Contents lists available at ScienceDirect

Engineering

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

Research
Medical Engineering—Article

Bayesian Inference and Dynamic Neural Feedback Promote the Clinical Application of Intelligent Congenital Heart Disease Diagnosis

Weimin Tan [a,#], Yinyin Cao [b,#], Xiaojing Ma [b,#], Ganghui Ru [a,#], Jichun Li [a], Jing Zhang [b], Yan Gao [b], Jialun Yang [b], Guoying Huang [b,*], Bo Yan [a,*], Jian Li [b,*]

[a] *Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China*
[b] *Cardiovascular Center & Clinical Laboratory Center, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai 201102, China*

ARTICLE INFO

Article history:
Received 26 March 2022
Revised 1 September 2022
Accepted 10 October 2022
Available online 16 January 2023

Keywords:
Congenital heart disease
Artificial intelligence
Deep learning
Model uncertainty

ABSTRACT

Congenital heart disease (CHD) is the leading cause of infant death. An artificial intelligence (AI)-based CHD diagnosis network (CHDNet) is an echocardiogram video-based binary classification model that judges whether echocardiogram videos contain heart defects. Existing CHDNets have shown performances comparable to or even better than medical experts, but their unreliability on cases outside of the training set has become the main bottleneck for their deployment. This is a common problem for most AI-based diagnostic approaches. Here, to overcome this challenge, we present two essential mechanisms—Bayesian inference and dynamic neural feedback—to respectively measure and improve the diagnostic reliability of AI. The former easily makes the neural network output its reliability instead of a single prediction result, while the latter is a computational neural feedback cell that allows the neural network to feed knowledge from the output layer back to the shallow layers and enables the neural network to selectively activate relevant neurons. To evaluate the effectiveness of these two mechanisms, we trained CHDNets on 4151 echocardiogram videos containing three common CHD defects and tested them on an internal test set of 1037 echocardiogram videos and an external set of 692 videos that were newly collected from other cardiovascular imaging devices. Each echocardiogram video corresponds to a unique patient and a unique visit. We demonstrate on various neural network architectures how the reliability obtained by Bayesian inference interprets and quantifies the significant performance difference between internal and external test sets of neural networks, and how the devised feedback cell helps the neural networks to maintain high accuracy and reliability, despite the input being corrupted by noise or when using an external test set.

1. Introduction

The incidence of congenital heart disease (CHD) is about 9/1000 live-born children. Among all subtypes of CHD, ventricular septal defect (VSD), atrial septal defect (ASD), and patent ductus arteriosus (PDA) are the most common [1,2]. ASD refers to the abnormal development of the atrial septum during the embryonic period, resulting in a defect of the septum between the left and right atriums. Depending on the location of the defect, it can be further classified as secondary ASD, primum ASD, sinus venosus ASD, or other rarer forms of ASD, among which secondary ASD is the most common type. VSD, which refers to abnormal traffic between the left and right ventricles of the heart, is the most common congenital heart malformation, accounting for almost 50% of all CHD. According to its anatomical site, VSD can be further divided into perimembrane, subcristal, intracristal, subarterial, and muscle defects. Perimembranous VSD is the most common type, accounting for about 60% of all VSDs. PDA refers to persistent opening of the duct connecting the descending department of the aorta and the pulmonary artery. In the fetal period, the fetal circulation system depends on the duct's existence, but the duct should close naturally after birth. If the duct remains open, the condition is known as PDA. The incidence of isolated PDA in term-born infants is about 1/2000 live births, accounting for 5%–10% of all CHD.

* Corresponding authors.
E-mail addresses: gyhuang@shmu.edu.cn (G. Huang), byan@fudan.edu.cn (B. Yan), lijianjulia@fudan.edu.cn (J. Li).

These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2022.10.015

Early identification and diagnosis of CHD is the key to reducing the natural mortality of the disease. A transthoracic echocardiogram is considered to be the first choice for the detection and diagnosis of CHD, because it is a non-invasive imaging modality [3–5]. Echocardiograms are the main basis for the formulation of surgical or interventional therapy and the evaluation of efficacy. Recent echocardiogram-based diagnosis approaches using deep learning [6–10] have been developed and have achieved unprecedented performance on the data collected, even exceeding an expert level; however, when they are applied to cases for which they have not been trained, the diagnostic results may suffer in accuracy. A feasible way to identify such cases is to measure the reliability of the diagnostic results.

Reliability is the ability of a diagnostic model to identify in-distribution (IND) and out-of-distribution (OOD) test samples as either low uncertainty (high confidence) or high uncertainty (low confidence) on the premise of giving correct diagnostic results. Measuring the reliability of diagnostic results and pursuing high diagnostic accuracy should be equally important goals. In clinical practice, doctors' willingness to use machine learning-based diagnosis approaches depends on their trust that those approaches can accurately and reliably diagnose CHD. This is a valid concern that applies to any diagnosis approach. Therefore, reliability measurement for CHD diagnosis approaches is becoming increasingly important. Recently, research efforts to construct prediction sets with coverage guarantees under various assumptions are emerging [11–13]. Most of these approaches provide theoretical coverage guarantees when the data distribution from which the predictions are built matches the data distribution from which the predictive model was generated. Conformal prediction, one of the best-known of these methods, can guarantee to cover new observations with high probability [12,14]. As a generalization in another direction, the method [11] provides risk-controlling prediction sets, which have low prediction risk with high probability over the randomness in the data.

Unlike conformal prediction, Bayesian probability theory provides math-based tools for reasoning about model uncertainty, but these tools often require prohibitive computational overhead. The recently reported Bayesian neural networks [15], Bayesian approximation [16], and variational inference [17,18] are examples of Bayesian inference approaches that can measure the reliability of a model's output and thus provide an objective assessment of whether the model's output should be adopted. In this study, we use the approximate method of Bayesian inference, which only requires the commonly used dropout method to model uncertainty and has the advantage of being fast and easy to implement. Intuitively, the diagnostic reliability on internal test sets should be higher than that on external test sets, so the reliability is helpful in identifying cases outside of the training data.

Measuring the reliability of diagnostic results is significant but insufficient, as improving both reliability and robustness on external test sets is more meaningful and requires deep learning-based diagnosis approaches. Thanks to long-term natural evolution, the human visual system exhibits incredibly high reliability and robustness, and its anti-interference ability is extremely strong. Neural feedback—a complex basic mechanism in the human visual cortex—can selectively activate relevant neurons and suppress nonrelevant distractive noises or patterns, which is useful in dealing with input images with distractors or cluttered backgrounds [19–21]. Recently, a feedback mechanism has been explored and applied to various vision tasks [22–29]. In cognitive theory, feedback connections linking visual areas of the cortex can transmit response signals from higher order to lower order areas [22,23]. This has recently inspired scholars [24,25] to design advanced deep architectures containing a feedback module. The feedback mechanism in these architectures transmits the information from the deep layers of the network back to the previous layer to guide the extraction of low-level coding information, enabling a top–down approach. The feedback network method [30] is the work most relevant to the present study, as it transmits deep features with semantic information into intermediate representations of input images to enable feedback in deep networks. However, it only passes on deeper information and does not update shallow representations. Therefore, there is an urgent need for a dynamic neural feedback cell that can be widely used in deep neural networks.

In this work, we propose a computational dynamic neural feedback cell that can feed knowledge from the output layer back to the shallow layers, allowing diagnosis models to change their features in shallow layers during feedforward inference. For three common congenital heart defects (VSD, PDA, and ASD), we demonstrate on various representative deep architectures that the feedback cell can significantly improve the architectures' reliability, robustness, and accuracy in distinguishing between normal hearts and common CHD defects, even if the input is severely damaged by noise or comes from an external set. Considering the high transferability of the proposed feedback cell, it is possible for this cell to improve other diagnostic models in terms of reliability, robustness, and accuracy.

2. Methods

2.1. Training and test data acquisition

Deep learning-based diagnosis approaches [9,10,31–34] often require adequate data and accurate annotations for training. The research progress of CHD diagnosis is hindered by the lack of large-scale real-world echocardiograms with well-annotated heart defect types according to the intraoperative final diagnosis. In our work, data for a total of 5880 infants, including 1213 with ASD, 1078 with VSD, 970 with PDA, and 2619 healthy controls (Fig. 1; Tables S1 and S2 in Appendix A), were acquired from 1 January 2015 to 30 June 2021 from a grade-A tertiary children's hospital. Every patient included had an echocardiogram video and still images, which are sufficient for training CHD diagnosis networks (CHDNets). We used a Philips iE 33 as the instrument, and the frequency of the sensor ranged from 3 to 8 MHz, or from 1 to 5 MHz. Two-dimensional (2D) imaging combined with color Doppler flow mapping displayed the location, size, and flow direction of the defect. According to the anatomy of these three CHDs, the atrial septum, ventricular septum, and left pulmonary artery were observed to determine whether they had defects. Two standard 2D views together with color Doppler flow mapping (a dual model) were acquired, with a parasternal short-axis view (PSSAX) of the aorta for patients with VSD and PDA, and a subxiphoid long-axis view (SXLAX) of two atria for patients with ASD. The diagnosis results of all the patients were confirmed by either at least two senior echocardiographists or an intraoperative final diagnosis. All datasets originated from patient studies approved by the Ethics Committee of the Children's Hospital of Fudan University (approval number: 258), and the study was conducted according to the *Declaration of Helsinki*. We received informed consent from the patients' parents or guardians and ensured that patient information would not be disclosed.

