Medical Image Analysis 90 (2023) 102953

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A clinically applicable AI system for diagnosis of congenital heart diseases based on computed tomography images

Xiaowei Xu [a,b,1], Qianjun Jia [a,b,c,1], Haiyun Yuan [a,b,d,1], Hailong Qiu [a,b,d,1], Yuhao Dong [a,b,c], Wen Xie [a,b,d], Zeyang Yao [a,b,d], Jiawei Zhang [a,b], Zhiqaing Nie [b], Xiaomeng Li [e], Yiyu Shi [f], James Y. Zou [g,h], Meiping Huang [a,b,c], Jian Zhuang [a,b,d,*]

[a] Guangdong Provincial Key Laboratory of South China Structural Heart Disease, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China
[b] Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China
[c] Department of Catheterization Lab, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
[d] Department of Cardiovascular Surgery, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China
[e] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region
[f] Computer Science and Engineering, University of Notre Dame, IN, 46656, USA
[g] Department of Computer Science, Stanford University, Stanford, CA, 94305, USA
[h] Department of Electrical Engineering, Stanford University, Stanford, CA, 94305, USA

ARTICLE INFO

Keywords:
Congenital heart disease
Diagnosis
Computed tomography
Artificial intelligence
Deep learning

ABSTRACT

Congenital heart disease (CHD) is the most common type of birth defect. Without timely detection and treatment, approximately one-third of children with CHD would die in the infant period. However, due to the complicated heart structures, early diagnosis of CHD and its types is quite challenging, even for experienced radiologists. Here, we present an artificial intelligence (AI) system that achieves a comparable performance of human experts in the critical task of classifying 17 categories of CHD types. We collected the first-large CT dataset from three different CT machines, including more than 3750 CHD patients over 14 years. Experimental results demonstrate that it can achieve diagnosis accuracy (86.03%) comparable with junior cardiovascular radiologists (86.27%) in a World Health Organization appointed research and cooperation center in China on most types of CHD, and obtains a higher sensitivity (82.91%) than junior cardiovascular radiologists (76.18%). The accuracy of the combination of our AI system (97.20%) and senior radiologists achieves comparable results to that of junior radiologists and senior radiologists (97.16%) which is the current clinical routine. Our AI system can further provide 3D visualization of hearts to senior radiologists for interpretation and flexible review, surgeons for precise intuition of heart structures, and clinicians for more precise outcome prediction. We demonstrate the potential of our model to be integrated into current clinic practice to improve the diagnosis of CHD globally, especially in regions where experienced radiologists can be scarce.

1. Introduction

Congenital heart disease (CHD) is a type of disease caused by abnormal heart structure, which is the most common type of birth defect (Van Der Linde et al., 2011). Accurate diagnosis is particularly important in CHD, which can be used for prevalence, interventions, surgery, and outcome prediction of CHD patients. Currently, computed tomographic (CT) has been widely used in the assessment of CHD (Stout et al., 2019; Choi et al., 2021). However, interpretation of these images remains challenging. The heart is a remarkably complex organ considering both its anatomical structures and function (periodic beat for blood circulation) (Mori et al., 2019), and CHD adds another layer of complexity with significant variations in heart structures and great vessel connections. Clinically, there are more than 20 types of CHD (or more than one hundred if subtypes are included) (Mazur et al., 2013;

* Corresponding author at: Guangdong Provincial Key Laboratory of South China Structural Heart Disease, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China.
E-mail addresses: jamesz@stanford.edu (J.Y. Zou), huangmeiping@126.com (M. Huang), zhuangjian@gdph.org.cn (J. Zhuang).
[1] Contributed equally.

https://doi.org/10.1016/j.media.2023.102953
Received 19 December 2022; Received in revised form 22 August 2023; Accepted 1 September 2023
Available online 6 September 2023
1361-8415/

Adebo, 2021), which makes the CHD diagnosis intractable (Han et al., 2015). Furthermore, due to the shortage of experienced cardiovascular radiologists, it is hard to provide timely and accurate diagnosis in clinical practice. The problem of accurate CHD diagnosis is aggravated in developing regions, because of the lack of experienced radiologists and because complex forms of CHDs are commonly seen due to the improper life habitats and environmental impact (Nicoll, 2018).

Echocardiography is the primary detection and monitoring method for CHD due to its accessibility, and magnetic resonance imaging (MRI) is the second most widely used method for diagnosis of CHD as it is free from radiation side effects (Han et al., 2013). Recently, computed tomography (CT) has also been widely used in clinic practice, especially in developing regions (Bonnichsen and Ammash, 2016). First, CT can provide high-resolution images of the complex structures of hearts with CHD, based on which radiologists can make a more detailed diagnosis. Note that MRI and echocardiography can provide motion analysis of the heart, while CT can only provide static heart structure information. However, CT is preferred for CHD, as CHD is characterized by complex structural variations. Second, surgeons can know the heart anatomy well based on the high-resolution CT images, which can help them make proper surgical planning. Last but most importantly, CT is quite cost-efficient, as CT machines are much cheaper than MRI machines, and its examination time is short, which makes CT much more affordable than MRI, especially in developing regions.

Recently, artificial intelligence (AI) has shown great potential in the diagnosis of diseases (Erickson, 2021). There are a large number of works covering a variety of medical data including histopathology images (Song et al., 2020; Coudray et al., 2018; Courtiol et al., 2019; Hollon et al., 2020), optical coherence tomography (De Fauw et al., 2018; Yim et al., 2020; Brown et al., 2018), electrocardiogram (Hannun et al., 2019; Attia et al., 2019; Ribeiro et al., 2020), CT images (Shi et al., 2020; Mei et al., 2020), X-ray (Lotter et al., 2021), electronic health record (Liang et al., 2019), and skin images (Liu et al., 2020; Soenksen et al., 2021). Recently, as a key step for the diagnosis of CHD, AI for segmentation of CT images in CHD has been studied (Pace et al., 2018; Xu et al., 2019). These works focus on limited segmentation categories which can only support diagnosis of limited types of CHD. A previous study has shown that methods based on deep learning can identify kinds of complicated cardiac malformations effectively in mouse models (Chu et al., 2020). Our pilot study has demonstrated that diagnosis of CHD with CT images based on deep learning is feasible on a small dataset but have not yet shown clinical applicability with acceptable performance and clinical usability (Xu et al., 2020).

In this paper, we propose an AI system to diagnose CHD based on CT images. Due to the high structural variations of CHD, a single end-to-end black-box network can hardly work as it typically requires millions of labeled scans (De Fauw et al., 2018). On one hand, the number of CHD CT scans is limited. On the other hand, 3D heart labeling of CT scans is quite time-consuming (1–2 h per scan by experienced radiologists). Thus, we combine deep learning and machine learning methods to perform segmentation, extraction of diagnosis-related features, and diagnosis, respectively, which mimics the workflow of experienced radiologists. To demonstrate the clinical applicability of this system, we compare the diagnosis of the AI system to those made by cardiovascular radiologists in routine clinical practice. Furthermore, we show how our AI system might be integrated into routine clinical workflows. Compared with existing AI-based diagnosis tools, our AI system has three distinct features. First, neural networks are good at capturing texture information in images, while CHD comes with significant structural variations that they cannot handle well. As such, we fuse graph based optimization used in conventional computer vision with neural networks for better segmentation and feature extraction. Second, due to a large number of CHD types and limited data, a single end-to-end black-box network is not possible. Thus, we incorporate domain knowledge with the extracted features to tackle the problem. Third, interpretation is important to make it acceptable by clinicians, and our AI system can produce 3D visualization of the heart for diagnosis interpretation.

Our contributions are summarized as follows:

- We propose an AI system for diagnosis of CHD which is usually with complex structural variations. As far as we know, this is the first work for automatic diagnose of CHD using a large-scale dataset;
- In order to tackle the problem of limit dataset size and large structural variations, we propose to combine deep learning and machine learning methods to perform segmentation and extraction of diagnosis-related features, respectively, which mimics the workflow of experienced radiologists;
- We collected a large dataset of 3750 CT images, 1282 of which is labeled as the training dataset. We conducted comprehensive experiments to evaluate the AI system in clinical practice, and experimental results show our method is comparable with junior cardiovascular radiologists, and has the potential to be used in clinical practice.

The remainder of the paper is organized as follows. Section 2 describes the collected dataset. Section 3 introduces the architecture of our AI system. Section 4 presents the experiments and results, followed by the discussion and conclusion in Section 5 and Section 6, respectively.

2. Related work

In recent years, we have witnessed a rapid development of cardiac image analysis with a variety of cardiac image datasets covering different modalities (CT, MRI, and Ultrasound) that have proliferated. For instance, the MICCAI'16 HVSMR (whole-Heart and great Vessel Segmentation from 3D cardiovascular MRI in congenital heart disease) challenge labeled four pediatric cardiac MRI images for segmentation of the blood pool and myocardium for pre-procedural planning of children (Pace et al., 2015). MICCAI'17 MM-WHS (multi-modality whole heart segmentation) challenge provides 120 multi-modality cardiac images acquired in the real clinical environment, which created an open and fair competition for various research groups to test and validate their methods. Meanwhile, MICCAI'17 ACDC (Bernard et al., 2018) (automated cardiac diagnosis challenge) supply a training dataset of 100 patients along with the corresponding manual references based on the analysis of one clinical expert and a testing dataset composed of 50 new patients, without manual annotations but with the patient information given above. In MICCAI'19, a CHD segmentation dataset (Xu et al., 2019) has been proposed. However, the segmentation categories is limited, which can only support diagnosis of limited types of CHD. Recently, we proposed imageCHD (Xu et al., 2020), a dataset for CHD diagnosis with labels on sufficient categories. However, this pilot study has demonstrated that diagnosis of CHD with CT images based on deep learning is feasible on a small dataset but have not yet shown clinical applicability with acceptable performance and clinical usability.

Recently, a crowd of deep learning based models have demonstrated great promise in medical image analysis. In the light of the MICCAI'17 MM-WHS (Zhuang and Shen, 2016) challenge and MICCAI'17 ACDC (Bernard et al., 2018) challenge, a variety of state-of-the-art heart segmentation models appeared and summarized in Zhuang et al. (2019). For instance, Yang et al. (2017) closely coupled the FCN with 3D operators, transfer learning and deep supervision mechanism to distill 3D contextual information and attack potential difficulties in training deep neural networks with hybrid loss. Meanwhile, some two-stage methods also attract more and more attention in cardiac image analysis for the extraction of the heart structures can focus the network on the located regions to improve the performance. Wang et al. (2018) introduced a two-stage modified U-Net architecture, which simultaneously detected an ROI from the full volume and segmented the voxels at

Table 1
Data characteristics[a].

Characteristics	Development dataset (Training set + validation dataset)	Test dataset	Total
Years	2007 to 2019	2007 to 2020	2007 to 2020
No. of patients	1,282	2,468	3,750
Age (in years): Median (25th, 75th percentiles)	4.1 (2.74, 0.25)	4.59 (1.90, 0.26)	4.45 (2.21, 0.29)
Female (%)	39.5%	41.2%	40.6%
CT machine type (3 types)			
Siemens' SOMATOM Definition Flash (dual-source CT)	458	1,214	1,672
Philips' Brilliance iCT (256-slice CT)	363	1,078	1,441
GE LightSpeed VCT (64-slice CT)	461	176	637
CHD types (17 types)			
Ventricular septal defect (VSD)	499	1,715	2,214
Single ventricle (SV)	9	11	20
Atrial septal defect (ASD)	633	1,335	1,968
Single atrium (SA)	5	6	11
Tetrology of fallot (ToF)	184	596	780
Double outlet right ventricle (DORV)	117	89	206
Transposition of great arteries (TGA)	59	124	183
Truncus arteriosus communis (TAC)	23	9	32
Pulmonary atresia (PuA)	55	131	186
Interrupted aortic arch (IAA)	45	15	60
Aortic arch hypoplasia (AAH)	91	86	177
Coarctation of the aorta (CoA)	180	205	385
Right aortic arch (RAA)	166	190	356
Anomalous pulmonary venous connection (APVC)	151	222	373
Pulmonary stenosis (PS)	618	787	1,405
Double superior vena cava (DSVC)	148	186	334
Patent ductus arteriosus (PDA)	448	636	1,084

[a] Note that some images may correspond to more than one type of CHD.

the original resolution. Payer et al. (2017a) employed two CNNs in an end-to-end manner including location CNN and segmentation CNN for the whole heart segmentation. Location segmentation firstly localized the center of all heart substructures, which can help the subsequent segmentation CNN to focus on the heart region. which led the network to focus on anatomically feasible configurations. Xu et al. (2019) used deep learning for segmentation first, and then extracted the connection information and apply graph matching to determine the categories of all anomalous vessels. Unlike previous methods, our pilot work (Xu et al., 2020) focused on CHD diagnosis which used deep learning and shape similarity to determine the diagnosis. However, the performance is still far from clinical usage.

3. Dataset

3.1. Ethical and information governance approvals

This work and the collection of data of retrospective data on implied consent received Research Ethics Committee (REC) approval from Guangdong Provincial People's Hospital, Guangdong Academy of Medical Science under Protocol No. 2019324H. It complies with all relevant ethical regulations. Deindentification was performed in which all CT files are transformed into NIfTI format, and sensitive information of the patients including name, birther day, admission year, admission number, and CT number is removed. Only de-identified retrospective data were used for research, without the active involvement of patients.

3.2. General information

Data was selected from a retrospective cohort of all patients who attended Guangdong Cardiovascular Institute at Guangdong Provincial People's Hospital, the only World Health Organization (WHO) appointed research and cooperation center in China, between January 1st, 2007 to June 1st, 2020, who received CT imaging as part of their routine clinical care. **Cases without heart surgery were excluded**. CT images containing severe artifacts or serious mismatch of slices were also excluded, which is usually caused by improper image acquisition of the CT operators. Three low-end to high-end types of CT machines are used including Siemens' SOMATOM Definition Flash, Philips' Brilliance iCT, and GE LightSpeed VCT. 17 types of CHD (Table 1) are considered, and for each considered type of CHD, data is divided according to their CT examination date. Particularly, due to the limited data and to ensure there are sufficient data of rare types of CHD, we assign at least 5 cases to each type. However, as one patient usually suffers more than one type of CHD, the number of patients assigned to each type of CHD varies significantly (Table 1).

