Modeling Cardiac Normality for Neonatal CHD Screening on a Portable Single-View Device: Challenges of a Small-Sample Pediatric Pilot

Rodrigo Medellin-Robles[1]*, Mohammad Abdullah Bukhari[1], Jose Miguel Torres-Martel[2], Maria Victorova[1], Isidro Gutierrez-Alvarez[2], Saul Tovar-Arriaga[2], and Annamarie Saarinen[1]

[1] Bloom Standard, Minneapolis, Minnesota, United States
[2] Autonomous University of Queretaro, Santiago de Queretaro, Queretaro, Mexico

Abstract. Congenital heart disease (CHD) is the most common congenital malformation and a leading cause of infant mortality. Diagnostic echocardiography requires trained specialists; it is therefore hard to deliver at the point of care. In real screening, confirmed CHD is rare and abnormality labels are correspondingly few; this scarcity limits supervised detection. We present a feasibility pilot of CHD screening from a single subcostal video acquired with the RAPID EVA device (Bloom Standard, USA). Instead of learning CHD from scarce positives, we model cardiac normality: we train only on healthy hearts and flag deviations, using no CHD labels. We evaluated 64 patients (9 CHD) with patient-level five-fold cross-validation and a patient-cluster bootstrap; all reported AUCs are patient-level. The normality autoencoder reached an AUC of 0.87, against 0.93 for a fully supervised spatiotemporal classifier. At the same per-frame granularity, supervision reached only 0.60. A demographic baseline indicates that the anomaly is driven by the cardiac image rather than by age, and occlusion attribution localizes it to the cardiac region rather than the image borders. Modeling cardiac normality is thus a practical direction toward point-of-care CHD screening, because it needs only the normal studies a birth facility already collects.

Keywords: Congenital heart disease · Point-of-care ultrasound · Anomaly detection · One-class learning · Small-sample learning · Interpretability.

1 Introduction

Congenital heart disease (CHD) is the most common congenital malformation and a leading cause of infant mortality [1]. Its incidence is about 9 per 1,000 births, and roughly a quarter of cases are critical and need intervention within the first year [2]. Early detection is therefore urgent, yet it remains hard. Pulse oximetry misses a quarter of critical defects and about half of all major defects [3]. Fetal echocardiography detects only 30 to 50 percent, limited by examiner skill [4]. Transthoracic echocardiography is definitive; however, it needs

* Corresponding author: rodrigo@bloomstandard.com, ORCID 0009-0008-4288-8622.

a specialist and a 17-view protocol lasting 30 to 60 minutes [5]. Scarce expertise is thus the main bottleneck where the CHD burden is highest.

Handheld and portable ultrasound has widened access at the point of care [6], and AI support can improve the consistency of acquisition and interpretation [7]. We study one such setting: a single subcostal view acquired with the RAPID EVA device (Bloom Standard, USA). This window gives quick access in newborns [8]. The coronal long-axis view is among the best for detecting atrial and ventricular septal defects, because shunt jets run parallel to the ultrasound beam; it also shows atrioventricular valve regurgitation and ventricular size and function [8]. The setting carries pediatric-specific challenges. Neonatal hearts change quickly with age and size; therefore an unmatched cohort can confound growth with disease. A single portable view shows only part of the anatomy. Confirmed CHD is rare and heterogeneous; furthermore, it is ascertained only after an abnormal screen, so positives and abnormality labels are scarce.

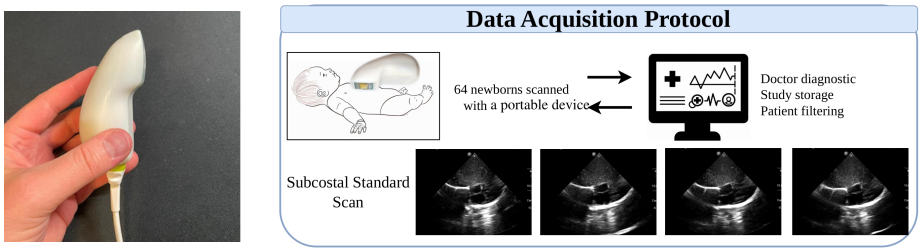
These challenges are a poor fit for supervised CHD detection. Jiang *et al.* [9] detect CHD from seven standard views. Hong *et al.* [10] detect secundum atrial septal defects from color-Doppler images. Tang *et al.* [11] use two-stage transfer learning for duct-dependent CHD in fetal echocardiography (see [12,13] for reviews). These systems need multiple standard views and labeled abnormal studies; both are scarce in single-view point-of-care screening.