2.2. Data labeling and quality control

We downloaded all echocardiograms in Digital Imaging and Communications in Medicine (DICOM) format, and keyframes in each echocardiogram video were manually selected by experienced echocardiographists. The process of data annotation was as

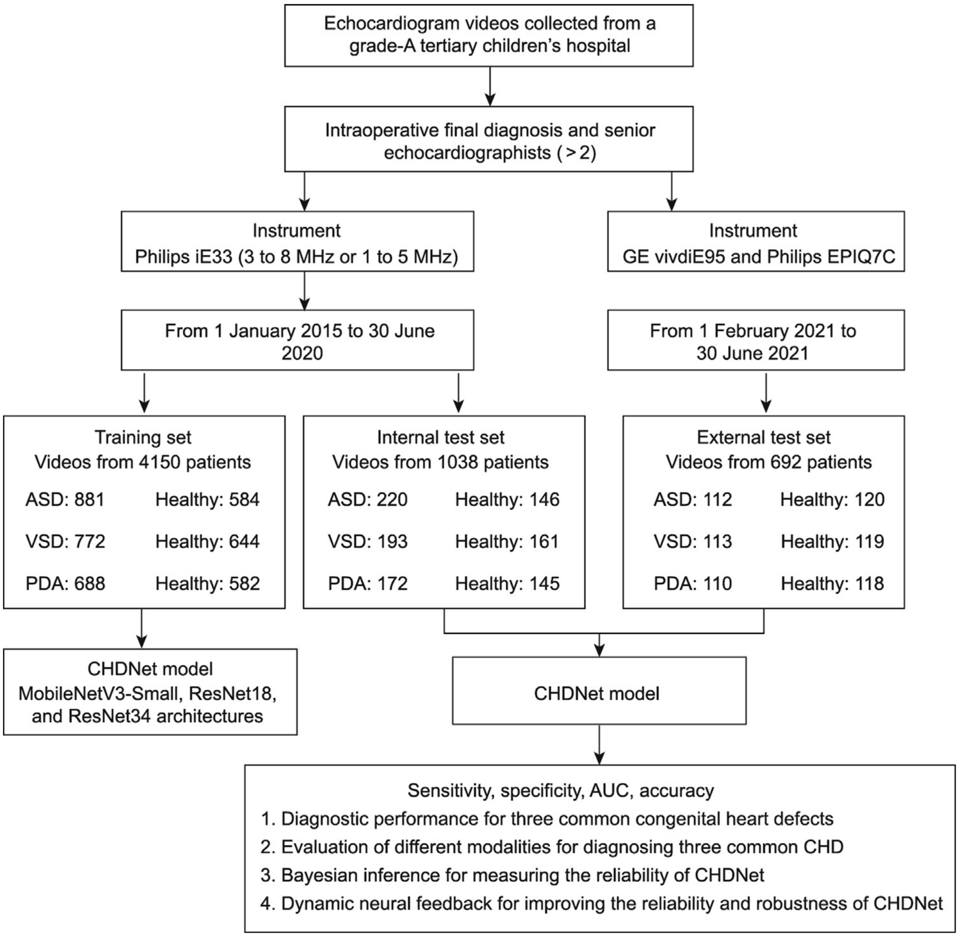
Echocardiogram videos collected from a grade-A tertiary children's hospital
Intraoperative final diagnosis and senior echocardiographists (> 2)
Instrument
Philips iE33 (3 to 8 MHz or 1 to 5 MHz)
Instrument
GE vivdiE95 and Philips EPIQ7C
From 1 January 2015 to 30 June 2020
From 1 February 2021 to 30 June 2021
Training set
Videos from 4150 patients
ASD: 881 Healthy: 584
VSD: 772 Healthy: 644
PDA: 688 Healthy: 582
Internal test set
Videos from 1038 patients
ASD: 220 Healthy: 146
VSD: 193 Healthy: 161
PDA: 172 Healthy: 145
External test set
Videos from 692 patients
ASD: 112 Healthy: 120
VSD: 113 Healthy: 119
PDA: 110 Healthy: 118
CHDNet model
MobileNetV3-Small, ResNet18, and ResNet34 architectures
CHDNet model
Sensitivity, specificity, AUC, accuracy
1. Diagnostic performance for three common congenital heart defects
2. Evaluation of different modalities for diagnosing three common CHD
3. Bayesian inference for measuring the reliability of CHDNet
4. Dynamic neural feedback for improving the reliability and robustness of CHDNet

Fig. 1. Flowchart of data acquisition for the training and evaluation of the CHDNet.

follows. Each echocardiogram was evaluated by means of a three-level assessment system. First-level assessment was conducted by a medical student with a bachelor's degree or higher in quality control training. Second-level assessment was performed by two junior echocardiographists, and third-level assessment was performed by two experienced echocardiographists with more than ten years of clinical experience. This three-level assessment system ensured that each echocardiogram had the correct diagnostic label and heart defect location. After completing the data labeling, 100 objects in the acquired data were randomly selected and checked by a third experienced echocardiographist with more than 20 years of clinical experience to minimize the impact of human error on the computational modeling process. Finally, 881 cases of ASD, 772 cases of VSD, and 688 cases of PDA were randomly selected for model training. The training also included their corresponding healthy control groups, with 584, 644, and 582 individuals, respectively.

2.3. Keyframe-based echocardiogram video diagnosis

The diagnostic procedure followed by the experienced echocardiographists involved selecting keyframes with the clearest view of the heart defect if present from the echocardiogram video, and then making diagnosis decisions according to the selected keyframes (Fig. 2(a)). The diagnosis decisions included whether the patient was healthy, what kind of congenital heart defect the patient had, and where the heart defect was located. This diagnostic procedure inspired us to design a keyframe-based CHD diagnosis model (Fig. 2(b), left). In the diagnosis model, the trained classification model is used first to determine whether the video contains heart defects (video frames can be combined into a batch form for parallel computation), and then potential video frames containing heart defects are selected by comparing the feature distances. Finally, a faster region-based convolutional neural network (faster-RCNN) is used to detect the position of the heart defect on the selected video frames. The diagnostic results are demonstrated in Figs. 2(c)–(e) and analyzed in the Results section.

This diagnostic procedure was a reasonable choice because, from the model aspect, the faster-RCNN is larger than the classification model in terms of network parameters and computational complexity. Besides, from an echocardiogram video containing congenital heart defects, only a small number of video frames contain heart defects and are used for the analysis. Therefore, considering both the model and the data, the method of first identifying and then detecting heart defects is a good choice for CHD diagnosis based on an echocardiogram video, as it makes a tradeoff between diagnostic accuracy and efficiency.

2.4. CHDNet training and evaluation

Based on the above analysis and the acquired echocardiogram data, we chose three representative neural network architectures to implement the CHDNet: ① MobileNetV3-Small [35], a typical classification network on the ImageNet database with a low resource use; ② ResNet18 [36], a classification network on the ImageNet database with well-known deep residual connection; and ③ a deeper version of ResNet18, ResNet34 [36]. After being