3.3. Clinical taxonomy

Considering the common types of CHD (Bhat et al., 2016) and the distribution of CHD types in our center, 17 types of CHD are considered in our AI system for CHD diagnosis. Particularly, the 17 types are selected due to the following considerations. First, most of the common CHD like ventricular septal defect (VSD), atrial septal defect (ASD), Tetrology of fallot (ToF), double outlet right ventricle (DORV), pulmonary atresia (PuA), transposition of great arteries (TGA), coarctation of the aorta (CoA), pulmonary stenosis (PS), aortic arch hypoplasia (AAH), and patent ductus arteriosus (PDA) are included. Second, less common ones are aggregated, e.g., four subtypes of pulmonary venous drainage: supracardiac, cardiac, infradiaphragmatic, and mixed connections are aggregated as anomalous pulmonary venous connection (APVC). Such aggregation also happens in VSD, ASD, and PS. Other rare types of CHD, like single atrium (SA), single ventricle (SV), and interrupted aortic arch (IAA), are also included in the dataset as they are usually associated with the common ones. Others, like dextrocardia, are not included as their quantity is quite limited in our center. The dataset represents the full variety of CT examinations and CT-based CHD diagnosis at Guangdong Cardiovascular Institute at Guangdong Provincial People's Hospital. The definition of the above types of CHD can be referred to related standards (Mazur et al., 2013; Adebo, 2021). Note that there exist overlaps in the definition, e.g., ToF, DORV, PuA, TAC, and TGA are several stages of the disease development, and

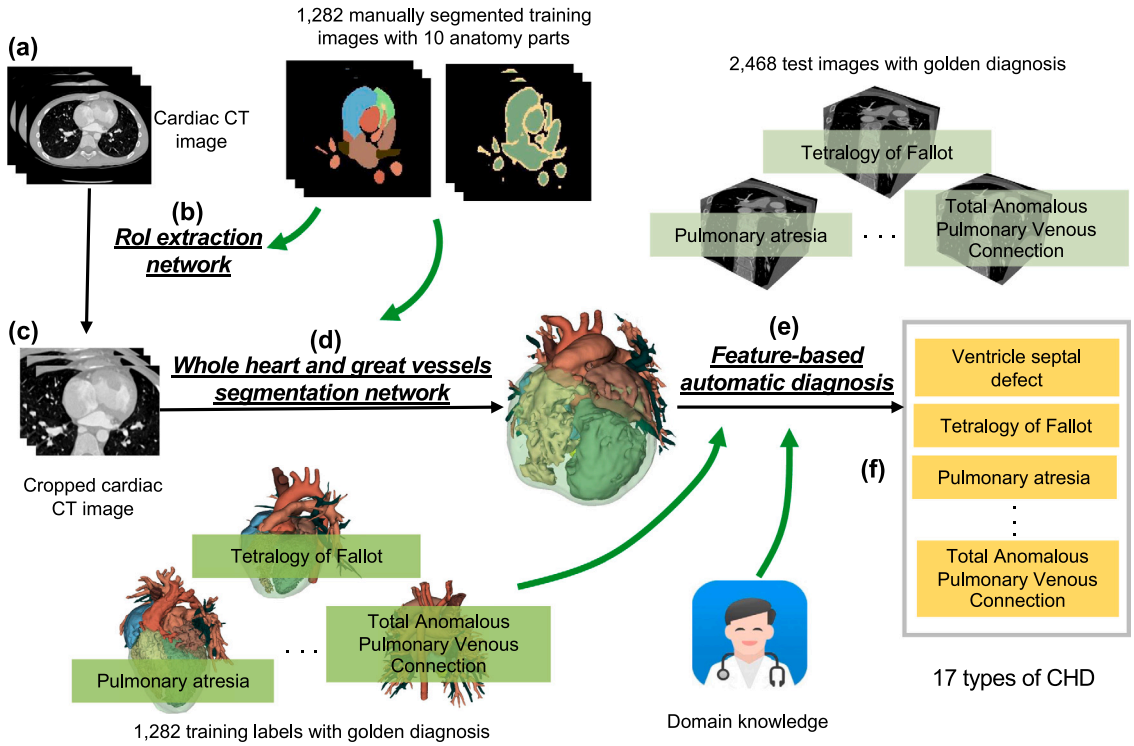

Fig. 1. Structure of our proposed AI system. (a) Cardiac CT images (512 × 512 × ~280); (b) RoI extraction network trained with manually segmented training images; (c) Resulting cropped cardiac CT image; (d) Whole heart and great vessels segmentation network trained with manually segmented training images; (e) Feature-based automatic diagnosis designed using domain knowledge and training labels with golden diagnosis by an overall consideration of CT reports, operative reports, discharge summaries, and expert review; (f) Predicted diagnosis; and our AI system is tested with 2468 images with golden diagnosis by an overall consideration of CT reports, operative reports, discharge summaries, and expert review.

sometimes one case can be assigned to both DORV and ToF. This situation also happens between CoA, AAH, and IAA (CoA indicates there is a limited stenosis part, and the stenosis part elongates in AAH, and finally the stenosis part becomes a thin line or disappears which indicates IAA happens. The clinic taxonomy of CHD in the current definition introduces much subjectivity and some confusion into the current diagnosis and treatment. Thus, for the above types of CHD, each case has two predicted classes either of which matching the golden diagnosis means correct prediction. However, we can still notice the complexity and challenges of CHD diagnosis even for radiologists, and not mention for our AI system. Also, the subjectivity and confusion show that AI-enabled precision medicine may help to mitigate such a situation.

3.4. Clinical labeling

Clinical labels of the adopted 17 types of CHD in our training dataset and test dataset were checked and assigned manually by senior radiologists. We have defined a golden standard which is an overall consideration of CT reports, operative reports, discharge summaries, and expert review. 17 resident physicians first collected CT reports, operative reports, and discharge summaries, and then extracted their diagnosis information in a universal diagnosis name (the description information of each type of CHD and other information are removed). To ensure there is no correlation, each resident physician just collects and extracts one kind of information, e.g., CT report, thus to avoid possible mistakes as he/she may be affected if already knowing the discharge summary. Then, four senior radiologists would review the corresponding CT image to check the diagnosis in person. If the radiologist is still uncertain of the diagnosis, at least three radiologists will then involve voting for the final diagnosis. Note that in the majority vote, radiologists consider the operative report as important proof.

3.5. Segmentation labeling

1282 CT images in the training dataset were manually segmented using an open-source segmentation software 3D Slicer (Pieper et al., 2004). The segmentation labels were chosen to distinguish the diagnosis of all the selected 17 types of CHD, and valves including atrioventricular valves, the aortic valve and the pulmonary valve are not labeled due to the fact that valves cannot be accurately discovered in CT images and are not relevant to almost all types of CHD. CHD suffers from serious structural variations, and its diagnosis is usually based on features like connections and shapes. Thus, minor segmentation errors between ventricles and atrium is acceptable. On the other hand, atrioventricular valves cannot be accurately discovered in CT images. 10 anatomical structures, including left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), aorta (AO), pulmonary artery (PA), pulmonary veins (PV), superior vena cava (SVC), and inferior vena cava (IVC) (Fig. 1). Particularly, we labeled AO, SVC, and IVC as follows. The beginning of the AO trunk is the aortic valve, and the ending of the descending AO is at the same axial plane as the lowest part of the myocardium, and the upper boundary of AO, including the brachiocephalic trunk, the left carotid artery, and the subclavian artery is at the same axial plane as the highest part of the lung. The lower boundary of IVC is at the same axial plane as the lowest part of the myocardium, while the upper boundary of SVC is at the same axial plane as the highest part of the lung. A three-stage workflow is adopted for segmentation labeling. First, 17 resident physicians whose major are medical imaging are recruited and trained to perform the initial segmentation using the segmentation module in 3D Slicer. Then, three junior radiologists manually review all the segmentation labeling, and each CT has been reviewed twice. Finally, if the three junior radiologists have concerns on some cases, a senior radiologist will further check the segmentation labels in detail. Note that in order to ensure quality, the associated diagnosis is also considered during the

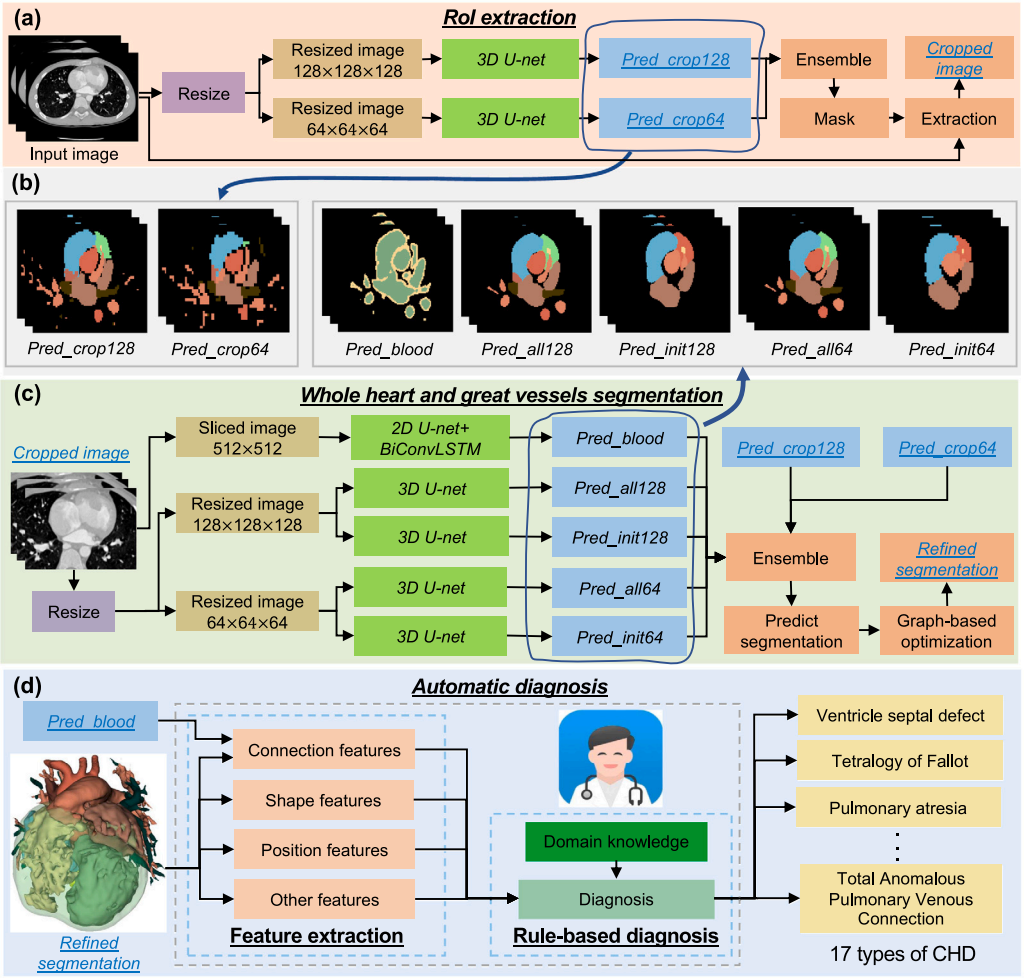

Fig. 2. Detailed structure of our AI system. It includes three modules: (a) RoI extraction, (c) whole heart and great vessels segmentation, and (d) automatic diagnosis.

check by radiologists. Each segmentation labeling takes about 1–1.5 h for resident physicians, and radiologists take about 10 min to check the labeling. For the diagnosis of RAA, we additionally labeled all the spines in 20 CT images to train another network to extract the features of the related position of AO and the spine in each 2D slice.

4. Methods

4.1. Overview

In clinical routine, diagnosis of CHD based on CT images is performed by two radiologists: a junior cardiovascular radiologist to write the report with the preliminary diagnosis and description, and a senior cardiovascular radiologist to check the report (Geijer and Geijer, 2018). Such a process can train the junior radiologists, save the time of senior radiologists, and at the same time ensure the report quality. Diagnosis of CHD requires three skills: anatomy identification, connection identification, and diagnosis recognition. The first task needs anatomy knowledge to locate structures and vessels, which may be time-consuming as some parts need extensive boundary tracing to determine their categories. The second task needs abundant experience to decide whether different parts are connected, which is challenging and time-consuming especially when CT images are of low quality or with large slice thickness. The third task is a complex decision process, where comprehensive clinical experiences and knowledge are needed to determine the CHD type with information/features extracted from the first two tasks.

Mimicking the workflow of radiologists, our AI system includes three modules: region of interest (RoI) extraction, whole heart and great vessel segmentation and automatic diagnosis (Fig. 1). The first two steps (Figs. 1(b)(d), 2(a)(c) and 3) obtain a multi-class segmentation of the heart and great vessels, which corresponds to anatomy identification and connection identification (Sections 4.2 and 4.3), while the step in Figs. 1(e) and 2(d) performs the diagnosis based on the segmentation, which corresponds to diagnosis recognition based on diagnostic criteria (Section 4.4).

4.2. RoI extraction

The RoI extraction module (Fig. 2(a)) extracts the target region in a CT image containing the heart and great vessels for efficiency. It uses two 3D U-net (Payer et al., 2017b) with different input sizes to segment the CT image with 10 anatomical structures including LV, RV, LA, RA, Myo, AO, PA, PV, SVC, and IVC, and majority voting ensemble is used for fusion. The ensemble of their outputs including $Pred\_crop128$ (output for an input size of 128^3) and $Pred\_crop64$ (output for an input size of 64^3) is used to obtain a cuboid crop. The detailed network structure is shown in Fig. 4(a). The input size of the 3D U-net networks includes $64 \times 64 \times 64$ and $128 \times 128 \times 128$, which are mainly for the ensemble on various scales thus with improved accuracy. Note that there may exist an optimal resolution for the training, and a larger resolution does not mean better performance (Sabottke and Spieler, 2020; Rukundo, 2023). Thus, we just performed ensemble on inputs with two resolutions considering the serious variations of structures

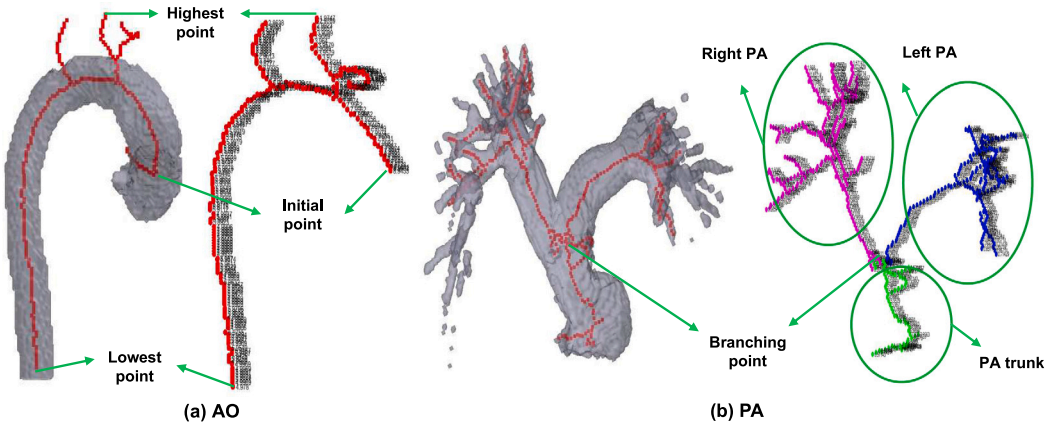

Fig. 3. Illustration of graph based optimization in (a) AO and (b) PA analysis. In (a), the 3D visualization of AO and the corresponding graph of the trunk are shown. The number near the red points of the graph represents the radius of AO at the point. By further analyzing the graph, more features of AO, e.g., how the radius varies along with AO, can be obtained. The same process is also performed on PA. In addition, the PA's graph is further divided into three parts: the trunk part, the left PA part, and the right PA part for further analysis.

and contrasts to obtain a robust performance. The number of resolution levels is 5, and the number of batch sizes is 64 and 48 for the input sizes of $64 \times 64 \times 64$ and $128 \times 128 \times 128$, respectively. We have made several improvements as follows. First, we adopt weighted Dice loss (Milletari et al., 2016) and cross-entropy loss to enhance the performance. Particularly, the weights of the background, LV, RV, LA, RA, AO, PA, MYO, SVC, IVC, PV are 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, respectively, for the segmentation of all anatomical structures. Second, we perform position correlation using several convolutional layers (Payer et al., 2017b), and the final output is a point-wise multiplication of the U-net output and the output of the position correlation process. Third, we adopted 3D instance normalization (Ulyanov et al., 2016) after each convolutional layer to mitigate the ingredient missing problem.