We therefore detect CHD by modeling cardiac *normality*. Instead of learning a boundary from few positives, we train only on the abundant healthy studies and flag CHD as a deviation, using no CHD labels. Normality (one-class) modeling has been used for anomaly detection in medical imaging [14,15] and echocardiography [16], but not, to our knowledge, for single-view subcostal CHD screening in neonates. This is a feasibility pilot, not a state-of-the-art claim. We frame small, imbalanced, single-site cohorts, biased ascertainment, and single-view coverage as pediatric-specific screening challenges; furthermore, we report what a normality-only model achieves under them. We address them with an age-matched, patient-level evaluation, paired testing, and occlusion-based attribution, so that a positive flag reflects pathology rather than population shift. The intended output is a referral decision and not a diagnosis. The flag sends a neonate to a formal echocardiogram; today that referral rests on pulse oximetry and clinical examination alone.

2 Materials and Methods

2.1 Acquisition Device and Protocol

Data were acquired with the RAPID EVA device (Bloom Standard, USA): a 64-element CMUT phased array producing a conic sector image at 3.5 MHz, with a 9–12 cm depth and 30 fps. All examinations used a single acoustic window; under a standardized protocol, two to three short subcostal coronal (long-axis) cine clips were obtained per patient by experienced pediatric cardiologists (Fig. 1).



(a) (b)

Fig. 1. (a) The RAPID EVA portable pediatric ultrasound device (Bloom Standard, USA), used for every acquisition in this study; the probe is operated single-handed at the cot side. (b) Single-view subcostal acquisition protocol. Each newborn is scanned through the subcostal window; studies are then stored and filtered for analysis. Two to three subcostal coronal (long-axis) cine clips are obtained per patient, with representative frames shown.

Where a reference diagnosis required a conventional cart-based system, those recordings were excluded from model development.[3]

2.2 Cohort

All training and evaluation use a single device version and an *age-matched* selection of 64 patients (55 normal, 9 CHD) contributing 168 standard subcostal scans. The nine CHD cases comprise four ventricular septal defects, three patent ductus arteriosus, and two pulmonary atresia, the lesions ascertained to date in this ongoing screening cohort; each CHD was confirmed by a reference cart-based echocardiographic examination following an abnormal portable screen. Positives are therefore conditioned on portable detectability, so the reported sensitivity is likely optimistic. Routine care at this site is universal pulse oximetry plus clinical examination. A cart-based echocardiogram is available in the same hospital; however, it is performed only for neonates flagged by that routine screen. Table 1 reports the demographics of both classes; the small between-class age gap means age alone carries little discriminative information, by design (Sec. 3).

2.3 Preprocessing and Representations

Recordings were standardized to 30 fps and cropped to the ultrasound sector with a *fixed* conic mask (identical for all videos, calibrated once on a class-balanced sample) that removes on-screen text and margins without per-image

[3] The parents of each neonate gave written informed consent before screening, and all echocardiographic data were de-identified before analysis. The study was approved by the National Center for Maternal and Child Health of Mongolia; approval was also granted by the Ethics Committee of the National University of Medical Sciences, Ulaanbaatar (No. 149, 2 April 2024).

Table 1. Demographics and clinical characteristics of the age-matched cohort. Data are mean ± SD or n (%); no values were imputed.

	Normal	**CHD**
n	55	9
Age, days	2.2 ± 0.8	3.4 ± 2.9
Gender, male	27 (49%)	5 (56%)
Height, cm	50.3 ± 1.7	49.7 ± 2.3
Weight, kg	3.5 ± 0.4	3.2 ± 0.6