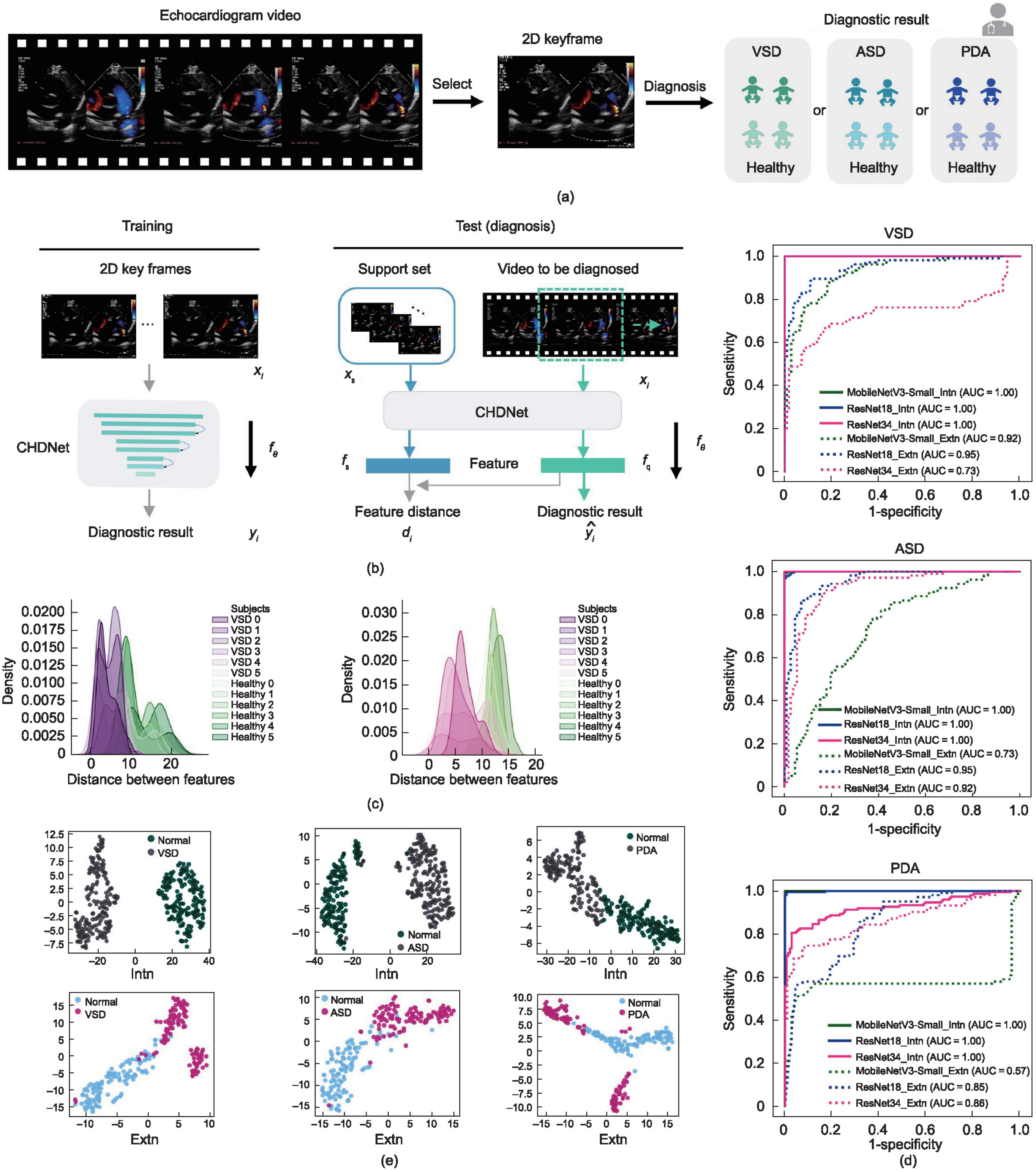

Fig. 2. Diagnosis of three common infantile congenital heart defects based on 2D keyframes. (a) The diagnosis process of experienced echocardiographists; 2D keyframes with a clear heart defect are selected from the echocardiogram video for diagnosis. (b) Overview of the proposed pipeline for CHD diagnosis. Pairs (x_i, y_i) of 2D keyframes and corresponding annotations based on an intraoperative final diagnosis are used to train the CHDNet to predict y_i from x_i. The trained CHDNet $f_\theta(\cdot)$ can then be used to diagnose previously unseen echocardiogram videos x_i, yielding $(d_i, \hat{y}_i)$ for the feature distance and diagnostic result, where x_s denotes the support set containing abnormal cases of heart defects. f_s is the meta-representation of the support set x_s. f_q is the feature representation of input video frames. (c) Statistics of the feature distance between the support set and the echocardiogram video to be diagnosed for VSD and ASD. (d) Receiver operating characteristic (ROC) curves of three representative neural network architectures for diagnosing three common congenital heart defects. (e) Visualization of the classification feature of CHDNet (ResNet18) on the internal and external test sets of VSD. Intn: internal; Extn: external.

trained, the CHDNet models were applied to previously unseen internal and external test sets (Fig. 2(b), right).

For the experiment in diagnosing three common congenital heart defects, MobileNetV3-Small, ResNet18, and ResNet34 were adopted to train on a 2D keyframe of the echocardiogram videos in the training set. The training set was compiled from keyframes (containing representative normal and abnormal frames) selected by experienced echocardiographists from echocardiographic videos. For all other experiments, ResNet18 was adopted to implement the CHDNet. All the experiments were implemented in Python using Pytorch. Source code for training and evaluation is available at https://drive.google.com/file/d/17plnqZVyBGADlYRX-MulZRu4te-zkbDBu/view?usp=sharing. All CHDNet models employed the same training setting—that is, network hyperparameters, training data, and data preprocessing. Each CHDNet model was trained three times by setting a different initial seed for the network parameters. During model training, we divided a small portion of the training set into a validation set. The model with the highest accuracy on the validation set was used as the final model.

2.5. Automatic selection of keyframes based on the trained CHDNet

An echocardiogram video often contains a large number of frames (> 50). It is important to select keyframes that clearly show the heart defects and recommend them to echocardiographists for further diagnosis. To this end, we first obtained the support prototype feature f_s^d, $d \in \{\mathrm{VSD}, \mathrm{PDA}, \mathrm{ASD}\}$ of each heart defect by averaging the classification features of all abnormal cases in the training set (Fig. 1(b), right), which can be formulated as follows:

$$f_s^d = \frac{1}{S}\sum\nolimits_{k=1}^{S} f_\theta(x_s(k)), d \in \{\mathrm{VSD}, \mathrm{PDA}, \mathrm{ASD}\} \tag{1}$$

where $x_s = \{x_s(k)|k = 1, 2, ..., S\}$ denotes the support set containing S abnormal echocardiograms of heart defects. $f_\theta(\cdot)$ denotes the trained CHDNet model with parameters θ. The support prototype feature f_s^d is the meta-representation for each heart defect.

Then, we computed the Euclidean distance between the support prototype feature and the classification feature of each frame of an echocardiogram video to be diagnosed. Finally, those frames with a minimal distance to the support prototype feature were selected as keyframes.

$$L_{\mathrm{keyframe}}(x_i) = \underset{x_i}{\mathrm{argmin}} \left|\left|f_s^d - f_\theta(x_i)\right|\right|_2, i = \{1, 2, \cdots, T\} \tag{2}$$

where T denotes the number of frames in an echocardiogram video. Typically, 3–4 frames were selected as the keyframes.

2.6. Keyframes-based heart defect detection

It is not enough to just show whether the current echocardiogram video contains a heart defect; it is also necessary to tell echocardiographists the specific location and size of the heart defect. Therefore, two experienced echocardiographists were invited to accurately mark the location and size of each heart defect in an echocardiogram using a bounding box (Tables S3 and S4 in Appendix A), which was then used to train the faster-RCNN model [37]—a classic detection network for heart defect detection. Finally, the trained faster-RCNN model was applied to the selected keyframes instead of to all frames in an echocardiogram video for diagnosis, thereby significantly reducing the diagnosis time.

The processing flow of the detection network was as follows. ResNet50 [36] was selected as the feature extractor. Afterwards, based on the learned representations from layer 1 to layer 4 of ResNet50, we used feature pyramid networks [38] to compute multiscale feature representations in order to deal with the problem of heart defects with different sizes in the echocardiograms. Then, a region proposal network based on the multiscale representations was used to produce regions of interests (RoI) to determine whether each predefined anchor was a heart defect or not, as well as to optimize the sizes and positions of these anchors. In our study, we set predefined anchors to have three aspect ratios, {1:2, 1:1, 2:1}, and five scales, $\{16^2, 32^2, 64^2, 128^2, 256^2\}$. This setting fully considered the actual size and contours of the heart defects in the echocardiograms. In the second stage, we used the RoI pooling operation to extract features based on the proposed regions. A class predictor was used to predict the region type and to further optimize the location and size of each bounding box. Finally, we used the non-maximum suppression (NMS) [39] algorithm to remove redundant prediction bounding boxes with a confidence lower than 0.3.

2.7. Reliability measurement

Conventional classification networks are trained using a cross entropy function, which can be described as follows. Pairs $\left\{\left(x_i^d, y_i^d\right)\right\}_{i=1}^N, d \in \{\mathrm{VSD}, \mathrm{PDA}, \mathrm{ASD}\}$ of input images and output labels are used to train the CHDNet by minimizing the binary cross entropy function $L_{\mathrm{ce}}(\theta)$.

$$L_{\mathrm{ce}}(\theta) = -\frac{1}{N}\sum\nolimits_{i=1}^{N} y_i \log\left(p(x_i)\right) + (1 - y_i)\log\left(1 - p(x_i)\right) \tag{3}$$

where N is the total number of training samples. $p(x_i) = \mathrm{Softmax}\left(\hat{y}_i = 1|x_i\right) = \frac{\exp\left(f_\theta^1(x_i)\right)}{\sum_{j=0}^{1}\exp\left(f_\theta^j(x_i)\right)}$ is the predicted probability of the sample x_i being abnormal, $\hat{y}_i$ denotes the diagnostic result.