4.3. Whole heart and great vessel segmentation

The whole heart and great vessel segmentation module then performs multi-task and multi-resolution segmentation including high-resolution blood pool segmentation, heart and initial part of great vessels segmentation, and whole heart and great vessel segmentation (Fig. 2(b). The corresponding segmentation outputs includes: (1) *Pred_blood* (output of 2D U-Net+BiConvLSTM with 2D sliced images as input for blood pool segmentation), (2) *Pred_all*128 (output of 3D U-Net with an input size of 128^3 for whole heart and great vessel segmentation), (3) *Pred_init*128 (output of 3D U-Net with an input size of 128^3 for heart and initial part of great vessels segmentation), (4) *Pred_all*64 (output of 3D U-Net with an input size of 64^3 for whole heart and great vessel segmentation), (5) *Pred_init*64 (output of 3D U-Net with an input size of 64^3 for heart and initial part of great vessels segmentation), (6) *Pred_crop*64, and (7) *Pred_crop*128, and their ensemble can produce a multi-class segmentation including 10 anatomical structures Fig. 5(a) and (c). In the ensemble, we just format all the outputs with different resolutions and sizes with the same location, resolution and size. Particularly, we first extracted the RoI (obtained in the RoI extraction module) parts of each segmentation result, which are then resized to 1283. Then, weighted voting (*Pred_crop*128: *Pred_crop*64: *Pred_all*128: *Pred_init*128: *Pred_all*64: *Pred_init*64 = 1: 1: 2: 2: 2: 2) is performed to get an overall segmentation. Next, the blood pool and its boundaries (*Pred_blood*) corresponding to the overall segmentation are removed, and the rest are assigned to the ten classes using region growing in which the results in the overall segmentation are the seed points. Finally, the largest islands in each class are regarded as the main part in the class using connection analysis, and other small islands are assigned according to their connection to other classes using major voting and domain knowledge.

Fig. 4 shows the network structure of the adopted 3D U-net and the combination of 2D U-net and BiConvLSTM. Two 3D U-net networks are used for segmentation of all classes, and segmentation of chambers and initial parts of great vessels, respectively. The combination of 2D U-net and BiConvLSTM is used for blood pool segmentation. Note that we have also performed the experiment using patch based sampling approach (Feng and Meyer, 2017) using 3D U-Net for blood pool segmentation, and we observe that the combination of 2D U-net and BiConvLSTM obtains a slightly higher performance compared with the patch based sampling approach. A possible reason is the serious variations of structures and contrasts in CTs of CHD, and radiologists usually perform zoom-in and zoom-out to find the boundaries. Thus, a large scale should be used to recognize the boundaries. So we adopted the combination of 2D U-net and BiConvLSTM for segmentation of blood pool and its boundaries. Particularly, the number of resolution levels of the 2D U-net is 8, and the BiConvLSTM network has 2 layers, each of which has 32 channels, and a sequence length of 7. Note that the blood pool is a combination of LV, RV, LA, RA, AO, PA, PV, SVC, and IVC. The initial parts of great vessels play a critical role in the diagnosis, and thus we design a particular task for it, and we set the input size for segmentation of the blood pool and its boundaries to 512×512 to obtain a high-resolution of the segmentation. We adopted the same optimizations as discussed in Section 4.2. In the configuration of weighted Dice loss, the weights of the background, LV, RV, LA, RA, MYO, and initial parts of great vessels are 1, 2, 2, 2, 2, 2, and 3, respectively, for the segmentation task of initial parts of great vessels. The weights of the background, the blood pool, and its boundary are 1, 2, and 16, respectively, in the blood pool segmentation task. In addition, considering the fact that 2D U-net+BiConvLSTM is rather memory-consuming, we divided it into 2D U-net and BiConvLSTM and trained them individually. To mitigate the loss of features, we selected the second last layer with a channel number of 32 as the input to the BiConvLSTM layers to preserve features as much as possible.

As great vessels usually have large variations, we performed graph-based segmentation optimization as shown in Fig. 3. First, we extracted the centerline of these vessels to form a graph in which each center point corresponds to one graph point. Particularly, if two center points are adjacent to each other, the two are connected in the graph. The radius of the inscribed sphere in the segmentation centered at the graph

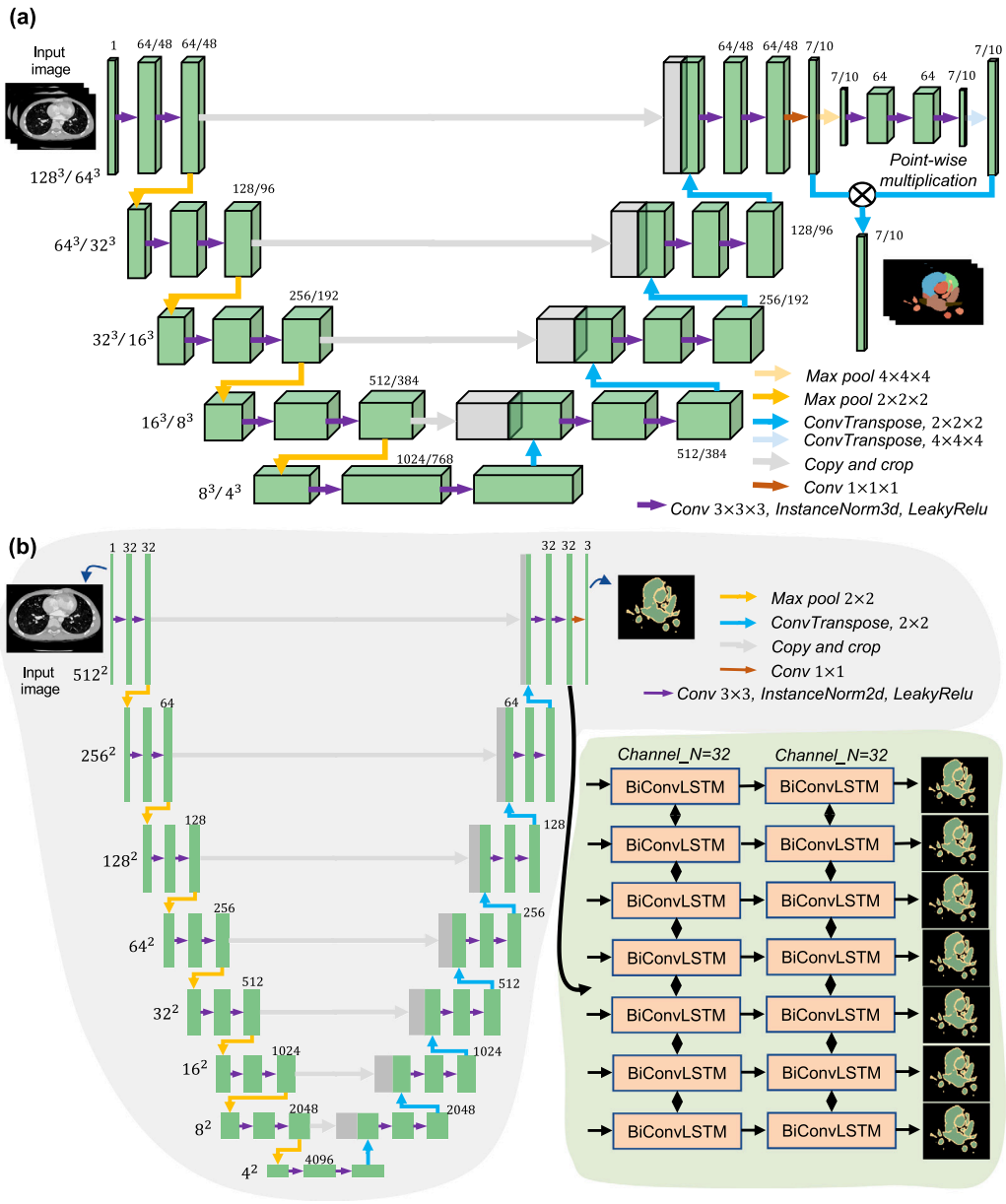

Fig. 4. Network structures of (a) 3D U-net for 3D image segmentation and (b) 2D U-net+ConvLSTM for 2D image segmentation. There are six instances of (a) and only one instance of (b). In (a), there are two kinds of inputs size (64 × 64 × 64 and 128 × 128 × 128) and two corresponding initial channel numbers (64 and 48). There are five resolution levels, and the number of channels doubles when the resolution level increases by one. There are two kinds of output classes: 11 for all classes and 7 for chambers and initial parts of great vessels and myocardium.

point is obtained as its weight, and the connection weight is defined as the average of the weights of the associated two points. Then, graph based optimization is used to obtain the main trunk of great vessels. For AO, three graph points are extracted including the initial point, the highest point, and the lowest point. The initial point is extracted based on the fact that it is usually the nearest point to LV and RV, while the highest and the lowest points are extracted based on their positions on the AO segmentation. Two max-flow paths from the lowest point to the highest point, and the initial point to the highest point, can be obtained and then combined to find the graph of the main AO part (AO without branching vessels) and its corresponding segmentation. For PA, a similar process is performed, and the branching point (separating the PA trunk, the left PA, and the right PA) is extracted to find the graph of the main PA part (the PA trunk, the left PA, and the right PA without branching vessels) and its corresponding segmentation. Other small branch vessels are then analyzed to be assigned to AO or PA based on their connection to AO and PA. In order to fulfill the above analysis, the initial parts of AO and PA are segmented as shown in Fig. 2(b) and Fig. 5(b) and (d).

4.4. Automatic diagnosis

The automatic diagnosis module analyzes the segmentation to perform CHD diagnosis. Due to the high structural variations, deep learning-based end-to-end diagnosis in a black-box style can hardly work. We instead make use of domain knowledge and define features for each CHD type. Particularly, radiologists perform CHD diagnosis based on related features and diagnostic criteria (Mazur et al., 2013; Adebo, 2021). Following this principle, we extracted related features (Table 2) by extensive discussion among our computer scientists, radiologists, and surgeons. These features are grouped into four types: connection, ratio, stenosis, relative position, and number of islands (illustrated in Fig. 6), where morphology analysis are performed extensively using graph based analysis as discussed in Fig. 3. Note that ratio and stenosis

Table 2
List of extracted features which are used for CHD diagnosis based on diagnostic criteria (Mazur et al., 2013; Adebo, 2021).

Disease group	Features and definition
VSD, SV ASD, SA	*conn_LV_RV*: whether LV and RV are connected *conn_LV_AO*: whether LV and AO are connected *conn_RV_AO*: whether RV and AO are connected *conn_RV_PA*: whether RV and PA are connected *conn_LV_RA*: whether LV and RA are connected *conn_RV_LA*: whether RV and LA are connected *conn_LA_RA*: whether LA and RA are connected *conn_LV_PA*: whether LV and PA are connected
ToF, DORV, TGA, TAC, PuA	All features for the disease group of VSD, SV, ASD, and SA *n_island_init_part*: number of islands connected to AO or PA in the initial parts of great vessels *ratio_PA_to_AO*: the ratio of the number of pixels of AO to PA within the initial part of great vessels *ratio_PA_to_low_AO*: the ratio of the radius of the lowest point of AO to the largest radius of the points of the PA trunk
RAA	*posi_AO_to_spine_arch*: the difference of the center of AO and that of the spine at the slice with the AO arch *posi_AO_to_spine_low*: the difference of the center of AO and that of the spine at the slice with the lowest AO points *posi_AO_to_spine_middle*: the difference of the center of AO and that of the spine at the middle slice of that with the AO arch and that with the lowest AO points
APVC	*conn_PV_LA*: whether PV and LA are connected *conn_PV_RA*: whether PV and RA are connected *conn_PV_SVC*: whether PV and SVC are connected *conn_PV_IVC*: whether PV and IVC are connected
IAA, AAH, CoA	*n_island_AO*: number of islands of the main part of AO *stenosis_AO_arch*: whether there exist a stenosis part in the arch *stenosis_desending_AO*: whether there exist a stenosis part in the descending AO (not include AO arch)
DSVC	*n_island_SVC*: number of islands of SVC *n_island_SVC_to_RA*: number of islands of SVC that is connected to RA *n_island_SVC_to_LA*: number of islands of SVC that is connected to LA *n_island_SVC_to_AO*: number of islands of SVC that is connected to AO *n_island_SVC_to_PA*: number of islands of SVC that is connected to PA
PDA	*conn_ao_pa*: whether AO and PA are connected
PS	*stenosis_PA_trunk*: whether there exist a stenosis part in the PA trunk *stenosis_PA_valve_upper*: whether there exist a stenosis part in the valve *stenosis_PA_valve*: whether there exist a stenosis part in the upper part of the PA valve *stenosis_PA_valve_down*: whether there exist a stenosis part below the part of the PA valve

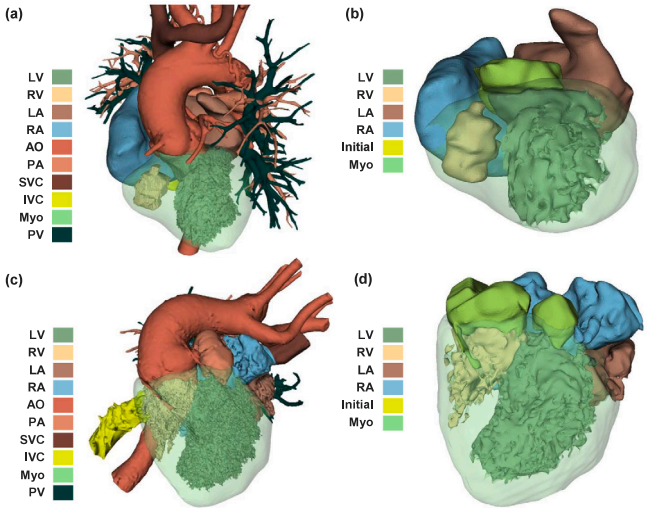

Fig. 5. Illustration of the labels of 10 anatomical structures and initial parts of great vessels. Examples of (a, c) labels with all 10 classes and (b, d) their corresponding labels with LV, RV, LA, RA, initial parts of great vessels, and MYO. Note that usually there are two separate parts (see (d)) in the initial part of great vessels as both AO and PA have initial parts, and there may exist only one (see (b)) if some diseases like PuA and TAC happen.

features work in the same manner which obtains a ratio to determine whether some feature exists. A threshold is used in the process, e.g., the ratio is larger than the threshold, then some feature exist, otherwise not. Such threshold is determined by an overall consideration of clinical practice, radiologists and surgeons. Automatic features extraction is then performed to obtain these defined features.

Note that usually the features defined by radiologists and surgeons cannot be directly implemented by computer scientists as some definitions are quite subjective. For example, the stenosis part in the descending AO and the AO arch indicates the existence of IAA and AAH, respectively. However, how to define stenosis is quite subjective, and there are many cases that radiologists cannot achieve consensus on whether a stenosis part should be reported (see Fig. 7(g-l)). The same problem also exists for the detection of the stenosis part of PA which is a critical feature for diagnosis of ToF, and we introduce some other parameters including the radius of the lowest points of AO to make a precise definition of "PA stenosis". Extraction of other features follows the same manner. As the segmentation is limited to detecting small vessels and areas with low contrast, we extract other features to assist the diagnosis. For example, in the diagnosis of DSVC, the number of islands of SVC that are connected to RA and LA are calculated and considered.