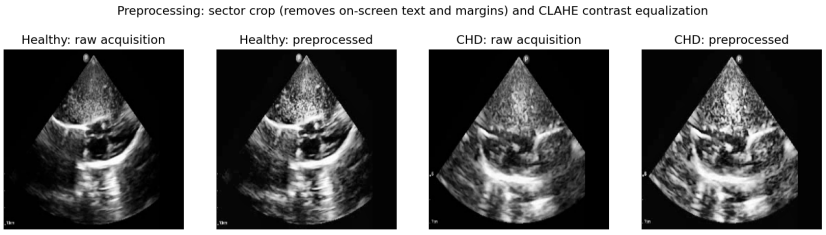


Fig. 2. Preprocessing on representative frames. In each pair, the raw single-view subcostal acquisition (left) is cropped to the ultrasound sector (on-screen text and margins removed), resized to 256×256, and contrast-equalized with CLAHE (right). A healthy and a CHD example are shown.

adaptation. Frames were then resized to 256×256 and contrast-equalized with CLAHE [17] (Fig. 2). The fixed, class-agnostic mask is deliberate; a per-image or per-class crop could leak label information, whereas a single calibrated mask removes device text and margins uniformly.

Each frame is mapped to a 512-d feature vector by a frozen ImageNet VGG16 [18] with global average pooling. We additionally evaluate three alternative backbones (ResNet-50 and ResNet-18 [19], Inception-V3 [20]) and an echo-specific video foundation model (EchoFM [21]) as ablations. EchoFM is used as a frozen extractor; subcostal clips are passed through the pretrained encoder with its classification head removed to obtain per-clip embeddings, which feed the same normality models.

2.4 Detection Methods

(a) Normality-only training (proposed). Models are fit on features of *healthy frames only* ($\approx$ 55,000 frames, a training-data volume and not an evaluation unit): a shallow autoencoder (512–256–64–256–512, ReLU, MSE reconstruction; Adam, learning rate 10^{-3}, batch size 256, 40 epochs), a one-class SVM [22] with a radial-basis-function (RBF) kernel, and an Isolation Forest [23]. The per-frame anomaly (or reconstruction) score is aggregated per patient by its mean. No abnormality labels are used in training.

(b) Supervised reference. A two-stage spatiotemporal classifier. Frozen per-frame features feed a trained temporal head (bidirectional GRU [24] with multi-head attention and masked temporal pooling) predicting normal/abnormal. As a granularity control, we also train *per-frame* supervised classifiers on the same features and folds (logistic regression, RBF-SVM, random forest, k-NN and gradient boosting). We report all five families rather than selecting one, because choosing a family by cross-validated AUC would reuse the labels we test on.

(c) Demographic baseline. Logistic regression on patient age, over the same patients and folds.

Occlusion attribution. A 40×40 patch (stride 12) filled with a blurred copy of the frame is slid over the image; the reconstruction anomaly is recomputed at each position, and the resulting map is averaged over the 14 most anomalous frames per clip.

2.5 Evaluation

All reported AUCs are *patient-level*. Methods use patient-level five-fold stratified group cross-validation (a subject's scans stay within one fold) and share the same folds and seed. Each AUC is computed on the pooled out-of-fold predictions (one score per patient, concatenated across all folds), rather than by averaging per-fold AUCs, which is unstable with one to two positives per fold. The $\approx$55,000 healthy frames are a training-data volume only, not a unit of evaluation, and do not enter the patient count. We report AUC with a *patient-cluster* bootstrap 95% confidence interval (resampling patients, not videos). Because marginal intervals are wide at this sample size, we additionally report *paired* bootstrap differences Δ between methods (same patients and folds) with a permutation test, and a leave-one-CHD-out jackknife for stability. Given the small positive count we report no winner and treat single differences as suggestive. The autoencoder architecture and the VGG16 representation were fixed before analysis. The training and evaluation code is available from the corresponding author on reasonable request. The de-identified dataset is available on the same basis, subject to approval by the participating institutions.

3 Results

At nine positives the marginal intervals are wide and small AUC gaps correspond to about one patient, so we lead with what is stable and read single differences as suggestive.

Main comparison. Table 2 lists all methods and families. The normality-only autoencoder trails the supervised classifier by only Δ=0.06 (95% CI $[-0.04, 0.18]$) yet uses no CHD labels, while every per-frame supervised classifier sits at 0.50–0.60; supervision helps only with temporal context. The autoencoder also exceeds the age prior (Δ=0.29, CI $[0.01, 0.57]$), and the occlusion analysis below argues the signal is cardiac rather than demographic.