Bayesian neural networks are suggested to provide a probabilistic interpretation for deep networks, but their feedforward inference suffers from expensive computation [40–42]. To overcome this issue, Gal and Ghahramani [16] proposed the use of a dropout technique to approximate Bayesian inference. Their work demonstrates that the posterior distribution of a model prediction can be obtained by means of Monte Carlo sampling using the dropout method. We follow their method here and allow the CHDNet to have the ability to output the uncertainty of the result—that is, the ability to approximate Bayesian inference. Instead of $L_{\mathrm{ce}}(\theta)$, we choose $L_{\mathrm{uce}}(\theta, \sigma)$ as the loss function in order to optimize the model parameters θ and an additional variance σ.

$$\begin{aligned} L_{\mathrm{uce}}(\theta, \sigma) &= \frac{1}{N}\sum\nolimits_{i=1}^{N} -\log p(y_i = 1|f_\theta(x_i), \sigma) \\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N} -\log \mathrm{Softmax}(y_i = 1|f_\theta(x_i), \sigma) \\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{\sigma^2} L_{\mathrm{ce}}(\theta) + \log \frac{\sum_{j=0}^{1}\exp\left(\frac{1}{\sigma_i{}^2} f_\theta^j(x_i)\right)}{\left(\sum_{j=0}^{1}\exp\left(f_\theta^j(x_i)\right)\right)^{\frac{1}{\sigma^2}}} \\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{\sigma^2} L_{\mathrm{ce}}(\theta) + \log\sigma \end{aligned} \tag{4}$$

where an explicit simplifying assumption $\left(\sum_{j=0}^{1}\exp\left(f_\theta^j(x_i)\right)\right)^{\frac{1}{\sigma^2}} \approx \frac{1}{\sigma}\sum_{j=0}^{1}\exp\left(\frac{1}{\sigma_i^2} f_\theta^j(x_i)\right)$ is utilized in the final transition [16].

In network optimization, directly optimizing σ is difficult due to numerical instability. Instead, we decided to optimize $\log\sigma$. Then, Eq. (4) is formulated as follows:

$$L_{\mathrm{uce}}(\theta, z) = e^z L_e(\theta) + z \tag{5}$$

with

$$z = \log \sigma \tag{6}$$

where the variance σ is used to measure the aleatoric uncertainty.

Epistemic uncertainty can be measured by computing the variance of the results of multiple models. Thanks to the advantages of the dropout technique, it is necessary to independently train M diagnosis models on the same dataset by setting different initial seeds for model initialization. Therefore, we can sample multiple different networks by keeping the dropout open during the test time.

$$\widehat{y}_i = E[P_j(y = 1|f_\theta(x_i, \sigma_i)], j = 1, 2, \cdots, M \tag{7}$$

where E denotes the mathematical expectation. $\widehat{y}_i$ denotes the expected model output. P_j is prediction probability of the jth model.

$$u_i = \frac{1}{M}\sum_{j=1}^{M}\left(P_j(y = c|f_\theta(x_i), \sigma_i) - \widehat{y}_i\right)^2 \tag{8}$$

where u_i denotes the epistemic uncertainty of the model output. The epistemic uncertainty u_i is an effective and objective metric for assessing the blind spot of the trained CHDNet. Its characteristics are shown in Fig. 3.

2.8. Dynamic feedback cell for improving model robustness

The main idea of the dynamic feedback cell is to obtain an attention map by using the feature-category transformation matrix $W_{\text{clf}} \in R^{C_{\text{out}} \times C}$ in the classification layer, and to then use the attention map to update the intermediate layer feature F_{in}^{i-1} in the classification model in order to obtain more robust classification features. The whole feedback process is shown in Algorithm 1.

Algorithm 1 Dynamic feedback cell

```
Input:   Input $F_{\text{in}}^{i-1} \in R^{H \times W \times C_{\text{in}}}$ and output $F_{\text{out}}^{i-1} \in R^{H \times W \times C_{\text{out}}}$
         features of a middle module (contains several
         convolutional layers) in a classification model,
         matrix $W_{\text{clf}} \in R^{C_{\text{out}} \times C}$ in classification layer, max
         iterations $T$. $C_{\text{in}}$ and $C_{\text{out}}$ denote feature channels. $C$
         denotes the number of categories.
Output:  Feedback feature $F_{\text{fb}}^{i-1}$
Target:  Updating the input feature $F_{\text{in}}^{i-1}$
for $i$ = 1 to $T$ do
    $F_{\text{fb}}^{\text{cls}} = \text{Softmax}\left(F_{\text{out}}^{i-1} \cdot W_{\text{clf}}\right)$
    $F_{\text{fb}}^{\text{cls}} \leftarrow F_{\text{fb}}^{\text{cls}}[\cdots, k]$% slice, where $k$ indexes the
    abnormal channel and use it as the attention map
    $F_{\text{fb}}^{\text{fuse}} = \left(F_{\text{fb}}^{\text{cls}} \otimes F_{\text{in}}^{i-1}\right) \oplus F_{\text{in}}^{i-1}$
    $F_{\text{fb}}^{i-1} = \text{Conv}_{1\times1}\left(F_{\text{fb}}^{\text{fuse}}\right)$
    $F_{\text{in}}^{i-1} \leftarrow F_{\text{fb}}^{i-1}$
end
```

The architecture of the dynamic neural feedback cell is illustrated in Fig. 4(a). For a shallow layer in the deep neural network, the input and output are denoted as F_{in}^{i-1} and F_{out}^{i-1}, respectively. The feedback cell begins with a classification layer:

$$F_{\text{fb}}^{\text{cls}} = \text{Softmax}\left(F_{\text{out}}^{i-1} \cdot W_{\text{clf}}\right) \tag{9}$$

where $F_{\text{fb}}^{\text{cls}}$ is the output of the classification layer and indicates the contribution of the feature F_{out}^{i-1} of the shallow layer to the classification results.

Then, $F_{\text{fb}}^{\text{cls}}$ and F_{in}^{i-1} are integrated by two sequential operations: point-wise multiplication ($\otimes$) and channel-wise concatenation ($\oplus$).

$$F_{\text{fb}}^{\text{fuse}} = \left(F_{\text{fb}}^{\text{cls}} \otimes F_{\text{in}}^{i-1}\right) \oplus F_{\text{in}}^{i-1} \tag{10}$$

where $F_{\text{fb}}^{\text{fuse}}$ is the fusion result and has twice the number of feature channels as F_{in}^{i-1}. To let $F_{\text{fb}}^{\text{fuse}}$ pass into the original network without any modification to the network, we use a convolutional layer with a kernel size of 1 × 1 ($\text{Conv}_{1\times1}$) to compress the channel of $F_{\text{fb}}^{\text{fuse}}$, resulting in the same number of channels as F_{in}^{i-1}.

$$F_{\text{fb}}^{i-1} = \text{Conv}_{1\times1}\left(F_{\text{fb}}^{\text{fuse}}\right) \tag{11}$$

where F_{fb}^{i-1} is the output of the feedback cell and can be regarded as the clean version of F_{in}^{i-1}. Irrelevant background noise and patterns in F_{in}^{i-1} are suppressed and do not exist in F_{fb}^{i-1}.

From Eqs. (9)–(11), we find that the key idea of the feedback cell is to make full use of the transformation matrix in the classification layer, because the matrix can project the high-dimensional feature vector extracted by the network on the input into the category space. This projection process is a dimensionality reduction process, which retains the feature information that is related to the classification and ignores feature information that is not related to the classification. Therefore, we propose the application of this transformation matrix to update the features of the middle layer of the network, thereby helping the network to better filter out features irrelevant to the classification—that is, to obtain a "clean feature."

2.9. Evaluation metrics

In this work, commonly used classification evaluation metrics were employed, including specificity, sensitively, accuracy, F1-score, and area under the curve (AUC). To evaluate the performance of the detection model, we counted the number of true positives (D_{tp}), false positives (D_{fp}), and false negatives (D_{fn}) in each type of heart defect. It should be noted that we determined whether a detection was accurate or not according to the degree of coincidence between the prediction box and the bounding box annotated by the doctors. Based on these statistics, we calculated two evaluation metrics—namely, recall and precision—using the following equations:

$$D_{\text{recall}} = \frac{D_{\text{tp}}}{D_{\text{tp}} + D_{\text{fn}}} \tag{12}$$

$$D_{\text{precision}} = \frac{D_{\text{tp}}}{D_{\text{tp}} + D_{\text{fp}}} \tag{13}$$

where D_{recall} and $D_{\text{precision}}$ are used to evaluate the detection performance.