5. Experiments

5.1. Experimental setting

Totally there are seven segmentation networks in the first and second stages of our AI system, which can be divided to two types of networks: 2D U-net (Ronneberger et al., 2015) + BiConvLSTM (Song

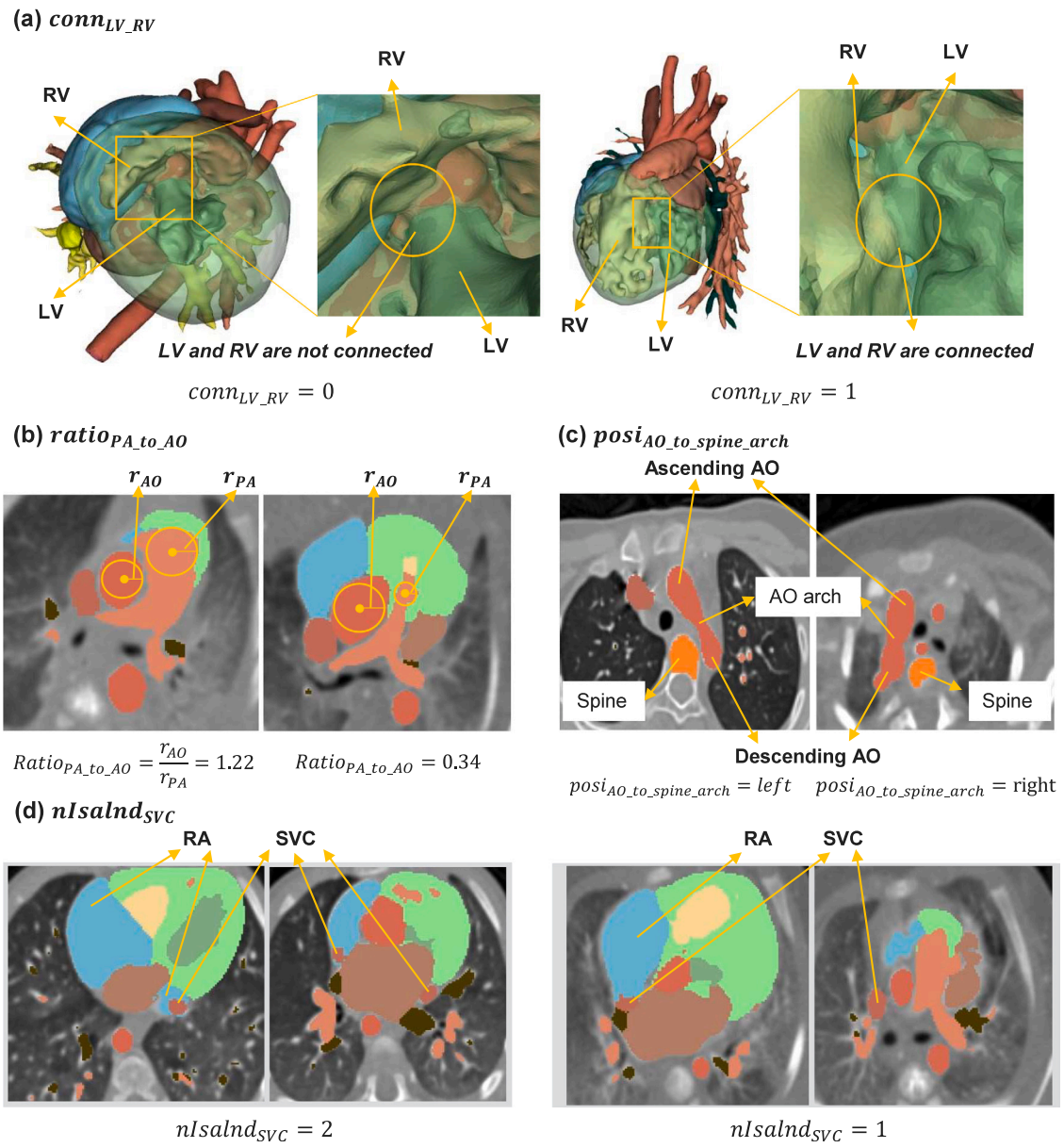

Fig. 6. Illustration of diagnosis related features. Examples of diagnosis features including (a) $conn_{LV\_RV}$, (b) $ratio_{PA\_to\_AO}$, (c) $posi_{AO\_to\_spin\_RV}$, and (d) $conn_{LV\_RV}$. Other diagnosis features in Table 2 works in the same manner.

et al., 2018) for 2D image segmentation and 3D U-net (Çiçek et al., 2016) for 3D image segmentation. The numbers of parameters of 3D U-net with an initial channel number of 64, 3D U-net with an initial channel number of 48, 2D U-net, and two-layer BiConvLSTM are 84.9 million, 51.2 million, 452.9 million, and 0.2 million, respectively. The total number of parameters of our AI system is 861.4 million. All the experiments were run on a machine with 4 Nvidia GPUs each with an 11 GB memory. We implemented all the networks using Pytorch (Paszke et al., 2019). Data augmentation is also adopted with the same configuration as in Seg-CNN (Payer et al., 2017b) for 3D U-net. Data normalization including random scaling, rotation, and deformation are adopted for 3D image segmentation tasks, while that including random rotation and flipping is used for the 2D image segmentation task. More details of the network parameters and hyperparameters are as shown in Table 3.

Both fixed training and testing sets and cross validation are used in our implementation. Overall, we adopted a fixed training and test sets as indicated in Table 1. The training set is for the training of the overall AI system including RoI extraction networks, segmentation networks, feature extraction and rule-based diagnosis. In the training process, RoI extraction networks and segmentation networks are trained and validated using three-fold cross-validation in which the training set (1282 CT images) is further divided into a sub training set (56.7%), a sub validation set (10%) and a sub test set (33.3%). All hyperparameters were selected using the segmentation validation dataset. The learning rate is 0.0002 for the first 50% of epochs, then 0.00002 for the next 25% of epochs, and 0.000002 afterward. Note that the design of feature extraction and rule-based diagnosis is based on the segmentation results by cross validation. In addition, three models are created during the three-fold cross-validation training for each segmentation network, and ensemble of the three models are adopted during inference for each segmentation network.

Three-fold cross-validation was adopted in the training and evaluation of each network, and the training data (1282 CT images) is divided into a segmentation training dataset (56.7%), a segmentation validation dataset (10%) and a segmentation test set (33.3%). Note that all hyperparameters were selected using the segmentation validation

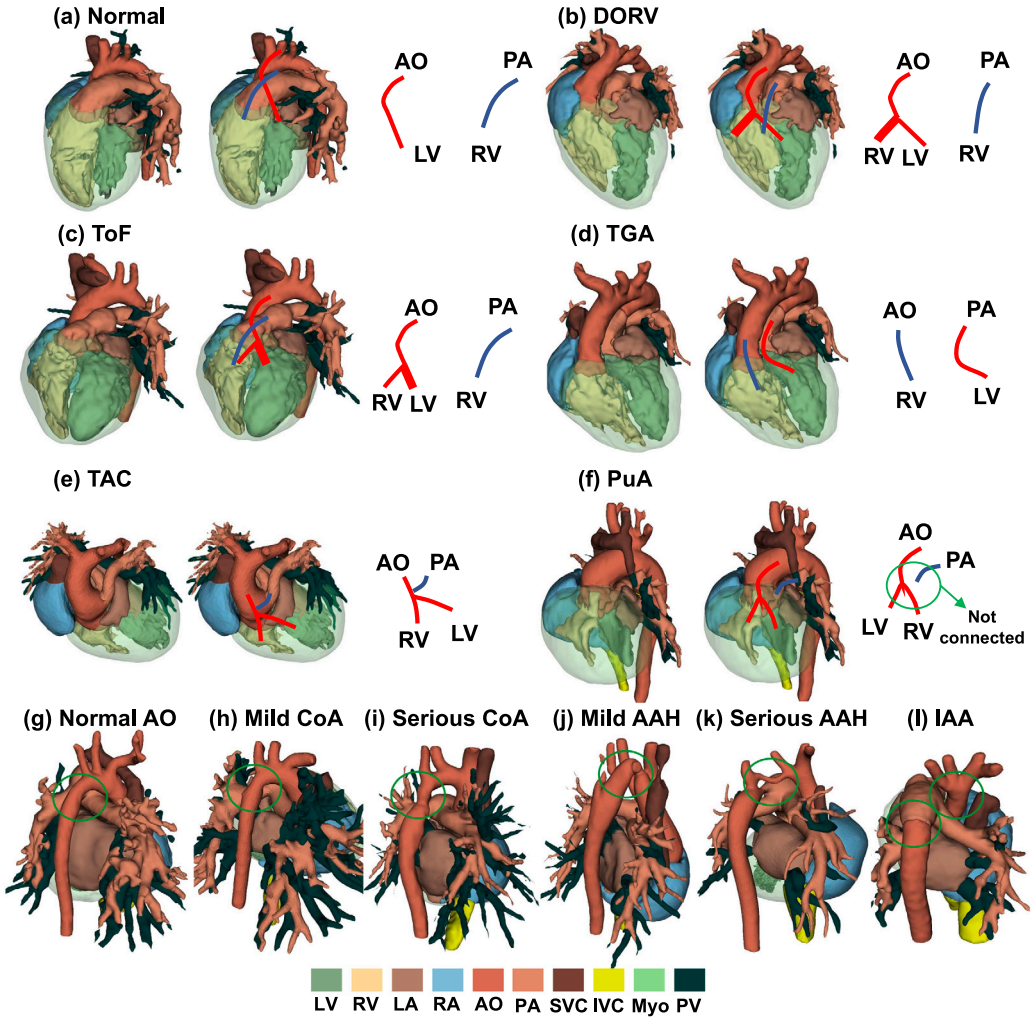

Fig. 7. Examples of types of CHD with large subjectivity. (a-f) ToF, DORV, PuA, TAC, and TGA are several stages of disease development, while (g-i) show the same situation of CoA, AAH and IAA. The current clinic taxonomy of CHD introduces much subjectivity and some confusion into current diagnosis and treatment in clinical practice. We can notice the complexity and challenges of CHD diagnosis for radiologists and our AI system.

Table 3
Network parameters and hyper parameters of the 3D U-net, 2D U-net and BiConvLSTM networks.

Networks	Input size	Output size	Initial channel #	# of scales	Initial learning rate	Batch size	Maximum # training epoch
3D U-net (Pred_crop64)	64^3	11×64^3	64	5	0.0002	4	15
3D U-net (Pred_crop128)	128^3	11×128^3	48	5	0.0002	4	15
3D U-net (Pred_all64)	64^3	11×64^3	64	5	0.0002	4	15
3D U-net (Pred_all128)	128^3	11×128^3	48	5	0.0002	4	15
3D U-net (Pred_init64)	64^3	11×64^3	64	5	0.0002	4	15
3D U-net (Pred_init128)	128^3	11×128^3	48	5	0.0002	4	15
2D U-net (blood)	512^2	3×512^2	32	8	0.0002	4	12
Two-layer BiConvLSTM (Pred_blood)	$7 \times 32 \times 512^2$	3×512^2	32	NA	0.0002	1	3

dataset. The learning rate is 0.0002 for the first 50% of epochs, then 0.00002 for the next 25% of epochs, and 0.000002 afterward.

We made the comparison with radiologists in the clinic process, which presents the real status of radiologists during working hours. Usually, a CT report of a patient is performed by two radiologists: a junior radiologist and a senior radiologist. The junior radiologist performs the initial diagnosis and description of the related types of CHD. While the senior radiologist checks the results and revises them if there exist errors. Usually, in our center it takes about 30–60 min for the junior radiologist to fulfill the report, and about 20–40 min for senior radiologists to check the report. Based on the working years, the 178 radiologists (Table 4) are divided into two groups: junior

Table 4
The experience and position of the 178 physicians for performance comparison with our AI system.

Radiologist index	Position	Years of experience
1–120	Junior radiologist in cardiovascular imaging	1–10
121–148	Senior radiologist n cardiovascular imaging	> 10

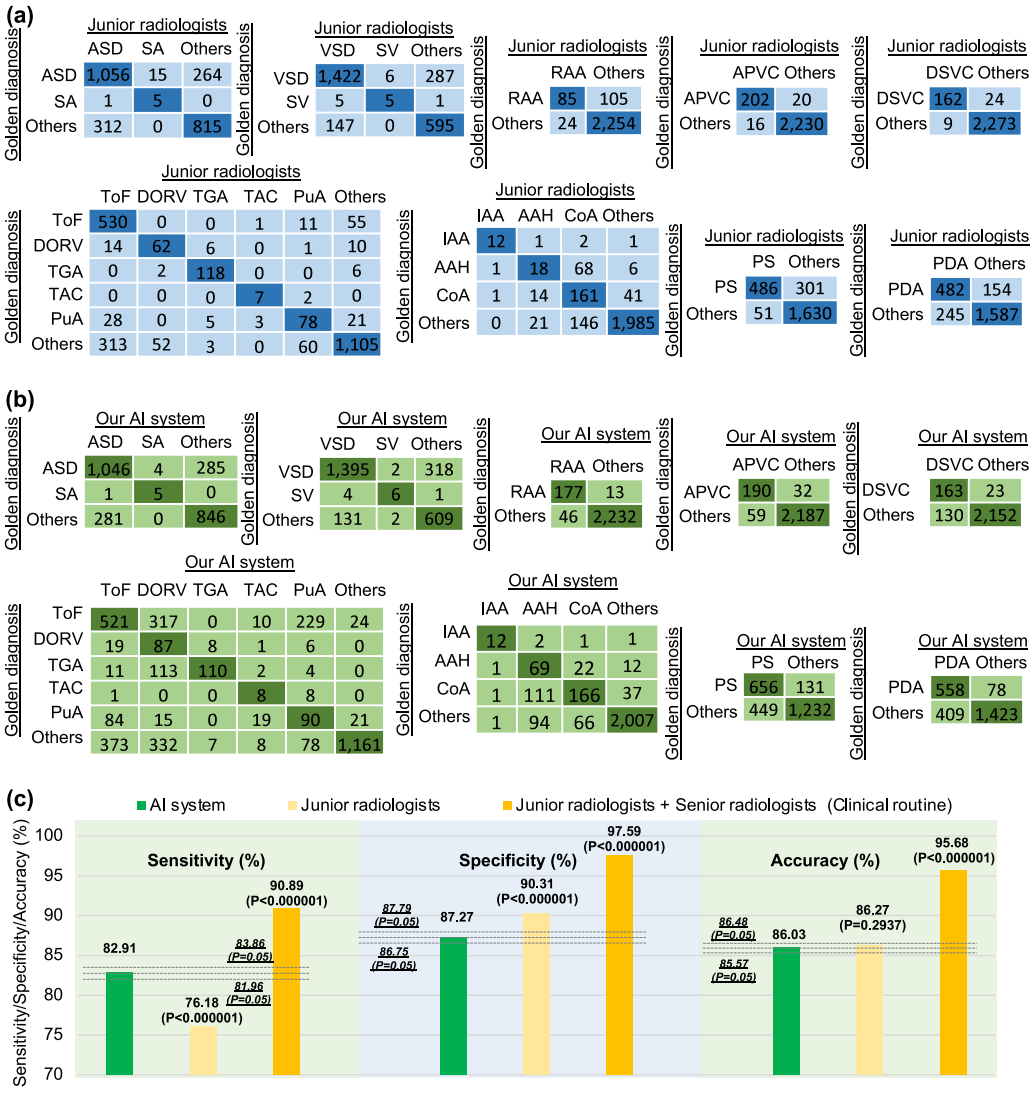

Fig. 8. Performance of our AI system and that of junior radiologists on an independent test dataset of 2468 patients covering 17 types of CHD. (a) Confusion matrices with patient numbers for CHD diagnosis of junior radiologists. For ease of discussion, we further divide the 17 types into 9 groups, where a patient cannot have more than one type of CHD within the same group. The numbers of correctly diagnosed cases are found on the diagonal. (b) Confusion matrices with patient numbers for CHD diagnosis of our AI system. (c) Sensitivity, specificity, and accuracy of our AI system, junior radiologists, and the combination of junior radiologists and senior radiologists. The 95% confidence intervals for the results of our AI system (83.86% and 81.96% on sensitivity, 87.79% and 86.75% on specificity, 86.48% and 85.57% on accuracy) are also marked using a two-sided exact binomial test, outside which the difference is considered to be statistically significant. The accuracy of our AI system is comparable ($P = 0.2937$) with that of junior radiologists with a significantly higher sensitivity ($P<0.000001$) and lower specificity ($P<0.000001$) than junior radiologists.

radiologists (1–10 years), and senior radiologists (> 10 years). We collected the CT report of the patients in the test dataset and extracted the 17 diagnosis classes for comparison. Note that each CT scan is only diagnosed by one radiologist, and thus each radiologist only performs diagnosis on a subset of all CT scans in our comparison.