Table 2. Patient-level five-fold cross-validation (n=64, 9 CHD). Sens/Spec at the Youden-optimal point, AUC with patient-cluster bootstrap 95% CI. Normality models use *no* abnormality labels. All five per-frame supervised classifiers are listed, so none is cherry-picked. Operating points are chosen in-sample and are optimistic at this n; we lead with AUC.

Family	**Method**	**Sens**	**Spec**	**AUC**	**95% CI**
Supervised	Spatiotemporal (BiGRU)	1.00	0.87	0.93	0.86–0.98
	Per-frame RBF-SVM	1.00	0.31	0.60	0.40–0.78
	Per-frame LogReg	0.44	0.84	0.58	0.33–0.81
	Per-frame RandomForest	0.56	0.78	0.57	0.31–0.82
	Per-frame k-NN	0.89	0.25	0.52	0.31–0.71
	Per-frame GradBoost	0.78	0.40	0.50	0.32–0.69
Normality-only	Autoencoder	0.89	0.82	**0.87**	0.76–0.96
	One-class SVM	0.89	0.78	0.83	0.69–0.94
	Isolation Forest	0.89	0.69	0.81	0.66–0.93
	EchoFM + Isolation Forest	1.00	0.51	0.64	0.49–0.77
Baseline	Demographics (age only)	0.33	1.00	0.58	0.28–0.85

Ablations. Table 3 summarizes robustness. Temporal modeling does not beat per-frame appearance for the normality branch, and backbone AUCs span only 0.85–0.93; at nine positives that spread is about one patient, so we read them as instability, not a ranking. With the anomaly detector held fixed, a generic ImageNet backbone still outperforms the echo-specific foundation model as a representation for the normality models.

Occlusion attribution. For the nine confirmed-CHD cases, occluding the central cardiac region reduces the anomaly score roughly twice as much as occluding the periphery (median heart/periphery ratio 2.1, range 1.0 to 2.9, with one case showing no cardiac preference); Fig. 3 shows two examples. We report this as a qualitative localization check and not as a discriminative test.

4 Discussion

We modeled cardiac normality with three anomaly detectors fit on healthy frames only. The autoencoder gave the highest normality point estimate and approaches the fully supervised spatiotemporal classifier (0.87 against 0.93), using no CHD labels; the three normality intervals overlap, so we claim no winner among them. The paired difference is not significant and the interval allows a supervised advantage of up to 0.18 AUC, so we claim no equivalence; more positives are needed to tell whether the gap is real.

At the same per-frame granularity the supervised classifier reached only 0.60. Temporal aggregation lifted it to 0.93 but did not help the normality branch. A single frame may not separate 9 positives from 55 normals, so the supervised

Table 3. Ablations (patient-level). Sensitivity and specificity at the Youden-optimal point. No winner is claimed at this n.

Axis	Setting	Sens	Spec	AUC
Granularity (normality)	per-frame AE	0.89	0.82	0.87
	sequence AE	1.00	0.65	0.83
	motion-only	1.00	0.64	0.84
Backbone (supervised model)	VGG16	1.00	0.87	0.93
	ResNet-50	1.00	0.76	0.88
	ResNet-18	0.78	0.85	0.87
	Inception-V3	1.00	0.58	0.85
Representation (Isolation Forest)	VGG16-avg	0.89	0.69	0.81
	EchoFM	1.00	0.51	0.64

model needs temporal context, whereas the normality model, fit on roughly 55,000 healthy frames, does not. We treat this asymmetry as suggestive.

This is the practical message of the study. To flag a possible CHD the model only needs to know what a normal heart looks like, and normal studies are abundant in any birth facility whereas confirmed CHD is rare; a supervised detector there would have very few positives to learn from.

Intended clinical role. A single normal/abnormal flag is deliberately coarse. It does not compete with echocardiographic diagnosis; it targets the referral decision, which is already binary and which at our site rests on universal pulse oximetry plus clinical examination. Pulse oximetry reached 75 percent sensitivity for critical and 49 percent for all major defects in 20,055 newborns [3], and is blind by construction to lesions that do not desaturate; seven of our nine positives are ventricular septal defects or patent ductus arteriosus. A neonate with a small septal defect, normal saturations and no murmur is discharged today without an echocardiogram; a coarse flag suffices to change that, because the downstream action is single and already available. We did not measure incremental yield: our positives were ascertained after an abnormal routine screen, so estimating it needs a prospective design. A cart-based system was also on site here; a portable screen is worth more where it is not.