2.10. Noise robustness evaluation

To test the robustness of the CHDNet model, we added Gaussian noise with a variance σ_{noise} from 0.01 to 0.05 to the input echocardiogram. The noise input is defined as follows:

$$x_{\text{noise}} = x_i + N\left(0, \sigma_{\text{noise}}^2\right) \tag{14}$$

where x_i denotes the input echocardiogram and $N\left(0, \sigma_{\text{noise}}^2\right)$ denotes the Gaussian distribution with 0 as the mean and σ_{noise} as the variance.

2.11. Data availability

Since the entire data set is very large, we only uploaded part of the data to Google Drive. Despite that, this part of the data can still

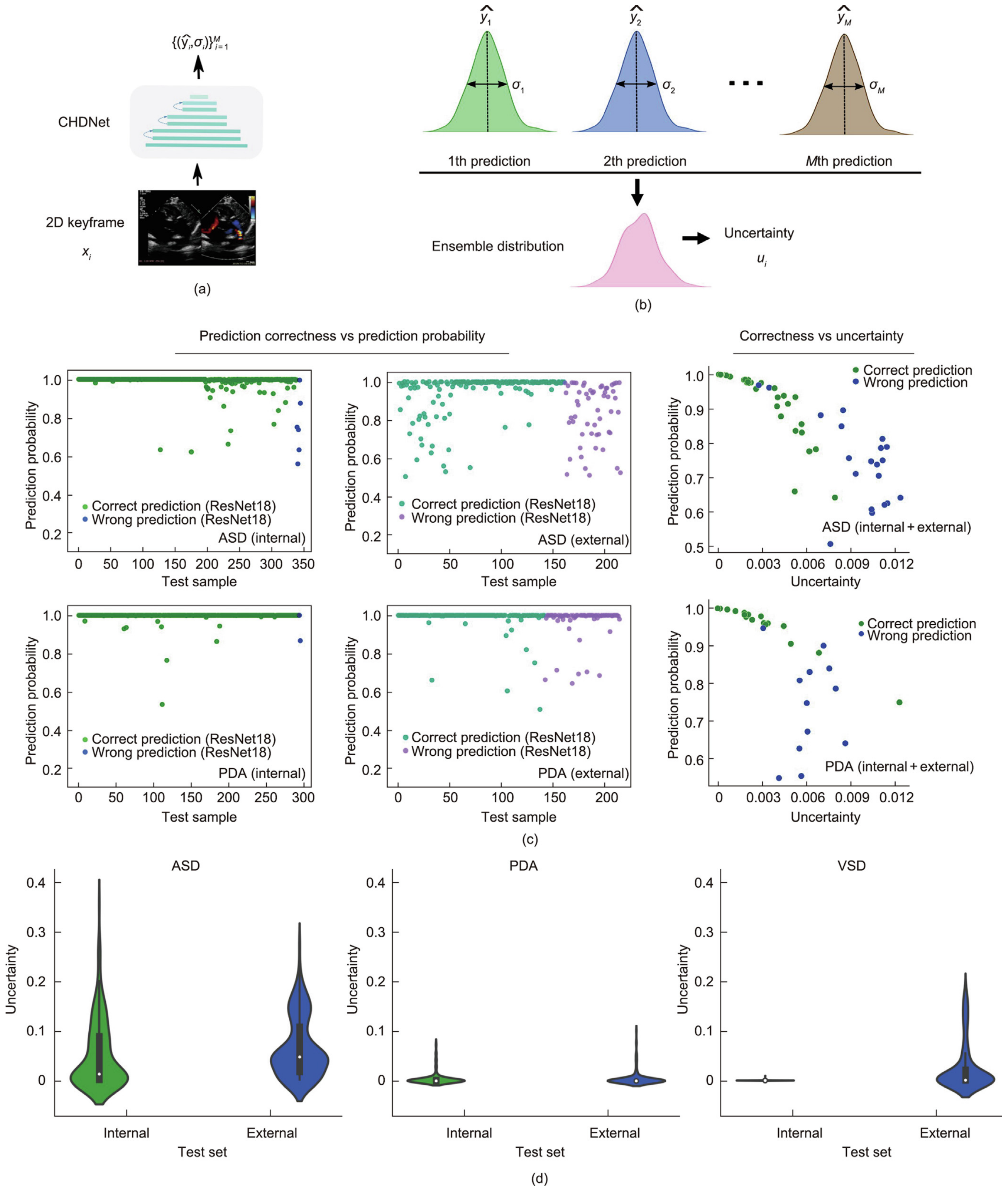

Fig. 3. Reliability of the CHDNet with Bayesian inference. (a) The output of the CHDNet (ResNet18) is modified to a vector $(\widehat{y}_i, \sigma_i)$, $i = \{1, 2, \cdots, M\}$, where σ is the variance of each output. (b) M distributions are ensembled to obtain the epistemic uncertainty u_i for input x_i. (c) Prediction correctness versus (vs) prediction probability (left), and prediction correctness vs uncertainty (right). (d) From left to right, statistics of uncertainty on the internal and external test sets of ASD, PDA, and VSD.

help us to understand how CHDNet model was developed and validated. The data is available at https://drive.google.com/file/d/16Z-uZw6JuIqKHUTvYDt2z_0xgq8Ua7e1J/view?usp=sharing.

2.12. Code availability

The Pytorch code of CHDNet (ResNet18) for diagnosing three common congenital heart defects and faster-RCNN for detecting the position and size of heart defects are freely available at https://drive.google.com/file/d/17plnqZVyBGADlYRXMulZRu4te-z-kbDBu/view?usp=sharing.

3. Results

3.1. Evaluation of representative neural network architectures for diagnosing three common congenital heart defects

Fig. 2 shows the diagnostic performance of three CHDNet models on the internal and external test sets for three common congenital heart defects (VSD, ASD, and PDA). The receiver operating characteristic (ROC) curves show that these three models consistently achieve high diagnostic performance on the internal test set, even reaching 100% AUC on VSD and ASD (Fig. 2(d)). On the external test set, the models' performance drops significantly; for example, the AUC of CHDNet (ResNet18) drops from 100% on the internal test set of PDA to 57.8% on the external test set (Tables S5–S7 in Appendix A). These results suggest that the data from different acquisition devices will partially cause the CHDNet model to fail, which is a common problem of most machine learning-based diagnosis approaches. Fig. 2(e) shows the visualization results of the learned classification features of heart defects when using the *t*-SNE algorithm to reduce the dimensionality. It is clear that the cases from the internal test set are better separated, while cases from the external test set are worse separated. The results of the CHDNets based on the two other architectures (MobileNetV3-Small and ResNet34) are shown in Fig. S1 in Appendix A. This observation is consistent with the diagnosis result mentioned above.

We next ask whether a CHDNet trained on a 2D keyframe can be applied to echocardiogram video-based CHD diagnosis. To achieve this, all abnormal cases of a certain congenital heart defect in the training set are used as the support set. Therefore, ASD, VSD, and PDA each correspond to a support set. The average classification feature of a support set is regarded as a support prototype feature that is a meta-representation of a CHD abnormal echocardiogram. The feature distance between the support prototype feature and the frame features of the echocardiogram video to be diagnosed can help us select the frames most likely to be abnormal. Fig. 2(c) shows statistics of the feature distance of 12 cases (six healthy and six abnormal) randomly selected from the external test sets of VSD and ASD. The support prototype features of VSD and ASD are closer to the features of patients and farther from those of healthy individuals. With selected abnormal keyframes, we can further use a detection model to detect the spatial position and size of the heart defect (Fig. S2(a) in Appendix A). The detection precision and recall on the internal set of ASD are respectively 0.955 and 0.907, and those on the external set are respectively 0.962 and 0.752 (Fig. S2(b)). The prediction box is highly coincident with the doctor's annotation (Figs. S2(c) and S3 in Appendix A). Similar results are observed in the other types of heart defects (VSD and PDA).

3.2. Bayesian inference for measuring the reliability of CHD diagnosis

We have shown the diagnostic performance of CHDNet models on the internal and external sets of three common congenital heart defects and the models' relationship with different echocardiogram modalities. Nevertheless, as for any diagnosis approach, the problem of measuring the reliability of diagnosis results must be solved, because echocardiographists are often unwilling to use a diagnosis model with unknown reliability. To this end, following Monte Carlo sampling [16] we embedded Bayesian inference into the CHDNet models and obtained the uncertainty of the diagnosis (Figs. 3(a) and (b)). We first drew the relationship between the prediction accuracy and the prediction probability (after Softmax) of the CHD cases (Fig. 3(c), left). For wrong prediction cases, the model still gives its prediction results with a relatively high probability—that is, with high confidence in its own predictions. Therefore, the prediction probability cannot well measure the reliability of diagnosis.