In the statistical analysis, the two-side exact binomial test is adopted based on the assumption that the performance of our AI system and radiologists are constant but unknown (De Fauw et al., 2018). Suppose every associated type of CHD in CT images is correctly diagnosed with a probability of P_{ai}, and correctly diagnosed by radiologists with a probability of $P_{radiologist}$. For N cases (one type of CHD in a CT image is defined as a case, thus there are more cases than CT images), the number of correctly diagnosed cases k is then binomially distributed with $P(k) = B(k|p, N)$. Suppose that our AI system obtains k_{ai} correctly diagnosed cases, and radiologists achieve $k_{radiologist}$ correctly diagnosed cases. Then, the probability that our AI system has a higher true performance than radiologists is defined as

$$P(p_{ai} > p_{radiologist}|k_{ai}, k_{radiologist}, N) = \frac{\int_0^1 B(k_{ai}|p_1, N)\int_0^{p_1} B(k_{radiologist}|p_2, N)dp_2dp_1}{\int_0^1 B(k_{ai}|p, N)dp\int_0^1 B(k_{radiologist}|p, N)dp}. \quad (1)$$

For all the experiments, a confidence level of 95% was used.

5.2. Diagnosis performance

The diagnosis performance is shown in Fig. 8. In the diagnosis of CHD, several types can exist in one patient. For ease of discussion, we further divide the 17 types into 9 groups (Fig. 8(a)), where a patient cannot have more than one type of CHD within the same group. Particularly, for the third (including T o F et al.) and fourth (including IAA et al.) groups, each case has two predicted classes either of which matching the golden diagnosis means correct prediction. Other groups

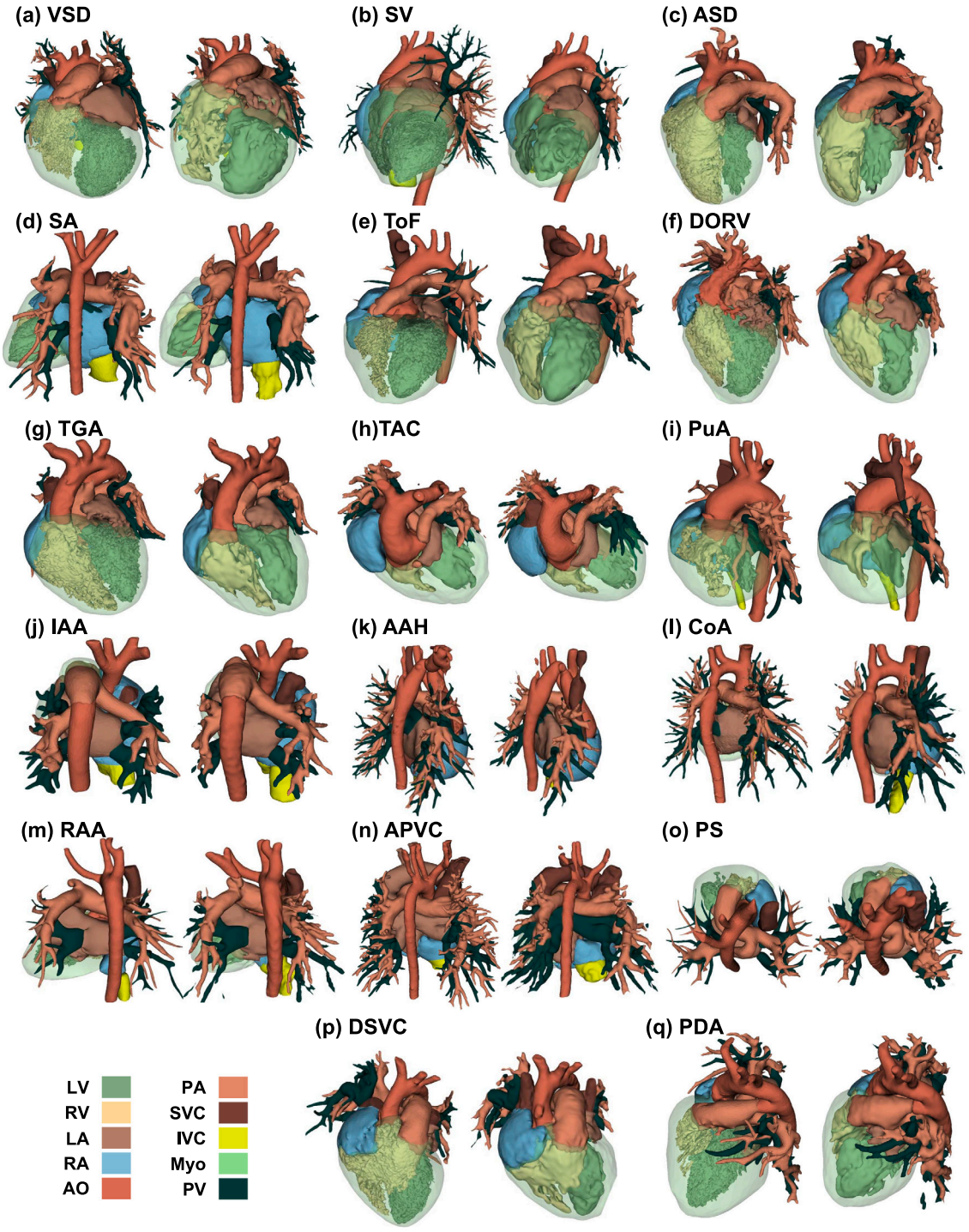

Fig. 9. 3D visualization of the hearts of our AI system. Examples of 3D visualization of the hearts of the ground truth (left) and in the whole heart and great vessels segmentation step of our AI system (right). Compared with the ground truth, our AI system can also provide a good segmentation of all parts. However, there are two main drawbacks of our AI system. First, some details are missing, e.g., the surfaces of LV and RV are smoothed. Second, some structures are missing, e.g., some thin PA and PV vessels are not recognized.

only make one prediction for each case. This also happens in clinic practice, and is due to the fact that there exists high subjectivity among these types in the third and fourth groups, which can be regarded as differential diagnosis (Arter and Jenkins, 1979). In the development of some diseases, the process is continuous and divided into several diseases. Thus, there is no precise boundary between these types of CHD, and in clinical practice, such misclassification is acceptable only if it can describe the anatomical structure and connection properly (Arter and Jenkins, 1979). The confusion matrix of the diagnosis performance of our AI system is shown in Fig. 8(b), and the overall sensitivity, specificity, and accuracy are 82.91%, 87.27%, and 86.03%, respectively (Fig. 8(c)). Deployed on a machine with four GPU cards (11 GB memory), our AI system can process each patient in about 15 s.

Segmentation performance is involved with the diagnosis performance of our AI performance. The quantitative segmentation results on the training dataset is shown in Table 5. The average Dice score, 95% Hausdorff distance and average surface distance are 73.67%, 6.84 mm, and 1.83 mm, respectively. The Dice score of LV (82.49%),

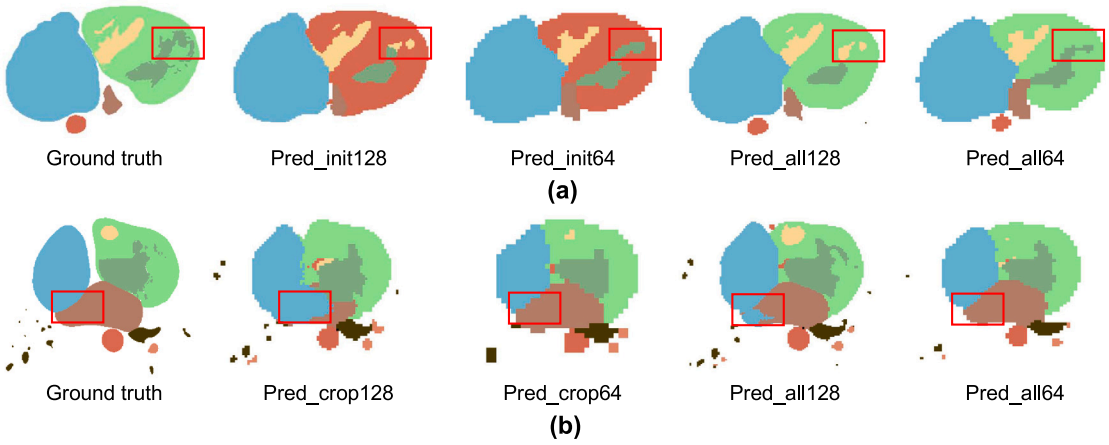

Fig. 10. Examples of segmentation results by different networks. Networks with low-resolution inputs (*Pred_all*64, *Pred_init*64 and *Pred_crop*64) can find some important structures which may be wrongly segmented by networks with high-resolution inputs (*Pred_all*128 and *Pred_init*128).

Table 5
Segmentation accuracy of our AI system on the training dataset. Three-fold cross-validation is adopted, and the training data (1282 CT images) is divided into a segmentation training dataset (56.7%), a segmentation validation dataset (10%) and a segmentation test set (33.3%). Dice score (DSC), 95% Hausdorff distance (HD)(mm), and average surface distance (ASD)(mm) are used for evaluation. Std is short for standard deviation.

	LV	RV	LA	RA	Myo	AO	PA	SVC	IVC	PV
DSC	82.49	78.66	83.04	86.42	84.08	75.09	80.45	52.14	51.26	63.10
Std	14.30	11.73	17.33	10.36	13.41	12.98	8.28	21.72	21.98	10.67
95% HD	4.00	4.44	4.48	5.15	4.73	8.58	3.42	14.85	10.91	7.88
Std	6.06	4.33	6.14	5.80	7.75	8.74	2.92	14.47	11.46	5.56
ASD	1.33	1.30	1.29	1.35	1.37	2.21	1.08	3.80	3.08	1.46
Std	2.59	1.34	2.41	1.66	3.16	3.69	0.84	7.69	7.20	1.26

RV (78.66%), LA (83.04%), RA (86.42%), Myo (84.08%), AO (75.09%), and PA (80.45%) is much higher than that of SVC (52.14%), IVC (51.26%), and PV (63.10%). For SVC and IVC, the main reason is that only a part of SVC and IVC are within the CT image, and only the initial part of them are labeled in the training dataset. Thus, the two parts are usually poorly segmented as shown in Fig. 9(d), (j), (m) and (n). However, such poor segmentation has no effect on the diagnosis accuracy as their connection to RA is usually correctly recognized. For PV, its boundary with LA is relatively hard to define, and it is too thin to be recognized precisely. Overall, the segmentation performance is moderate which is mainly due to the low contrast. The qualitative segmentation results are shown in Fig. 9. The overall segmentation is good, however, with many noticeable difference in details. For LV and RV, the details around the papillary muscles cannot be accurately recognized. For AO and PA, most of the thin vessels cannot be detected. We can also notice that the boundaries cannot be accurately determined, such as that between AO and LV, and that between PA and RV, etc. We can also notice the low Dice scores of IVC, SVC and PV, and the reasons are three-fold. First, they have relatively smaller volumes compared with others, and their boundaries with others are also more complex and hard to precisely detect. Thus, the segmentation errors at boundaries have a much larger influence on the Dice scores. Second, during CT image acquisition, the iodinated contrast agents are mainly used to enhance the contrast of LV, LA, RV, RA, AO and PA. Other parts have a relatively lower contrast, thus presenting a greater challenge for segmentation. Third, their boundaries with some errors have almost no effect on diagnosis. For example, longer IVC and SVC will not influence the features during feature extraction. Imprecise boundaries between PV and other parts are also not important, as their connections are critical for diagnosis.

The segmentation performance of our AI system on the training dataset without graph based optimization is shown in Table 6. The optimal segmentation performance of LV, RV, LA, RA, Myo, AO, PA, SVC, IVC, and PV is obtained by the *Pred_all*128 network, which is slightly lower than that with graph based optimizations shown in Table 5. We can also notice that *Pred_all*128 and *Pred_init*128 obtain higher performance than *Pred_all*64 and *Pred_init*64, respectively, which is mainly due to the resolution. However, by diving into the details, we find that *Pred_all*64 and *Pred_init*64 can find some important structures which are wrongly segmented by *Pred_all*128 and *Pred_init*128. As shown in Fig. 10(a), *Pred_init*64 and *Pred_all*64 can detect LV correctly, while *Pred_init*128 and *Pred_all*128 cannot. Fig. 10(b) shows the same phenomenon for a part of LA, which indicates that networks with low-resolution inputs is also critical for segmentation thus helpful for the final ensemble. *Pred_init*128 and *Pred_init*64 obtain similar performance on LV, RV, LA, RA, and Myo with *Pred_all*128 and *Pred_all*64, respectively, as their corresponding input resolutions and network structures are the same. We can notice that the accuracy of the initial parts of AO and PA is relatively low as their boundaries are not easily detected. The boundary of the blood pool is rather low, as a high weight is assigned to the boundary in the loss function. Note that the initial parts of AO and PA and the boundary of the blood pool are mainly used for feature extraction, so their segmentation accuracy is not critical.

Feature detection performance is shown in Table 7. Some thresholds are also given to make the features suitable for detection evaluation, which is also adopted in the diagnosis. For example, *n_island_init_part* larger than 1 indicates that there exist two large vessels, usually AO and PA; otherwise, only one large vessel exists (TAC happens). We can easily notice that there is a high correlation between the detection performance of these features and the final diagnosis performance, as a combination of these features determines the diagnosis. The overall performance is good, but there are still some low sensitivity and specificity values. For example, *posi_AO_to_spine_middle* obtains a low sensitivity as the middle part of the descending AO varies in position and has a weak correlation with RAA. Another low sensitivity of *stenosis_AO_arch* is due to the fact that the definition of stenosis is quite objective, making it relatively harder to detect. The low specificity of *conn_RV_AO* is because the connection parts between LV, RV, AO, and PA are hard to obtain their clear boundaries, especially when structure variations exist. Most of the errors happen when segmentation errors exist in critical areas (like the connection parts between LV, RV, AO, and PA), and please refer to Section 6 for more discussion of these cases.

5.3. Diagnosis performance compared with radiologists

Two groups of radiologists including 120 junior radiologists and 28 senior radiologists (Table 4) are considered. For the same 2468

Table 6
Segmentation accuracy of our AI system on the training dataset without graph based optimization. Three-fold cross-validation is adopted, and the training data (1282 CT images) is divided into a segmentation training dataset (56.7%), a segmentation validation dataset (10%) and a segmentation test set (33.3%). Dice score (DSC), 95% Hausdorff distance (HD)(mm), and average surface distance (ASD)(mm) are used for evaluation. BP is short for blood pool which includes LV, RV, LA, RA, AO, PA, SVC, IVC and PV. BBP stands for the boundary of the blood pool, and init. is short for initial parts of AO and PA. Std is short for standard deviation.