On the age-matched cohort, age alone was near chance (0.58), confirming the match removed the age confound. The image methods exceed this age prior, and the occlusion analysis supports a cardiac origin for the anomaly score.

With the anomaly detector held fixed (Isolation Forest), a generic ImageNet representation outperformed the echo-specific foundation model (0.81 against 0.64), which we attribute to domain mismatch: EchoFM was pretrained on adult, cart-based, multi-view studies [21].

We used occlusion instead of a gradient saliency map, which sanity checks have shown to be model-agnostic [25]: occlusion perturbs the actual model and measures the causal effect on the anomaly score. Still, the residual risk of shortcut

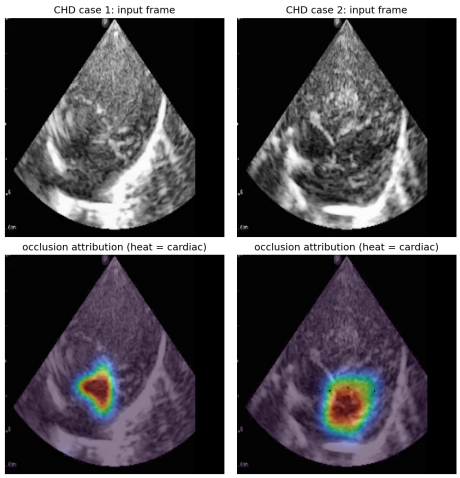


Fig. 3. Occlusion-based attribution of the normality-only anomaly score for two confirmed-CHD cases (top: input frame; bottom: overlay). Heat marks regions whose occlusion most reduces the reconstruction anomaly, which concentrates on the cardiac region rather than the sector border, consistent with a score driven by cardiac appearance rather than acquisition artifacts.

learning from single-device acquisition [26,27] is real; the occlusion result and the fixed mask reduce it but do not remove it.

The evaluation protocol is a key strength: all methods share the same patient-level folds and seed, the sector mask is fixed and class-agnostic to avoid label leakage, and the supervised baseline is a generous upper bound rather than a strawman. The limitations follow from the setting. The small positive count (9 CHD) yields wide marginal intervals, so we claim no winner and rely on paired differences and a leave-one-CHD-out jackknife (AUC 0.850 to 0.898 for the autoencoder, 0.925 to 0.939 for the supervised model). Results are single-site, from a single device version, with acquisition by experienced pediatric cardiologists; the non-specialist point-of-care use the device is intended for is therefore a design goal, not a claim demonstrated here. Five of the nine positives are patent ductus arteriosus or pulmonary atresia, whose direct assessment from the subcostal window is limited [8], so the model may be responding to secondary chamber-level consequences rather than to the primary lesion. These are not incidental caveats but the pediatric-specific challenges any neonatal point-of-care CHD screener must confront; stating them plainly, and showing what a normality-only model can still do under them, is part of this paper's contribution.

Future work will grow the cohort, validate across devices and sites, and explore video-based normality modeling and prospective studies with non-specialist operators; cross-version transfer to a second device iteration was exploratory and inconclusive.

5 Conclusion

We presented a feasibility pilot of modeling cardiac normality for CHD screening from single-view subcostal video on a portable pediatric ultrasound device. Training on healthy hearts only, and using no abnormality labels, the approach reached an AUC of 0.87 and approaches a fully supervised spatiotemporal detector (0.93); the paired difference is not significant at this sample size, and we claim no equivalence. At equal per-frame granularity the supervised classifier reached 0.60, and age alone reached 0.58 on the age-matched cohort. At nine positives, these results are suggestive rather than conclusive. Still, in the label-scarce reality of portable CHD screening, modeling normality is a practical direction, because it only requires the normal studies that a birth facility already collects.

Acknowledgments. We are indebted to Dr. Ulziikhishig Byambabayar and Dr. Erdenejargal Badarch of the Neonatal Department, First Maternity Hospital, Ulaanbaatar, Mongolia. They performed every echocardiographic acquisition reported here, ran the screening program at the bedside, and made the diagnoses. We further thank the Bloom Standard team for their efforts.