We next asked whether there was a close relationship between the variance σ of the model output and the prediction correctness of the test samples (Fig. S4 in Appendix A). Intuitively, a small variance indicates high confidence, while a large variance signifies low confidence in the CHD diagnosis. We observed that the variance σ could not help us to intuitively understand which prediction of a case was reliable or unreliable. Therefore, we visualized the uncertainty of the cases (Fig. 3(c), right). The uncertainty of the cases with incorrect predictions is clearly higher than that of the cases with correct predictions. This finding suggests that uncertainty can be a candidate metric to reflect the confidence of the prediction results. Fig. 3(d) shows the uncertainty of the internal and external test sets of VSD, ASD, and PDA. The uncertainty of the external test set is higher than that of the internal test set, which indirectly suggests that the diagnosis of cases in the external test set is more difficult than that of cases in the internal test set.

3.3. Dynamic neural feedback for improving the reliability and robustness of CHD diagnosis

Bayesian inference can obtain the uncertainty of the CHDNet models as a measure of the reliability, but this is insufficient. Improving both the reliability and the robustness of the CHDNet models is essential, which is true for any diagnosis approach based on machine learning. CHDNet models are often trained on a specific imaging device with specific parameter settings. When a trained CHDNet is applied to unseen cases acquired from different devices or different parameter settings, its diagnostic performance may be greatly affected, showing low reliability and robustness. Nevertheless, our human visual system exhibits incredibly high reliability and robustness, and its anti-interference ability is extremely strong. Inspired by the neural feedback mechanism in the human visual cortex, we propose a computational dynamic neural feedback cell that feeds the deep knowledge of the classification layer (which can distinguish different categories) back to the shallow layers, so as to eliminate the noise or patterns that are irrelevant to the diagnosis task (Fig. 4(a)). The proposed feedback cell can be applied to existing deep neural networks and to different layers of a deep network.

To test the ability of the proposed feedback cell to selectively activate relevant neurons, different degrees of Gaussian noise (σ_{noise} = 0.01–0.05) were added to the input images during the test. With an increase in the noise, the diagnostic accuracy on the internal and external test sets gradually decreased except on the external set of PDA without feedback. (Fig. 4(b)), indicating that the carefully trained CHDNet still showed low robustness. This observation can be explained using the network features (Fig. 5(a)). The feedback cell helps the CHDNet model to suppress background noise that is irrelevant to the goal of CHD diagnosis, resulting in clean representations. After the CHDNet model was embedded with the feedback cell, it showed better robustness and higher

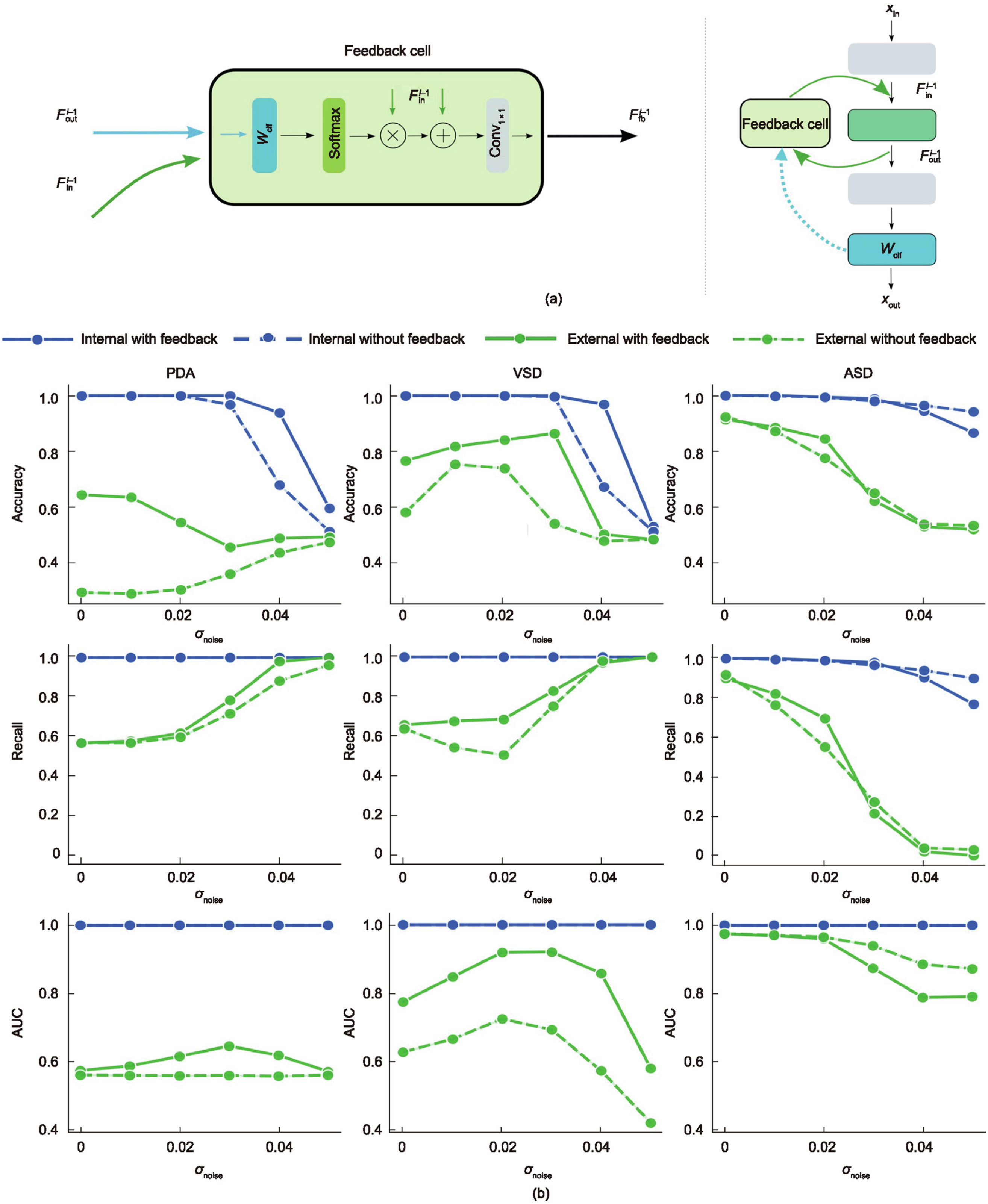

Fig. 4. Dynamic neural feedback mechanism. (a) Schematic of the feedback cell, which updates the intermediate layer feature F_{in}^{i-1} in the classification model to obtain more robust classification features by using the feature-category transformation matrix W_{clf} in the classification layer. (b) From top to bottom are the accuracy, recall, and AUC of the CHDNet with or without the feedback cell under the interference of different noise levels (σ_{noise} = 0.01–0.05). From left to right are the diagnostic performance of the internal and external test sets of PDA, VSD, and ASD.