Networks		Anatomical structure												
		LV	RV	LA	RA	Myo	AO	PA	SVC	IVC	PV	BP	BBP	Init.
Pred_all128	DSC	79.91	72.56	81.92	85.26	83.57	73.41	79.57	53.76	52.59	63.42	–	–	–
	Std	18.49	22.79	19.18	13.19	8.95	13.53	9.41	21.40	23.62	10.96	–	–	–
	95% HD	4.90	5.09	5.42	5.77	4.91	8.70	3.51	15.06	10.03	7.78	–	–	–
	Std	7.20	6.37	7.23	6.91	5.27	5.54	3.33	14.55	11.00	6.02	–	–	–
	ASD	1.64	1.48	1.43	1.40	1.46	2.30	1.14	3.07	3.32	1.52	–	–	–
	Std	3.23	3.07	2.55	1.62	1.58	1.25	1.21	6.33	7.41	1.71	–	–	–
Pred_init128	DSC	80.06	72.33	81.40	85.19	80.35	–	–	–	–	–	–	–	64.64
	Std	18.20	22.71	19.84	13.21	9.64	–	–	–	–	–	–	–	14.78
	95% HD	4.76	5.07	5.71	5.91	3.55	–	–	–	–	–	–	–	6.98
	Std	6.98	6.00	7.83	7.38	3.49	–	–	–	–	–	–	–	5.50
	ASD	1.67	1.51	1.53	1.39	1.17	–	–	–	–	–	–	–	1.93
	Std	3.23	2.81	3.27	1.58	1.68	–	–	–	–	–	–	–	2.38
Pred_all64	DSC	76.66	69.34	78.69	83.13	80.05	69.97	77.62	51.10	50.70	54.35	–	–	–
	Std	17.87	21.57	18.97	13.32	8.59	10.75	8.87	20.07	22.22	9.79	–	–	–
	95% HD	4.87	5.08	5.60	5.72	4.37	7.46	3.82	14.82	9.84	9.07	–	–	–
	Std	6.86	5.89	7.63	7.32	5.44	5.96	4.76	14.48	11.29	6.44	–	–	–
	ASD	1.73	1.51	1.50	1.41	1.13	1.80	1.17	3.19	3.19	2.28	–	–	–
	Std	3.12	2.55	2.46	1.63	1.64	1.45	1.74	6.62	7.54	2.29	–	–	–
Pred_init64	DSC	76.50	69.11	78.50	83.05	77.38	–	–	–	–	–	–	–	62.34
	Std	17.79	21.35	18.73	12.90	9.26	–	–	–	–	–	–	–	13.56
	95% HD	5.03	5.29	5.55	5.89	3.80	–	–	–	–	–	–	–	7.03
	Std	6.96	5.86	7.27	7.21	4.04	–	–	–	–	–	–	–	5.21
	ASD	1.79	1.59	1.53	1.40	1.20	–	–	–	–	–	–	–	1.91
	Std	3.33	2.68	2.74	1.66	1.81	–	–	–	–	–	–	–	2.29
Pred_blood	DSC	–	–	–	–	–	–	–	–	–	–	86.26	28.98	–
	Std	–	–	–	–	–	–	–	–	–	–	6.58	7.21	–
	95% HD	–	–	–	–	–	–	–	–	–	–	3.51	4.32	–
	Std	–	–	–	–	–	–	–	–	–	–	2.27	2.40	–
	ASD	–	–	–	–	–	–	–	–	–	–	0.97	0.51	–
	Std	–	–	–	–	–	–	–	–	–	–	0.70	0.49	–

'–' indicates the corresponding result is not applicable.

Table 7
Detection performance of extracted features which are used for CHD diagnosis based on diagnostic criteria (Mazur et al., 2013; Adebo, 2021) **on the test dataset.** TP, FN, TN, and FP are short for true positive, false negative, true negative, and false positive, respectively.

Features	TP	FN	TN	FP	Sensitivity (%)	Specificity (%)	Accuracy (%)
conn_LV_RV	1,407	319	133	609	81.52	82.08	81.69
conn_LV_AO	2,268	15	73	112	99.34	60.11	96.43
conn_LV_PA	335	54	383	1,696	86.12	81.55	82.27
conn_RV_AO	1,208	276	421	563	81.40	57.22	71.76
conn_RV_PA	1,806	336	122	204	84.31	62.58	81.44
conn_LV_RA	5	1	4	2,458	83.33	99.84	99.80
conn_RV_LA	5	1	4	2,458	83.33	99.84	99.80
conn_LA_RA	1,056	285	281	846	78.75	75.07	77.07
n_island_init_part > 1	2,098	184	60	126	91.94	67.20	90.07
ratio_PA_to_AO >0.5	96	42	408	1,922	69.57	82.49	81.77
ratio_PA_to_low_AO >1.2	970	188	251	1,059	83.77	80.84	82.21
posi_AO_to_spine_arch (right)	177	13	46	2,232	93.16	97.98	97.61
posi_AO_to_spine_low (right)	154	36	56	2,222	81.05	97.54	96.27
posi_AO_to_spine_middle (right)	124	66	170	2,108	65.26	92.54	90.44
conn_PV_LA	2,344	2	12	110	99.91	90.16	99.43
conn_PV_RA	65	12	31	2,360	84.42	98.70	98.26
conn_PV_SVC	78	2	12	2,376	97.50	99.50	99.43
conn_PV_IVC	49	3	11	2,405	94.23	99.54	99.43
n_island_AO > 1	12	4	3	2,449	75.00	99.88	99.72
stenosis_AO_arch	69	35	57	2,307	66.35	97.59	96.27
stenosis_desending_AO	368	51	163	1,886	87.83	92.04	91.33
n_island_SVC >1	138	8	13	2,309	94.52	99.44	99.15
n_island_SVC_to_RA > 1	2,268	200	0	0	91.90	100.00	91.90
n_island_SVC_to_LA >1	26	6	51	2,385	81.25	97.91	97.69
n_island_SVC_to_AO > 1	14	5	41	2,408	73.68	98.33	98.14
n_island_SVC_to_PA >1	13	5	37	2,413	72.22	98.49	98.30
conn_ao_pa	558	78	409	1,423	87.74	77.67	80.27
stenosis_PA_trunk	696	91	449	1,232	88.44	73.29	78.12
stenosis_PA_valve_upper	243	52	178	1,995	82.37	91.81	90.68
stenosis_PA_valve	329	67	209	1,863	83.08	89.91	88.82
stenosis_PA_valve_down	155	34	156	2,123	82.01	93.15	92.30

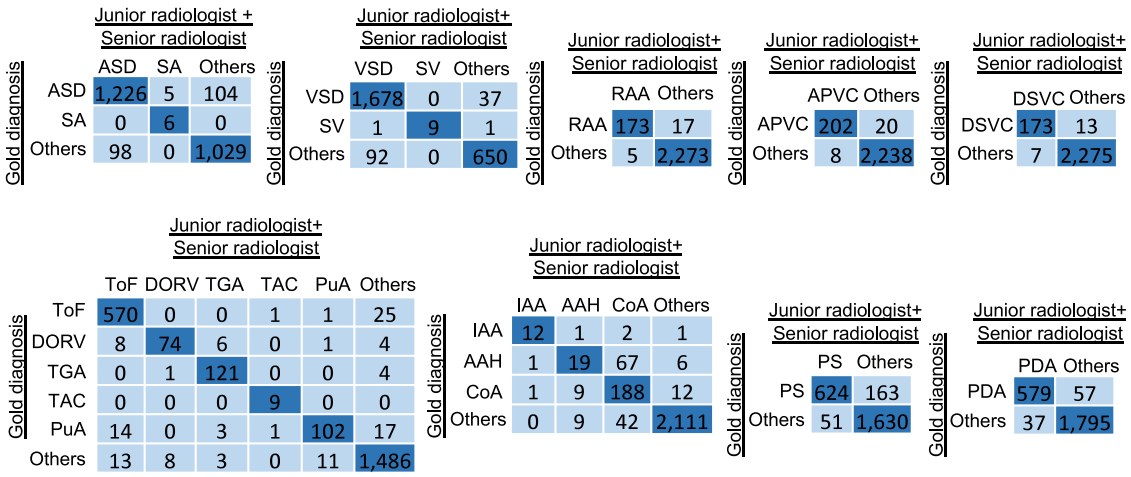

Fig. 11. Confusion matrices with patient numbers for CHD diagnosis of the combination of junior radiologists and senior radiologists. For ease of discussion, we further divide the 17 types into 9 groups, where a patient cannot have more than one type of CHD within the same group. The numbers of correctly diagnosed cases are found on the diagonal.

patients diagnosed by our AI system, the sensitivity, specificity, and accuracy achieved by junior radiologists are 76.18%, 90.31%, and 86.27%, respectively, while those of the combination of junior radiologist and senior radiologist are 90.89%, 97.59%, and 95.68%, respectively. Confusion matrices of the diagnosis of junior radiologists and the combination of junior radiologists and senior radiologists are shown in Fig. 8(a) and Fig. 11, respectively. Significant thresholds (83.86% for higher performance and 81.96% for lower performance for sensitivity, 87.79% for higher performance and 86.75% for lower performance for specificity, 86.48% for higher performance and 85.57% for lower performance for accuracy) were derived using a two-side exact binomial test incorporating uncertainty from both the radiologists and our AI system. The accuracy of our AI system is comparable with that of junior radiologists ($P = 0.2937$) and is lower than the combination of junior radiologists and senior radiologists ($P<0.000001$). Particularly, our AI system has a significantly higher sensitivity ($P<0.000001$) but lower specificity ($P<0.000001$) than junior radiologists. The combination of junior radiologists and senior radiologists obtain both higher sensitivity ($P<0.000001$) and specificity ($P<0.000001$) than our AI system. Considering that in our center junior radiologists usually takes 20-40 min to diagnose each patient, our AI system is much faster than junior radiologists. Some challenging cases misdiagnosed by our AI system is shown in Fig. 12 and Fig. 13.

5.4. Performance of the combination of AI and senior radiologists

Three senior radiologists (Table 4) are considered to perform the diagnosis refinement based on the diagnosis of our AI system. 300 CT images including 100 on 64 rows CT, 100 on 256 rows CT and 100 on dual source CT from the test dataset are used, and the results are shown in Fig. 14. The performance of our AI system is also reported as a reference which is much lower than the combination of junior radiologists and senior radiologists, and the combination of our AI system and senior radiologists. A similar phenomenon also exists in Fig. 8(c), where the performance of junior radiologists is much lower than the combination of junior radiologists and senior radiologists. The 95% confidence intervals for the results of the combination of junior radiologists and senior radiologists (87.23% and 91.99% on sensitivity, 98.04% and 99.08% on specificity, 96.49% and 97.75% on accuracy) are also marked using a two-sided exact binomial test, outside which the difference is considered to be statistically significant. The sensitivity, specificity and accuracy of our AI system is a significantly lower ($P<0.000001$, <0.000001, <0.000001, respectively) than that of the other two, which is expected. The sensitivity, specificity and accuracy of the combination of our AI system and radiologists is comparable ($P = 0.5198$, 0.4354, 0.9600, respectively) with that of the combination of junior radiologists and senior radiologists.

5.5. Diagnosis performance on new CT machines

There exists a large number of CT machines with varying manufacturers, models, and slice counts. It is therefore desirable that an AI-based diagnosis system can be machine-agnostic: when CT images are obtained from a machine that is different from where the training data is acquired, the performance can still be maintained. Our system should handle new machines well as it has three decoupled steps and the degradation in one step will have a limited impact on the overall performance. To evaluate the performance of our AI system on new machines, we divided the training and testing datasets according to CT machine types. The three stages in our AI system are trained using the dataset from one CT machine only and then evaluated using the dataset from all three machines. The results are reported in Fig. 15. As expected, the performance of our AI system is the highest when the training dataset and the test dataset are from the same machine. The accuracy of our AI system on 256-slice CT machines and dual-source CT machines is similar, and has a large gap with that on 64-slice CT machines. Compared with our AI system trained using datasets from all three machines, the one trained using the dataset from only one of them achieves slightly lower (less than 2%) accuracy when the test dataset is from a different machine, which shows that our AI system can be generalized to new CT machines with limited performance loss.

6. Discussion

As far as we know, this is the first AI system to diagnose CHD using a large-scale dataset, in which the processing mimics the clinic flow of radiologists. Particularly, unlike existing works using black-box networks either in an end-to-end or multi-stage fashion, our AI system combines deep learning and graph based optimization for diagnosis to tackle two grand challenges in CHD diagnosis: complex structural variations that cannot be handled by neural networks well, and a large number of types that require dedicated domain knowledge to classify. Our AI system is evaluated using CT images of 2468 patients who have undergone surgery to verify the diagnosis. It shows comparable performance with junior radiologists on 17 types of CHD and three types of CT machines. Particularly, our AI system achieves higher sensitivity and lower specificity than junior radiologists, which is acceptable as a false positive is less harmful than a false negative. In addition, our AI

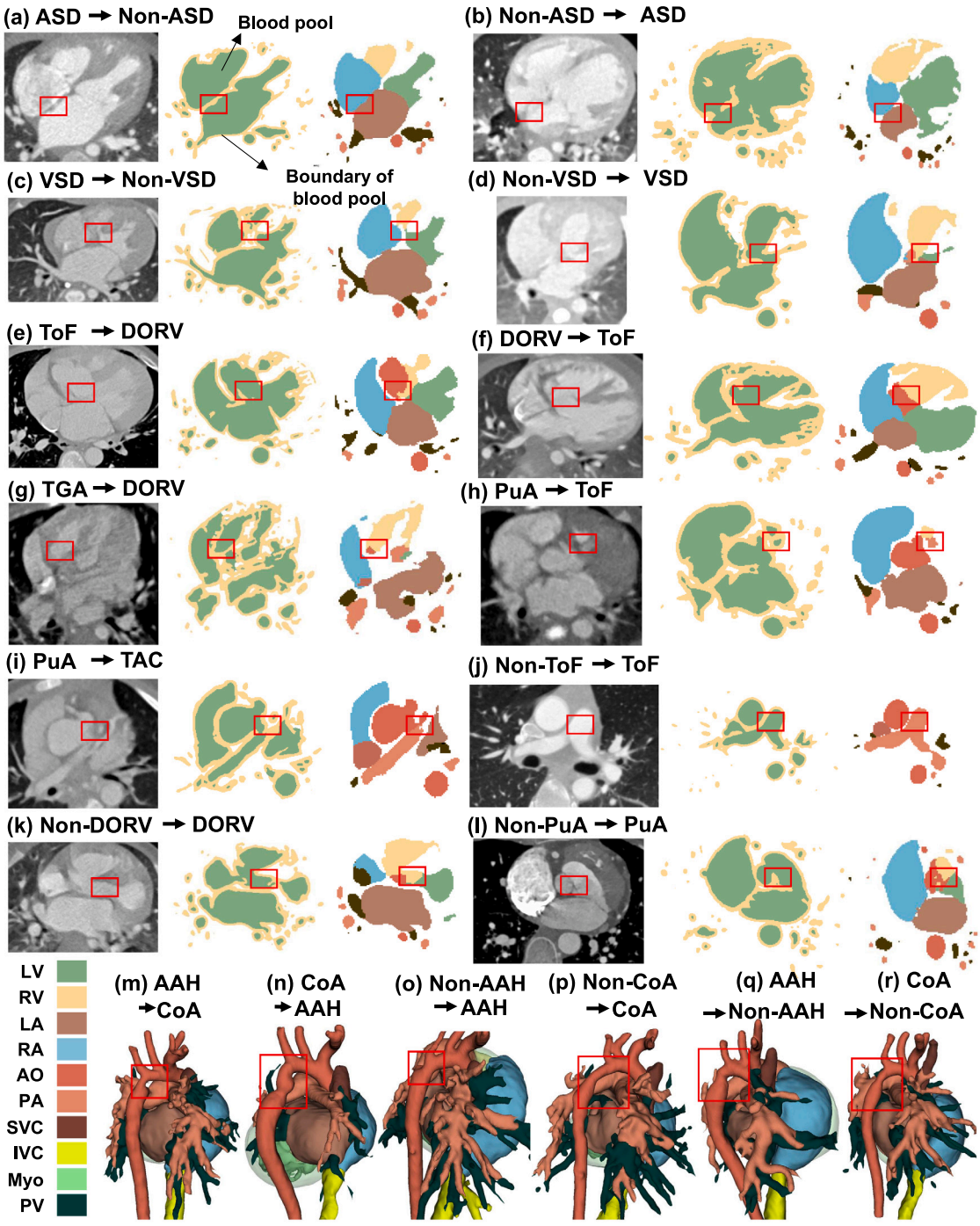

Fig. 12. Representative examples of challenging cases missed by our AI system. For each case, the associated 2D CT slice, 2D segmentation of the blood pool, 2D image or 3D model of the final whole heart and great vessels segmentation are provided. (a, c, d) low contrast and complex structure make accurate segmentation hard; (b) Downsampling may lose important information, e.g., some thin structures; (e-f) Complex structures make it hard to locate the boundary of each anatomy precisely; (g-l) Vagueness exists between several types of CHD, and thus, it is hard to precisely define the difference between them clearly. (m-r) Stenosis structures are a common feature in the diagnosis of CHD, and the subject/unclear definition of 'stenosis' introduces difficulty in detecting such features.

system can be generalized to new CT machines with limited accuracy loss. Our AI system is more efficient than junior radiologists with a largely reduced diagnosis time, which can help improve the clinical quality and save related costs at the same time.