Disclosure of Interests. R. Medellin-Robles, M. A. Bukhari, M. Victorova and A. Saarinen are employees of Bloom Standard, which develops the device used in this study. The remaining authors declare no competing interests. Bloom Standard had no role in the evaluation protocol, fixed before analysis, or in the decision to report negative results.

References

1. Du, C., Liu, D., Liu, G., Wang, H.: A meta-analysis about the screening role of pulse oximetry for congenital heart disease. BioMed Res. Int. **2017**, 2123918 (2017)
2. Klausner, R., Shapiro, E.D., Elder, R.W., Colson, E., Loyal, J.: Evaluation of a screening program to detect critical congenital heart defects in newborns. Hosp. Pediatr. **7**(4), 214–218 (2017)
3. Ewer, A.K., Middleton, L.J., Furmston, A.T., et al.: Pulse oximetry screening for congenital heart defects in newborn infants (PulseOx): a test accuracy study. Lancet **378**(9793), 785–794 (2011)
4. Komatsu, M., Sakai, A., Komatsu, R., et al.: Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning. Appl. Sci. **11**(1), 371 (2021)
5. Lopez, L., Saurers, D.L., Barker, P.C.A., et al.: Guidelines for performing a comprehensive pediatric transthoracic echocardiogram: recommendations from the American Society of Echocardiography. J. Am. Soc. Echocardiogr. **37**(2), 119–170 (2024)
6. Baribeau, Y., Sharkey, A., Chaudhary, O., et al.: Handheld point-of-care ultrasound probes: the new generation of POCUS. J. Cardiothorac. Vasc. Anesth. **34**(11), 3139–3145 (2020)
7. Akkus, Z., Aly, Y.H., Attia, I.Z., et al.: Artificial intelligence (AI)-empowered echocardiography interpretation: a state-of-the-art review. J. Clin. Med. **10**(7), 1391 (2021)

8. Kurath-Koller, S., Koestenberger, M., Hansmann, G., Cantinotti, M., Tissot, C., Sallmon, H.: Subcostal echocardiographic imaging in neonatal and pediatric intensive care. Frontiers in Pediatrics **9**, 471558 (2021)
9. Jiang, X., Yu, J., Ye, J., et al.: A deep learning-based method for pediatric congenital heart disease detection with seven standard views in echocardiography. World J. Pediatr. Surg. **6**(3), e000580 (2023)
10. Hong, W., Sheng, Q., Dong, B., et al.: Automatic detection of secundum atrial septal defect in children based on color Doppler echocardiographic images using convolutional neural networks. Front. Cardiovasc. Med. **9**, 834285 (2022)
11. Tang, J., Liang, Y., Jiang, Y., et al.: A multicenter study on two-stage transfer learning model for duct-dependent CHDs screening in fetal echocardiography. npj Digit. Med. **6**(1), 143 (2023)
12. Wang, A., Doan, T.T., Reddy, C., Jone, P.N.: Artificial intelligence in fetal and pediatric echocardiography. Children **12**(1), 14 (2025)
13. Vasile, C.M., et al.: Exploring the potential of artificial intelligence in pediatric echocardiography. J. Clin. Med. **12**(9), 3209 (2023)
14. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: IPMI, LNCS **10265**, pp. 146–157. Springer (2017)
15. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Medical Image Analysis **69**, 101952 (2021)
16. Wahlang, I., Maji, A.K., Saha, G., Chakrabarti, P., Jasinski, M., Leonowicz, Z., Jasinska, E.: Deep learning methods for classification of certain abnormalities in echocardiography. Electronics **10**(4), 495 (2021)
17. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing **39**(3), 355–368 (1987)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)
21. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q.: EchoFM: foundation model for generalizable echocardiogram analysis. arXiv:2410.23413 (2024)
22. Schölkopf, B., et al.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7), 1443–1471 (2001)
23. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation Forest. In: ICDM, pp. 413–422 (2008)
24. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP, pp. 1724–1734 (2014)
25. Adebayo, J., et al.: Sanity checks for saliency maps. In: NeurIPS (2018)
26. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (2020)
27. DeGrave, A.J., Janizek, J.D., Lee, S.-I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**, 610–619 (2021)