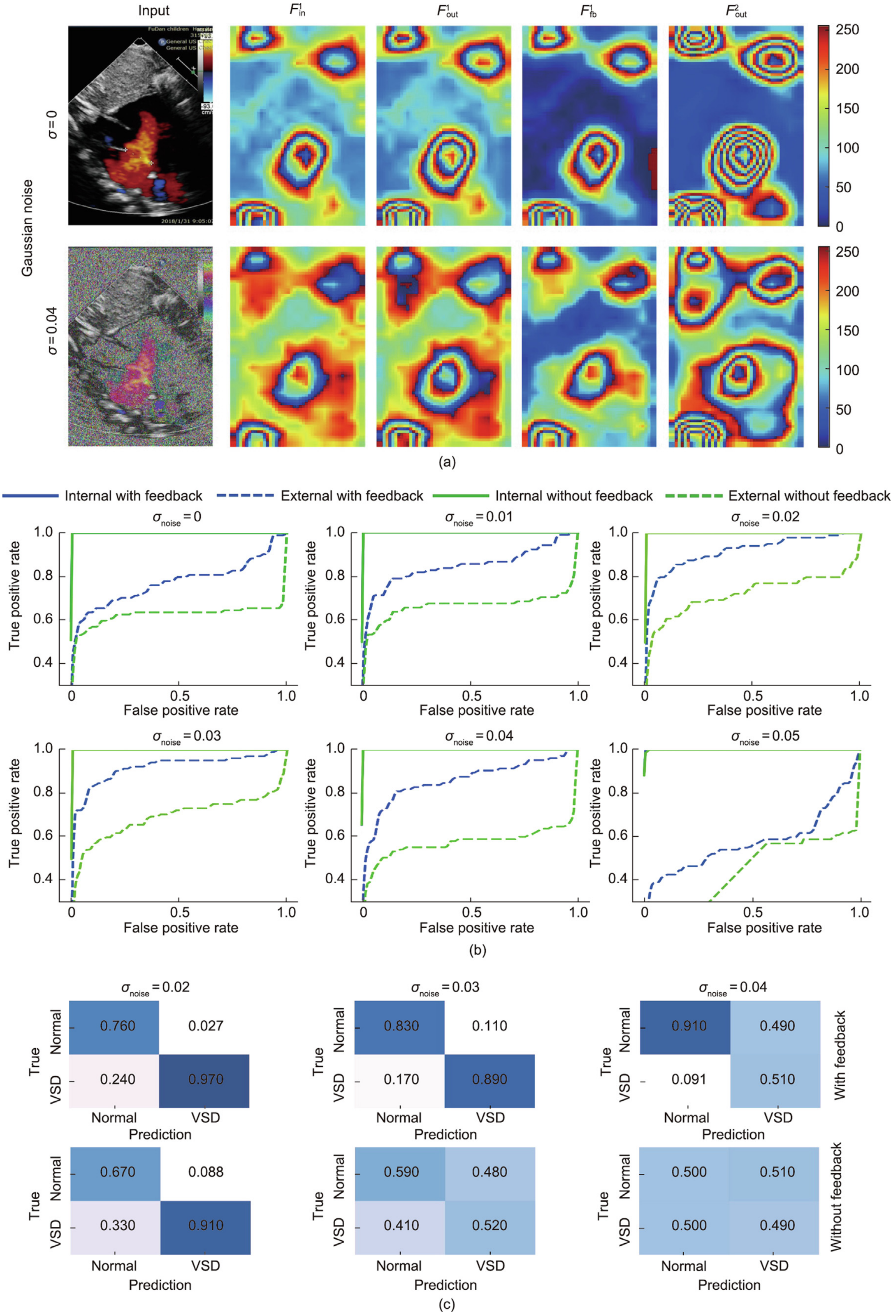

Fig. 5. Characterization of the feedback cell. (a) Comparison of network features before and after the feedback cell is included in the model. Spatial regions that are not relevant for the diagnosis are suppressed by the feedback cell. (b) ROC curve with σ_{noise} = 0.01–0.05. Comparison of the CHDNet (ResNet18) model with and without the feedback cell on the internal and external test sets of VSD. (c) Confusion matrix of the CHDNet model with and without the feedback cell under Gaussian noise σ_{noise} = 0.02–0.04.

Table 1
Comparison of classification accuracy between our method and LassoNet [43].

Method	PDA		VSD		ASD	
	Internal test set	External test set	Internal test set	External test set	Internal test set	External test set
ResNet18	1.000	0.321	1.000	0.583	0.997	0.921
ResNet18 + feedback cell	**1.000**	**0.638**	**1.000**	**0.782**	**1.000**	**0.945**
LassoNet (decoder networks) [43]	0.972	0.548	0.981	0.674	0.974	0.933
LassoNet (tree-based classifiers) [43]	0.978	0.565	0.983	0.685	0.976	0.935

diagnostic performance in terms of accuracy, recall, and AUC on all congenital heart defects (Figs. 4(b), 5(b), and 5(c)).

3.4. Comparison with another method

We compared our feedback cell with a recent method—LassoNet [43]—that can improve the model prediction accuracy on an external dataset or after adding noise. Lemhadri et al. [43] proposed the LassoNet method by modifying the loss function with constraints to make neural networks with feature sparsity. This method can help the network to select a subset of the features, resulting in a sparse network. We used the published office code and slightly modified it to adapt it to our task. A comparison of our method with LassoNet is shown in Table 1 [43]. The LassoNet using decoder networks and tree-based classifiers shows slightly lower performance on the internal dataset but better performance than the Resnet18 model without the feedback cell on the external dataset. Overall, we find that ResNet18 with the feedback cell achieves better performance on the internal and external test sets.

3.5. Comparison with human experts

An echocardiogram is the most important and cost-effective test for cardiovascular disease screening and evaluation. At present, the echocardiograms in most hospitals in China are checked and reported by professional echocardiographists, but the level of echocardiographists in different hospitals varies greatly. In addition, cardiologists, neonatal physicians, and intensive care unit physicians are often required to perform an initial echocardiographic evaluation when an echocardiographists is not available in time. Especially in critically ill infants and children, the presence and type of CHD can influence the choice of further drug and fluid therapy. Therefore, rapid and accurate echocardiogram interpretation is very important for these doctors. In clinical work, we have often found that some doctors can perform a cardiogram but cannot give a good diagnosis.

We invited 23 doctors, including eight echocardiographists and 15 clinicians (cardiovascular specialist doctors from different hospitals, neonatologists, and pediatric intensive care doctors), to evaluate echocardiographic images from internal (n = 50) and external (n = 115) test sets, respectively. The doctors gave only a diagnostic result, while the CHDNet performed positioning and preliminary measurements along with a diagnosis. Finally, the accuracy of each evaluation was compared and analyzed. Table S8 in Appendix A shows the comparison result. In the interpretation of the internal test sets, our model has the absolute advantage. In the interpretation of the external test set, our model does not prevail in accuracy compared with experienced echocardiographists and cardiologists at the Children's Hospital of Fudan University. However, the accuracy of our model—especially for the ASD and VSD prediction models—is significantly higher than that of sonographers and cardiologists with relatively poor medical skills and physicians in our hospital's neonatology department and intensive care unit.

The CHDNet model with the feedback cell has been used by cardiologists and clinicians for more than one year, helping sonographers with relatively backward medical standards to improve their work efficiency and professional capabilities, and helping medical staff outside the field of ultrasound imaging to better carry out clinical work.

4. Discussion

For three common CHDs, we have demonstrated the diagnostic performance of CHDNet implemented with three representative neural network architectures (MobileNetV3-Small [35], ResNet18 [36], and ResNet34 [36]) based on acquired echocardiogram data. We have observed a significant difference in diagnostic performance on the internal and external test sets. Moreover, we have conducted an ablation study to discuss the impact of different echocardiogram modalities on CHDNet performance, showing that the color modality is better than the gray modality and that combining both modalities is the best choice. All these results suggest that existing advanced deep models have enough regression power to implement CHD diagnosis. Nevertheless, the most urgent task in developing a CHD diagnosis model is to resolve clinical relevance, rather than just pursuing a high diagnostic performance on internal tests. What sets our work apart is that we make the CHDNet output its diagnostic reliability by embedding Bayesian inference, which allows echocardiographists to identify diagnostic cases in which results might not be correct. We have shown that the variance of distribution of CHDNet output is not accurate in measuring the reliability of the diagnosis, whereas the uncertainty of ensemble distribution of the CHDNet output can well measure it.

The connections of feedback are more contained than those of feedforward in the human visual cortex [44,45]. This fact is contrary to the current popular deep neural networks that basically only contain feedforward inference during evaluation and whose connection parameters remain unchanged even when the inputs are varied. The powerful ability of feedback inference is little studied in existing works on machine learning. In this work, we propose a computational dynamic neural feedback cell to improve the reliability, robustness, and accuracy of CHD diagnosis. This feedback cell can be easily embedded into existing deep neural networks and improves them significantly without adding major computational complexity. When the input echocardiogram is seriously disturbed by noise, the CHDNet with a feedback cell significantly outperforms the CHDNet without a feedback cell, as shown by the evaluation results for three common congenital heart defects.

To further understand the ability of the feedback cell, we visualized the features of a neural layer in the CHDNet model before and after the feedback cell. The visualized feature map shows that the feedback cell grants the CHDNet model the ability to selectively activate relevant neurons and suppress nonrelevant distractive noises or patterns for the task of CHD diagnosis. It should be noted that the feedback cell is significantly different from a recurrent neural network (RNN), gated recurrent unit (GRU), or long short-term memory (LSTM). These recurrent cells use the information of the previous moment, whereas the feedback cell uses the transformation matrix of the classification layer to update the intermediate layer features. The purpose of RNN, GRU, and LSTM is to exploit the relevance of words in the sequence, whereas the purpose of the feedback cell is to remove background noise that is irrelevant to the classification goal.

However, CHDNet models with Bayesian inference and dynamic neural feedback cannot be directly applied to clinical diagnosis. A great deal of work is still necessary, including data fairness, ethical compliance, clinical trials, and so forth. Bayesian inference enables the CHDNet to output the reliability of the diagnosis, which has been evaluated on the acquired large-scale real-world internal and external tests of three common congenital heart defects. Nevertheless, before clinical practice, there is still a need for a comprehensive and wide evaluation on external test sets, including different races, different ages, and diverse routine ultrasound imaging equipment.