Interpretation is a key point in the application of AI to healthcare (Castelvecchi, 2016; Tomsett et al., 2020). Our AI system can provide 3D models to radiologists for visualization. Examples of 3D models in 17 types of CHD are shown in Fig. 9, in which the anatomical morphology and their connections are well presented. In this way, clinicians can view these 3D models (Fig. 16) in many ways, e.g., 3D view of all parts, 2D view with a particular plane, 3D view with parts of interest, 2D view with the corresponding 2D CT image. The main benefits are as follows. First, a comprehensive description of the diagnosis is presented such that surgeons involved can get a much more vivid impression of the heart than just a report on the types of CHD. For example, the severity of aortic stenosis can be explained in detail with morphology in 3D visualization, and the hole in VSD can be visualized with precise morphology and relative position in reference to other

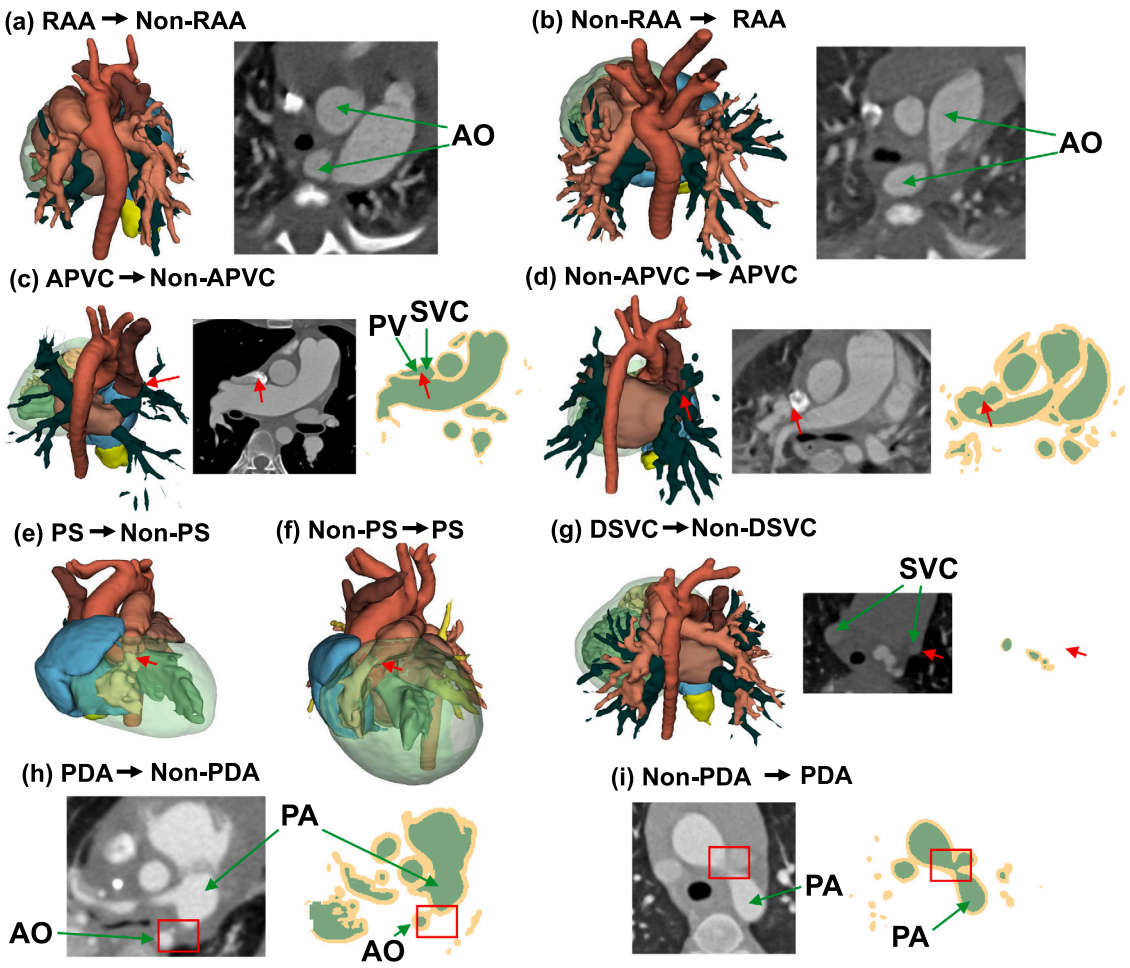

Fig. 13. Representative examples of challenging cases missed by our AI system (continued). For each case, the associated 2D CT slice, 2D segmentation of the blood pool, 2D image or 3D model of the whole heart and great vessels segmentation are provided. In (a-b), the descending AO is almost in the middle of the spine at the vertical direction, which introduces difficulty in determining the boundary between RAA and non-RAA. In (c-d), the boundaries between PV and SVC are not correctly recognized. In (e-f), the boundary between LV and PA is not correctly segmented. In (g), the boundary of one of the SVC vessels is with low contrast. In (h), the PDA vessel is too thin to recognize. In (i), the boundary is not correctly segmented.

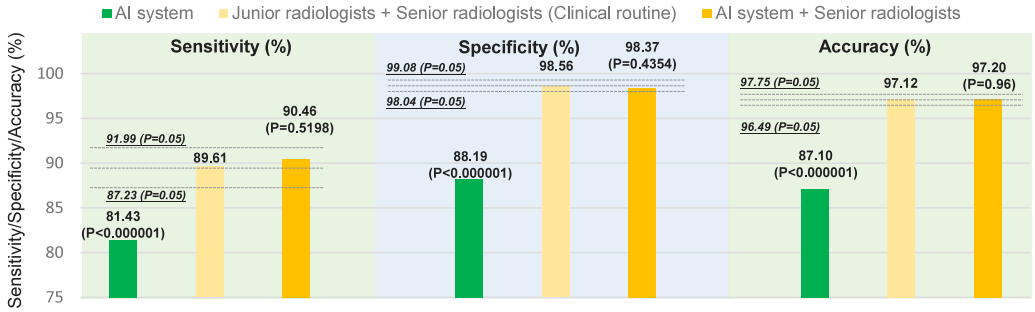

Fig. 14. Performance of our AI system, the combination of junior and senior radiologists, and the combination of our AI system and senior radiologists on a dataset of 300 patients. The 95% confidence intervals for the results of the combination of junior radiologists and senior radiologists (87.23% and 91.99% on sensitivity, 98.04% and 99.08% on specificity, 96.49% and 97.75% on accuracy) are also marked using a two-sided exact binomial test, outside which the difference is considered to be statistically significant. The sensitivity, specificity and accuracy of our AI system is a significantly lower (P<0.000001, <0.000001, <0.000001, respectively) than that of the other two, which is expected. The sensitivity, specificity and accuracy of the combination of our AI system and radiologists is comparable (P = 0.5198, 0.4354, 0.9600, respectively) with that of the combination of junior radiologists and senior radiologists.

important structures. Second, in surgery planning, surgeons can view the 3D heart model from any point and angle to get a much better understanding of the spatial relationship between each structure. This procedure can help surgeons get better prepared. Third, 3D models in 3D visualization can help improve the accuracy of outcome prediction, which is important especially for serious diseases like CHD (Yao et al., 2021). With 3D visualization, more parameters of the heart (e.g., the volume of the left ventricle, the length of the descending aorta (AO), the diameter ratio between the pulmonary artery and the right pulmonary artery) can be extracted and combined with other clinical data for better prediction (Yao et al., 2021). Overall, 3D visualization of heart models can help make a more personalized diagnosis, treatment, and outcome prediction for patients with possibly improved efficiency and accuracy.

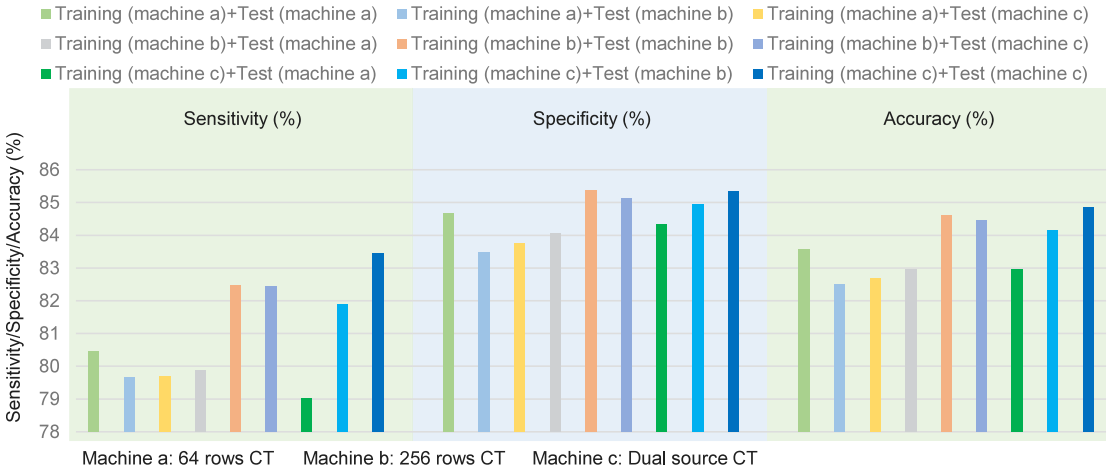

Fig. 15. Performance of our AI system when generalized to a new CT machine. Our AI system is trained on one CT machine and tested on other CT machines, and three machines and totally nine combinations are considered. When trained on machine a, our AI system has an accuracy loss of 0.89% and 1.09% on machine b and machine c respectively compared with that on machine a. When trained on machine b, our AI system has an accuracy loss of 1.64% and 0.17% on machine a and machine c respectively compared with that on machine b. When trained on machine c, our AI system has an accuracy loss of 1.88% and 0.69% on machine a and machine b respectively compared with that on machine c. The results indicate that our AI system can be generalized to new CT machines with limited performance loss (0.17%–1.88%). In addition, the sensitivity and specificity of our AI system are still significantly higher ($P<0.000001$) and lower ($P<0.000001$) than that of junior cardiovascular radiologists, respectively. Thus, our AI system has the potential to be deployed in clinic as sensitivity is more important than specificity for CHD diagnosis.

Challenging cases (Figs. 12 and 13) reveal various reasons that lead to the wrong diagnosis from our AI system. First, the crops used in 3D segmentation (with a resolution of 128 × 128 × 128) is downsampled from the original image with a resolution of $512 \times 512 \times (200-300)$ to make the computation feasible. Such downsampling may lose important information. For example, from Figs. 12(b) and 13(g-i) we can see that some small structures are missing, and disconnected vessels may appear to be connected to each other. Second, low contrast and complex structure make accurate segmentation hard. The low contrast in Figs. 12(a-d) and 13(c-d) makes the boundary hard for even experienced radiologists to recognize. The complex structure makes it hard to locate the boundary of each anatomical structure precisely. For example, in Fig. 12(e-f), the segmented boundaries of AO, LV, and RV are different from the ground-truth, which results in misdiagnosis. Third, vagueness exists between several types of CHD. For example, ToF, DORV, PuA, and TGA are several stages during disease development. Thus, as shown in Fig. 12(g-l), it is hard to precisely define the difference between them. Even experienced radiologists may have different opinions on the same patient (e.g., the same patient can be diagnosed with ToF or DORV). Fourth, stenosis structures are a common feature in the diagnosis of CHD; yet the degree of stenosis to which point the disease is considered to exist is usually rather subjective among radiologists (Figs. 12(m-r) and 13(e-f)). However, the subject or unclear definition of 'stenosis' introduces difficulty in detecting such features. The same situation also exists for other types of CHD, like RAA (Fig. 13(a-b)). Note that the above problem commonly exists in clinic practice and is known as differential diagnosis (Liu et al., 2020).

To tackle the above problems, efforts from both AI and medicine communities are crucial. In AI, highly accurate segmentation using high-resolution images is required especially in three aspects: low contrast areas that have a great impact on the overall structure, boundaries between different anatomies, and small structures. Domain knowledge and graph based optimization are extensively needed to tackle this problem. In medicine, clear and preferably quantitative definitions of each CHD type are desired to facilitate the automatic diagnosis. With our AI system, more features of the heart and great vessels can be obtained as desired, which can be used to establish more precise definitions of CHD types. In this way, precision medicine (Ginsburg and Phillips, 2018) on CHD can be achieved with more detailed types and thus possibly better treatment and outcome. In conclusion, deep collaboration between the two communities is necessary to advance the field.

Considering the need for efficient discussion and comprehensive interpretation of the diagnosis of complex diseases like CHD, our AI system can potentially improve the current workflow in several ways. First, our AI system can work as a junior radiologist to provide senior radiologists a variety of assistance including a 3D visualization model and diagnosis with extracted features. With such information, senior radiologists can check the diagnosis of simple cases easily while mainly focusing on challenging cases, which can lead to improved efficiency and accuracy and reduced cost. An illustration of graphical interface of our AI system is shown in Fig. 16, in which senior radiologists can view shapes and connections of each anatomical structure (Fig. 16(a)), diagnosis-associated features, anatomies' segmentation and corresponding 2D CT images (Fig. 16(b)). Second, the diagnosis and its 3D model can enable patients to visualize their heart structure and allow easier communication between them and clinicians. Third, surgeons can get a better understanding of the morphology, structures, and connections of the heart when viewing the 3D visualization model using virtual reality (VR). Traditionally, surgeons need to review the CT images in person which is timely and not intuitive. With 3D visualization models, they can view the heart in detail from any desired angle. Fourth, better outcome prediction can be possibly achieved as our AI system can extract more features such as McGoon index, total neopulmonary artery index, and pulmonary vein index. These features can be combined with other clinical data for potentially more accurate outcome prediction. Fifth, our AI system can facilitate education and training in CHD diagnosis and treatment. Overall, our AI system can bring tremendous benefits to the current clinic workflow.

