's disease level to ensure that CHD will not be missed.

4.2 Limitation

Our research has some limitations that need to be pointed out. First, our experimental group is a single-center investigation. We only studied clinical patients from Shenzhen Hospital of Peking University. Future research needs a multi-center investigation to verify this conclusion. In addition, the tabular data assessment used in this study does not include the analysis of the original image. Third, this study is limited by the format of the pre-training model and the resolution of the patient's CT images. Only four pictures were randomly selected for each patient to be arranged into a sequence and sent into the model training. In fact, higher granularity can be used, such as 16 pictures. If only 5, 6, 11, and 12 pictures are not enough to make up a sequence, the subsequent pictures will be filled up with blank pictures to form a same-length vector.

5 CONCLUSION

The use of the Transformer can improve the accuracy of machine learning in diagnosing CHD using CT angiography coronary volume rendering images, and the use of ViST models can further improve the accuracy of prediction and can maintain good performance in both specificity and sensitivity compared with ViT. Introducing the algorithm of computer image into CHD diagnosis using CT angiography can make some preliminary diagnoses, which may be particularly beneficial to doctors with limited experience.

ACKNOWLEDGMENTS

This work was supported in part by the Research and Development Program of Shenzhen under Grant KCXFZ-202002011010487 and Grant WDZC20200818121348001. We would like to thank Peking University Shenzhen Hospital for their cooperation and help in CTA data extraction.

Ethical Approval

The study was approved by Ethics Committee of Peking University Shenzhen Hospital.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Stanaway J D, Shepard D S, Undurraga E A, *et al.* The global burden of dengue: an analysis from the Global Burden of Disease Study 2013[J]. The Lancet infectious diseases, 2016, 16(6): 712-723.
[2] Ferreira-González I. The epidemiology of coronary heart disease[J]. Revista Española de Cardiología (English Edition), 2014, 67(2): 139-144.
[3] He T, *et al.* Diagnostic models of the pre-test probability of stable coronary artery disease: a systematic review. Clinics (Sao Paulo). 2017;72(3):188–96
[4] G. Litjens *et al.*, "A survey on deep learning in medical image analysis", Med. Image Anal., vol. 42, pp. 60-88, Dec. 2017.
[5] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
[6] Li Y, Wu Y, He J, *et al.* Automatic coronary artery segmentation and diagnosis of stenosis by deep learning based on computed tomographic coronary angiography[J]. European Radiology, 2022, 32(9): 6037-6045.
[7] Huang G, Liu Z, Van Der Maaten L, *et al.* Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[9] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[10] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3367-3375, Jun. 2015.
[11] Ö. Çiçeke, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation", Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent, pp. 424-432, 2016.
[12] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, 2014, [online] Available: https://arxiv.org/abs/1412.6980.
[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929-1958, 2014.
[14] J. M. Wolterink, R. W. van Hamersvelt, M. A. Viergever, T. Leiner and I. Išgum, "Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier", Med. Image Anal., vol. 51, pp. 46-60, Jan. 2019.
[15] Y. Xu, G. Liang, G. Hu, Y. Yang, J. Geng and P. K. Saha, "Quantification of coronary arterial stenoses in CTA using fuzzy distance transform", Computerized Med. Imag. Graph., vol. 36, no. 1, pp. 11-24, 2012.
[16] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2625-2634, Jun. 2015.
[17] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[18] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[19] Zreik M, Van Hamersvelt R W, Wolterink J M, *et al.* A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography[J]. IEEE transactions on medical imaging, 2018, 38(7): 1588-1598.