The effectiveness of the proposed dynamic neural feedback mechanism has been evaluated on both a non-enhanced and an enhanced CHDNet model through anti-noise experiments. This evaluation is not enough for clinical practice, because a clinical echocardiogram includes not only noise but also blur, jitter, intra- and inter-patient variations, and so forth. Therefore, further evaluation with clinical trials is necessary.

Taken together, our results show that CHDNet can—in combination with Bayesian inference and dynamic neural feedback—achieve better accuracy, higher reliability, and stronger robustness in diagnosing three common congenital heart defects. The two technologies introduced in this work can be easily embedded in existing deep neural network-based diagnosis approaches, improving their performance and reliability. We predict that the current explosion of echocardiogram data richness and the ability of CHDNet to dynamically adapt to various CHD cases with neural feedback will release the great clinical potential of CHDNet and make such learning approaches prevalent in the clinic.

Acknowledgments

We gratefully acknowledge support for this work provided by the Children's Hospital of Fudan University: Big Data and Artificial Intelligence Project (2020DSJ07), the National Natural Science Foundation of China (U2001209; 61902076; and 81670281), scientific research project of Shanghai Municipal Health Commission (20204Y0100), and Natural Science Foundation of Shanghai (21ZR1406600).

Authors' contribution

Bo Yan, Jian Li, Guoying Huang, Weimin Tan, Yinyin Cao, Xiaojing Ma, and Ganghui Ru conceived of the technique and the whole study; Weimin Tan, Yinyin Cao, Xiaojing Ma, and Ganghui Ru implemented the algorithm; Yinyin Cao, Jing Zhang, Yan Gao, Jialun Yang, Xiaojing Ma, and Jian Li collected and labeled data; Weimin Tan, Jichun Li, and Ganghui Ru conceived of deep learning-based CHD diagnosis; Bo Yan, Weimin Tan, and Ganghui Ru designed the validation experiments; Weimin Tan and Ganghui Ru trained the network and performed the validation experiments; Bo Yan, Jian Li, and Guoying Huang supervised the project; and all authors had access to the study and wrote the paper.

Compliance with ethics guidelines

Weimin Tan, Yinyin Cao, Xiaojing Ma, Ganghui Ru, Jichun Li, Jing Zhang, Yan Gao, Jialun Yang, Guoying Huang, Bo Yan, and Jian Li declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.10.015.

References

[1] Erikssen G, Liestøl K, Seem E, Birkeland S, Saatvedt KJ, Hoel TN, et al. Achievements in congenital heart defect surgery: a prospective, 40-year study of 7038 patients. Circulation 2015;131(4):337–46. Discussion 346.

[2] Luo H, Qin G, Wang L, Ye Z, Pan Y, Huang L, et al. Outcomes of infant cardiac surgery for congenital heart disease concomitant with persistent pneumonia: a retrospective cohort study. J Cardiothorac Vasc Anesth 2019;33(2):428–32.

[3] Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, et al. Deep learning in medical ultrasound analysis: a review. Engineering 2019;5(2):261–75.

[4] Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. Engineering 2020;6(3):291–301.

[5] Sedghi S, Huang B. Real-time sssessment and diagnosis of process operating performance. Engineering 2017;3(2):214–9.

[6] O'Neill S. Handheld ultrasound advances diagnosis. Engineering 2021;7(11):1505–7.

[7] Cui Z, Yang B, Li RK. Application of biomaterials in cardiac repair and regeneration. Engineering 2016;2(1):141–8.

[8] Li C, Pisignano D, Zhao Y, Xue J. Advances in medical applications of additive manufacturing. Engineering 2020;6(11):1222–31.

[9] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580(7802):252–6.

[10] Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med 2021;27(5):882–91.

[11] Bates S, Angelopoulos A, Lei L, Malik J, Jordan M. Distribution-free, risk-controlling prediction sets. J Assoc Comput Mach 2021;68(6):1–34.

[12] Sadinle M, Lei J, Wasserman L. Least ambiguous set-valued classifiers with bounded error levels. J Am Stat Assoc 2019;114(525):223–34.

[13] Affenit RN, Barns ER, Furst JD, Rasin A, Raicu DS. Building confidence and credibility into CAD with belief decision trees. In: Proceedings of the Medical Imaging 2017: Computer-Aided Diagnosis; 2017 Mar 3; Orlando, FL, USA.

[14] Scheffe H, Tukey JW. Non-parametric estimation. I. validation of order statistics. Ann Math Stat 1945;16(2):187–92.

[15] McClure P, Kriegeskorte N. Robustly representing uncertainty in deep neural networks through sampling. In: Proceedings of the Second Workshop on Bayesian Deep Learning (NIPS 2017); 2017 Dec 4-9; Long Beach, CA, USA.

[16] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 16–24; New York City, NY, USA. JMLR: W&CP; 2016. p. 1050–9.

[17] Vishnevskiy V, Walheim J, Kozerke S. Deep variational network for rapid 4D flow MRI reconstruction. Nat Mach Intell 2020;2(4):228–35.

[18] Piray P, Daw ND. A model for learning based on the joint estimation of stochasticity and volatility. Nat Commun 2021;12(1):6587.

[19] Lazar A, Lewis C, Fries P, Singer W, Nikolic D. Visual exposure enhances stimulus encoding and persistence in primary cortex. Proc Natl Acad Sci USA 2021;118(43):e2105276118.

[20] Chariker L, Shapley R, Hawken M, Young LS. A theory of direction selectivity for macaque primary visual cortex. Proc Natl Acad Sci USA 2021;118(32):e2105062118.

[21] Ferro D, van Kempen J, Boyd M, Panzeri S, Thiele A. Directed information exchange between cortical layers in macaque V1 and V4 and its modulation by selective attention. Proc Natl Acad Sci USA 2021;118(12):e2022097118.

[22] Gilbert CD, Sigman M. Brain states: top-down influences in sensory processing. Neuron 2007;54(5):677–96.

[23] Hupé JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature 1998;394(6695):784–7.

[24] Stollenga MF, Masci J, Gomez F, Schmidhuber J. Deep networks with internal selective attention through feedback connections. In: Proceedings of the 27th International Conference on Neural Information Processing Systems; 2014 Dec 8–13; Montreal, Canada. Cambridge, MA: MIT Press; 2014. p. 3545–53.

[25] Ozawa T, Ycu EA, Kumar A, Yeh LF, Ahmed T, Koivumaa J, et al. A feedback neural circuit for calibrating aversive memory strength. Nat Neurosci 2017;20(1):90–7.

[26] Williams MA, Baker CI, Op de Beeck HP, Shim WM, Dang S, Triantafyllou C, et al. Feedback of visual object information to foveal retinotopic cortex. Nat Neurosci 2008;11(12):1439–45.

[27] Cao C, Huang Y, Yang Y, Wang L, Wang Z, Tan T. Feedback convolutional neural network for visual localization and segmentation. IEEE Trans Pattern Anal Mach Intell 2019;41(7):1627–40.

[28] Cao C, Liu X, Yang Y, Yu Y, Wang J, Wang Z, et al. Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. Washington, DC: IEEE Computer Society; 2015. p. 2956–64.

[29] Haris M, Shakhnarovich G, Ukita N. Deep back-projection networks for super-resolution. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. Washington, DC: IEEE; 2018. p. 1664–73.

[30] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia. JMLR.org; 2017. p. 1183–92.

[31] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115–8.

[32] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24 (10):1559–67.

[33] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.

[34] Arnaout R. Toward a clearer picture of health. Nat Med 2019;25(1):12.

[35] Howard A, Sandler M, Chen B, Wang W, Chen L, Tan M, et al. Searching for mobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. Washington, DC: IEEE; 2019.

[36] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. Washington, DC: IEEE; 2016. p. 770–8.

[37] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39 (6):1137–49.

[38] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. Washington, DC: IEEE; 2017. p. 936–44.

[39] Liu Z, Hu J, Weng L, Yang Y. Rotated region based CNN for ship detection. In: Proceedings of the 2017 IEEE International Conference on Image Processing; 2017 Sep 17–20; Beijing, China. Washington, DC: IEEE; 2017. p. 900–4.

[40] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021;599 (7883):91–5.

[41] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95.

[42] Krygier MC, LaBonte T, Martinez C, Norris C, Sharma K, Collins LN, et al. Quantifying the unknown impact of segmentation uncertainty on image-based simulations. Nat Commun 2021;12(1):5414.

[43] Lemhadri I, Ruan F, Abraham L, Tibshirani R. LassoNet: neural networks with feature sparsity. J Mach Learn Res 2021;22:1–29.

[44] Suway SB, Schwartz AB. Activity in primary motor cortex related to visual feedback. Cell Rep 2019;29(12):3872–84.e4.

[45] Marques T, Nguyen J, Fioreze G, Petreanu L. The functional organization of cortical feedback inputs to primary visual cortex. Nat Neurosci 2018;21 (5):757–64.