Our AI system has several limitations. First, our study only uses retrospective data from a single center. For CTs in each type of CHD, retrospective data is used for the training of our AI system, and the collected data in the next years are used for validation. We further divided the data according to the CT machine types to mimic data collection from multiple sites, and the results indicate that our AI system is applicable to CT images collected from different machines. Though the retrospective data is collected in just one center, the data size is large considering its quantity (totally more than 3000) and the relatively low existence of CHD. Second, serious bias exists between different types of CHD in the collected dataset. Less-common types of CHD usually have a lower representation in the dataset, especially when surgery is

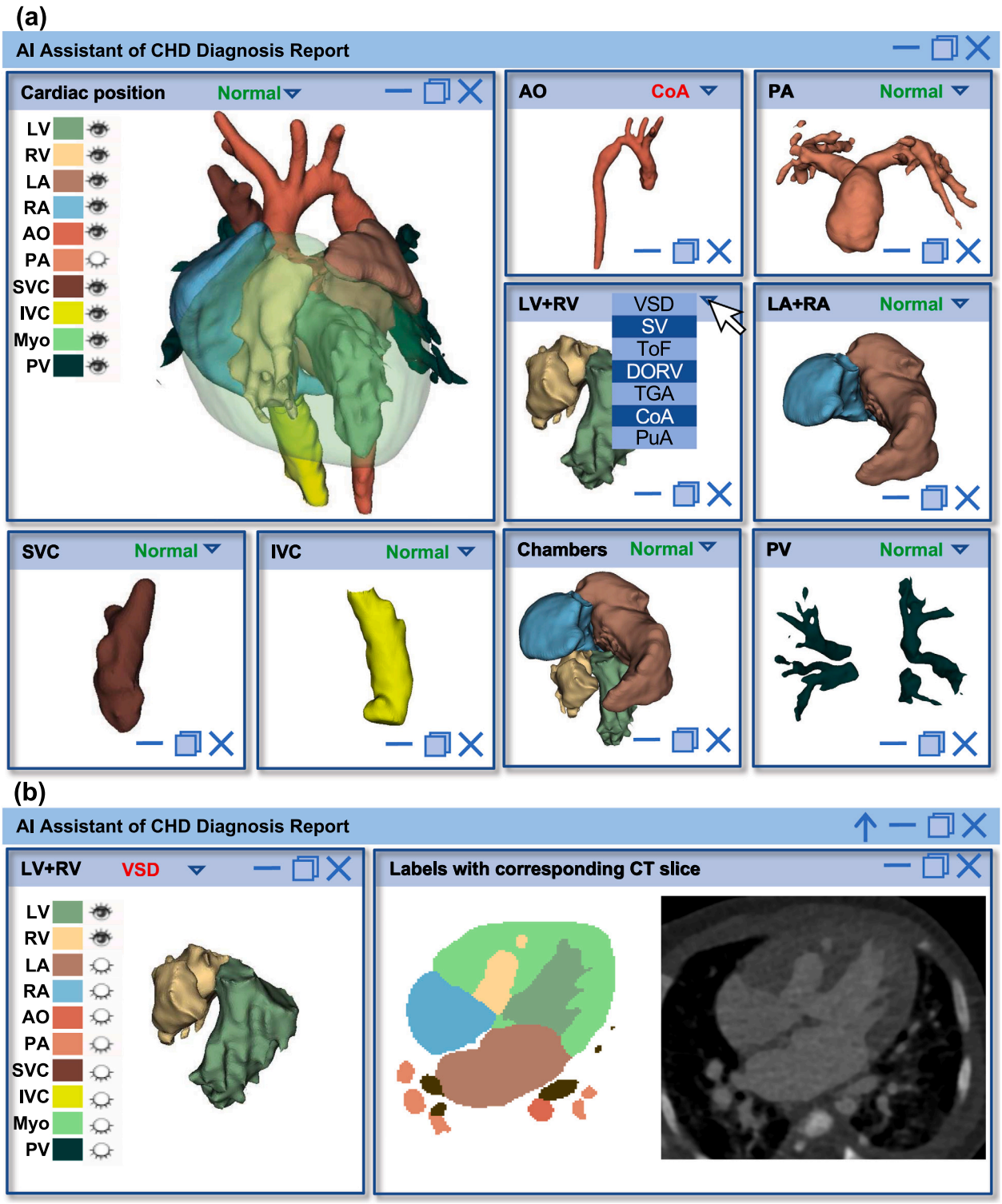

Fig. 16. Example of an interface for radiologists to check the diagnosis based on our AI system. Clinicians can view these 3D models in many ways including (a) 3D view of all parts, and (b) 3D view with parts of interest and 2D view with the corresponding 2D CT image. In this way, 3D visualization of heart models can help make a more personalized diagnosis, treatment, and possible outcome prediction for patients with improved efficiency and accuracy.

a prerequisite to have a case included in the dataset (so that the label can be ensured to be correct). Third, our AI system can only diagnose 17 types of CHD. Due to the complexity of CHD, some types are merged and some rare types are not included, which is partially because the quantity of these subtypes is quite limited. Detailed sub-types of CHD are not discussed. Forth, our AI system can only make the diagnosis but cannot provide a report including a detailed description of related types of CHD, e.g., the shape of AO, the structure of some anomalous vessels. Our AI system can provide a 3D model of the heart which may help to mitigate this problem. However, it still takes effort to integrate it into the current clinical practice. For example, we can obtain a 2D image from some angel and some point in the 3D model to replace the word description of the details of IAA in the report. To tackle the above problem, future work can validate our AI system in multiple sites in an

online manner. More data needs to be collected to allow our AI system to cover more rare types of CHDs or subtypes of CHDs, which would not be possible without contribution from the entire community. Fifth, the overall implementation is relatively complex which is not easy to deploy on resource-limited hardware, and many of the setup choices are just based on common configuration or experience. Thus, further analysis of hyper-parameters in our design including the number of networks, segmentation optimization strategies, the number of features, and diagnosis rules needs to be performed to make the system compact.

7. Conclusion

In this paper, we presented an AI system for CHD diagnosis based on CT images. Experiments on a dataset of more than 3750 CHD patients over 14 years demonstrates that our AI system achieves comparable performance with junior radiologists. In addition, as timely treatment is critical for the prognosis of many types of CHD, our AI system can facilitate automatic screening as our AI system has a higher sensitivity than junior radiologists. Our AI system also shows good results when applied to new types of CT machines. Further more, our AI system can potentially facilitate training of surgeons with 3D visualization of the heart. Our AI system opens up new opportunities for precision medicine research and treatment of the leading birth defect in the world, with a great potential to benefit developing regions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Science and Technology Planning Project of Guangdong Province, China (No. 2019B020230003), Guangdong Peak Project (No. DFJH201802), Science and Technology Projects in Guangzhou, China (No. 202206010049) and the National Natural Science Foundation of China (No. 62006050, No. 62276071), Guangdong Special Support Program-Science and Technology Innovation Talent Project (No. 0620220211), Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010157, 2022A1515011650), and Guangzhou Science and Technology Planning Project (No. 2021020 80188).

References

Adebo, D.A., 2021. Pediatric Cardiac CT in Congenital Heart Disease. Springer.

Arter, J.A., Jenkins, J.R., 1979. Differential diagnosis—prescriptive teaching: A critical appraisal. Rev. Educ. Res. 49 (4), 517–555.

Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., Satam, G., Pellikka, P.A., Enriquez-Sarano, M., Noseworthy, P.A., Munger, T.M., et al., 2019. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Med. 25 (1), 70–74.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging 37 (11), 2514–2525.

Bhat, V., BeLaVaL, V., Gadabanahalli, K., Raj, V., Shah, S., 2016. Illustrated imaging essay on congenital heart diseases: multimodality approach Part I: clinical perspective, anatomy and imaging techniques. J. Clin. Diagn. Res.: JCDR 10 (5), TE01.

Bonnichsen, C., Ammash, N., 2016. Choosing between MRI and CT imaging in the adult with congenital heart disease. Curr. Cardiol. Rep. 18, 1–10.

Brown, J.M., Campbell, J.P., Beers, A., Chang, K., Ostmo, S., Chan, R.P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J., et al., 2018. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. 136 (7), 803–810.

Castelvecchi, D., 2016. Can we open the black box of AI? Nat. News 538 (7623), 20.

Choi, A.D., Thomas, D.M., Lee, J., Abbara, S., Cury, R.C., Leipsic, J.A., Maroules, C., Nagpal, P., Steigner, M.L., Wang, D.D., et al., 2021. 2020 SCCT guideline for training cardiology and radiology trainees as independent practitioners (level II) and advanced practitioners (level III) in cardiovascular computed tomography: A statement from the society of cardiovascular computed tomography. Cardiovasc. Imaging 14 (1), 272–287.

Chu, Q., Jiang, H., Zhang, L., Zhu, D., Yin, Q., Zhang, H., Zhou, B., Zhou, W., Yue, Z., Lian, H., et al., 2020. CACCT: An automated tool of detecting complicated cardiac malformations in mouse models. Adv. Sci. 7 (8), 1903592.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Med. 24 (10), 1559–1567.

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al., 2019. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Med. 25 (10), 1519–1525.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Med. 24 (9), 1342–1350.

Erickson, B.J., 2021. Artificial intelligence in medicine: Technical basis and clinical applications. In: Artificial Intelligence in Medicine. Elsevier, pp. 19–34.

Feng, X., Meyer, C., 2017. Patch-based 3d u-net for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. (MICCAI).

Geijer, H., Geijer, M., 2018. Added value of double reading in diagnostic radiology, a systematic review. Insights Imaging 9 (3), 287–301.

Ginsburg, G.S., Phillips, K.A., 2018. Precision medicine: from science to value. Health Aff. 37 (5), 694–701.

Han, B.K., Lesser, A.M., Vezmar, M., Rosenthal, K., Rutten-Ramos, S., Lindberg, J., Caye, D., Lesser, J.R., 2013. Cardiovascular imaging trends in congenital heart disease: A single center experience. J. Cardiovasc. Comput. Tomogr. 7 (6), 361–366.

Han, B.K., Rigsby, C.K., Leipsic, J., Bardo, D., Abbara, S., Ghoshhajra, B., Lesser, J.R., Raman, S.V., Crean, A.M., Nicol, E.D., et al., 2015. Computed tomography imaging in patients with congenital heart disease, part 2: technical recommendations. An expert consensus document of the society of cardiovascular computed tomography (SCCT): endorsed by the society of pediatric radiology (SPR) and the North American society of cardiac imaging (NASCI). J. Cardiovasc. Comput. Tomogr. 9 (6), 493–513.

Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Med. 25 (1), 65–69.

Hollon, T.C., Pandian, B., Adapa, A.R., Urias, E., Save, A.V., Khalsa, S.S.S., Eichberg, D.G., D'Amico, R.S., Farooq, Z.U., Lewis, S., et al., 2020. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nature Med. 26 (1), 52–58.

Liang, H., Tsui, B.Y., Ni, H., Valentim, C.C., Baxter, S.L., Liu, G., Cai, W., Kermany, D.S., Sun, X., Chen, J., et al., 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature Med. 25 (3), 433–438.

Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al., 2020. A deep learning system for differential diagnosis of skin diseases. Nature Med. 26 (6), 900–908.

Lotter, W., Diab, A.R., Haslam, B., Kim, J.G., Grisot, G., Wu, E., Wu, K., Onieva, J.O., Boyer, Y., Boxerman, J.L., et al., 2021. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Med. 27 (2), 244–249.

Mazur, W., Siegel, M.J., Miszalski-Jamka, T., Pelberg, R., 2013. CT Atlas of Adult Congenital Heart Disease. Springer Science & Business Media.

Mei, X., Lee, H.-C., Diao, K.-y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al., 2020. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Med. 26 (8), 1224–1228.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. (3DV), IEEE, pp. 565–571.

Mori, S., Tretter, J.T., Spicer, D.E., Bolender, D.L., Anderson, R.H., 2019. What is the real cardiac anatomy? Clin. Anat. 32 (3), 288–309.

Nicoll, R., 2018. Environmental contaminants and congenital heart defects: A re-evaluation of the evidence. Int. J. Environ. Res. Public Health 15 (10), 2096.

Pace, D.F., Dalca, A.V., Brosch, T., Geva, T., Powell, A.J., Weese, J., Moghari, M.H., Golland, P., 2018. Iterative segmentation from limited training data: Applications to congenital heart disease. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 334–342.

Pace, D.F., Dalca, A.V., Geva, T., Powell, A.J., Moghari, M.H., Golland, P., 2015. Interactive whole-heart segmentation in congenital heart disease. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 80–88.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2017a. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 190–198.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2017b. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 190–198.

Pieper, S., Halle, M., Kikinis, R., 2004. 3D slicer. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821). IEEE, pp. 632–635.

Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P., Andersson, C.R., Macfarlane, P.W., Meira Jr., W., et al., 2020. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Commun. 11 (1), 1–9.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rukundo, O., 2023. Effects of image size on deep learning. Electronics 12 (4), 985.

Sabottke, C.F., Spieler, B.M., 2020. The effect of image resolution on deep learning in radiography. Radiology: Artif. Intell. 2 (1), e190015.

Shi, Z., Miao, C., Schoepf, U.J., Savage, R.H., Dargis, D.M., Pan, C., Chai, X., Li, X.L., Xia, S., Zhang, X., et al., 2020. A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. Nature Commun. 11 (1), 1–11.

Soenksen, L.R., Kassis, T., Conover, S.T., Marti-Fuster, B., Birkenfeld, J.S., Tucker-Schwartz, J., Naseem, A., Stavert, R.R., Kim, C.C., Senna, M.M., et al., 2021. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Sci. Transl. Med. 13 (581), eabb3652.

Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.-M., 2018. Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European Conference on Computer Vision. (ECCV), pp. 715–731.

Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al., 2020. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nature Commun. 11 (1), 1–9.

Stout, K.K., Daniels, C.J., Aboulhosn, J., Bozkurt, B., Broberg, C., Colman, J., Crumb, S.R., Dearani, J., Fuller, S., Gurvitz, M., et al., 2019. 2018 AHA/ACC guideline for the management of adults with congenital heart disease. Circulation 139 (14), e637–e697.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., Kaplan, L., 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns 1 (4), 100049.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

Van Der Linde, D., Konings, E.E., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J., Roos-Hesselink, J.W., 2011. Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis. J. Am. Coll. Cardiol. 58 (21), 2241–2247.

Wang, C., MacGillivray, T., Macnaught, G., Yang, G., Newby, D., 2018. A two-stage 3D unet framework for multi-class segmentation on full resolution image. arXiv preprint arXiv:1804.04341.

Xu, X., Wang, T., Shi, Y., Yuan, H., Jia, Q., Huang, M., Zhuang, J., 2019. Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 477–485.

Xu, X., Wang, T., Zhuang, J., Yuan, H., Huang, M., Cen, J., Jia, Q., Dong, Y., Shi, Y., 2020. Imagechd: A 3d computed tomography image dataset for classification of congenital heart disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 77–87.

Yang, X., Bian, C., Yu, L., Ni, D., Heng, P.-A., 2017. Hybrid loss guided convolutional networks for whole heart parsing. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 215–223.

Yao, Z., Hu, X., Liu, X., Xie, W., Dong, Y., Qiu, H., Chen, Z., Shi, Y., Xu, X., Huang, M., et al., 2021. A machine learning-based pulmonary venous obstruction prediction model using clinical data and CT image. Int. J. Comput. Assist. Radiol. Surg. 16 (4), 609–617.

Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., et al., 2020. Predicting conversion to wet age-related macular degeneration using deep learning. Nature Med. 26 (6), 892–899.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. 58, 101537.

Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. 31, 77–87.