Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33150

Lightweight Transformer Architectures for Embedded Congenital Heart Disease Screening

Zhixiang Sun

School of Electronic Engineering, Xidian University, Xi'an, China
ss4042@hw.ac.uk

Abstract. Using phonocardiogram (PCG) analysis to screen congenital heart disease (CHD) at an early stage can help with earlier referral and treatment. This is especially so in low-resource settings where echocardiography is not easily available. Recent transformer-based models have given good results in heart-sound classification, as self-attention is able to capture long-range temporal dependencies across the cardiac cycles. Standard transformers, however, require substantial computation and are hard to deploy on microcontrollers or portable digital stethoscopes, as these have strict limits on memory, power consumption, and latency. This review focuses on lightweight Transformer strategies for embedded CHD screening, such as sparse or linearized attention, CNN-Transformer hybrid architectures, quantization, and knowledge distillation. It also discusses how these methods influence the diagnostic accuracy, robustness to noisy pediatric recordings, model size, and real-time inference. It also identifies several problems that are yet to be solved, such as interpretability, dataset bias, variation in neonatal heart sounds, and the practical difficulty of using multimodal fusion in a point-of-care device. Overall, a lightweight Transformer design provides a path for developing accessible, hardware-efficient AI systems for early CHD screening.

Keywords: transformer, congenital heart disease, phonocardiogram, model compression, edge AI

1. Introduction

Congenital heart disease (CHD) is still a major pediatric cardiovascular problem, and detection as early as possible is especially important in settings where there is little access to echocardiography [1]. In this setting, the analysis of phonocardiogram (PCG) is a good option, since it is cheap, non-invasive, and easily compatible with the use of portable digital stethoscopes. However, manual auscultation is still subjective and depends on the clinical experience. Therefore, on the other hand, researchers have been developing automated PCG-based screening systems in order to support the frontline clinicians and also increase the pre-screening capacity.

The early automated heart-sound pipelines were mostly based on explicit segmentation and probabilistic models. Springer et al. used a logistic regression-hidden semi-Markov model for the robust heart-sound segmentation, in order to standardize the feature extraction that follows [2]. Later, the convolutional neural network (CNNs) improved the classification, because it was able to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33150

learn from the time-frequency representations, such as the mel-frequency cepstral coefficient map, rather than rely fully on handcrafted features [3].

Transformer models offered a different advantage because self-attention can capture long-range dependencies across full cardiac cycles and model patterns that local convolution alone cannot easily represent [4]. Recent heart-sound studies, therefore, have begun to combine convolutional modules with transformer encoders in order to retain local acoustic detail while improving global sequence modeling [5,6]. Even so, a standard Transformer still has attention complexity with respect to sequence length, which makes embedded deployment difficult on microcontrollers with tight constraints on memory, energy consumption, and latency [4,7]. This paper examines how a lightweight Transformer design may reduce the gap between strong classification performance and practical real-world deployment. It focuses on architectural simplification, hardware-aware optimization, interpretability, and the additional constraints that are imposed by the use cases of pediatric and congenital screening.

2. Structural and optimization strategies

2.1. Revisiting self-attention for quasi-periodic signals

PCG signals often do not require full pairwise attention because heart sounds are quasi-periodic and show repeated local structures across adjacent cycles. Efficient variants of Transformer therefore seek to lower the cost of attention by using sparsity, low-rank approximation, or shorter sequences [7]. In practice, one strategy that works well is to use a light-weight convolutional front end before the Transformer, so that the encoder will process a shorter and more informative sequence [5,8]. A representative example is the Convolution and Transformer Encoder Neural Network (CTENN) of Cheng and Sun. The model uses one-dimensional convolution to extract local features and a Transformer encoder to model longer-range patterns, and it showed strong performance on multiple heart-sound benchmarks while making preprocessing simpler [5]. The larger lesson is that lightweight design in this domain is not only about cutting parameters; it also means reorganizing the pipeline so that attention is applied where it provides the most useful information.

2.2. Hardware-aware compression and the INT8 standard

Even with the architectural simplification, many models are still too large or too slow for use in embedded screening devices. Therefore, the hardware-aware optimization is important. Quantization reduces the precision of parameters and the usage of memory. Pruning, together with the teacher-student compression, can also cut the computational cost [7]. Among these methods, 8-bit integer quantization is especially useful because the mature embedded inference toolchains already support it. Knowledge distillation also shows the promise of heart-sound tasks. Song et al. found that a distillation framework could compress deep learning models for heart-sound classification and still keep the competitive performance. This makes it more suitable for the resource-constrained hardware [9]. However, the aggressive compression must not be too aggressive, because the subtle acoustic cues, especially high-frequency murmur information, may be important for CHD screening.

2.3. Hybrid architectures and noise robustness

The hybrid CNN-Transformer systems are intended to combine the merits of both families of models, where convolution captures the morphology and local spectral texture and attention

DOI: 10.54254/2755-2721/2026.GU33150

modeling longer temporal dependencies [5,10]. These architectures are attractive for embedded CHD screening as they can provide strong performance if they do not apply a fully dense Transformer to the whole signal.

Still, strong benchmark accuracy does not necessarily mean that clinical deployment will be robust. Reviews of heart-sound deep learning have always pointed to environmental noise, device heterogeneity, and domain shift as persistent challenges [11]. In the case of pediatric screening, background crying, motion artefacts, and inconsistent sensor placement can distort the same patterns that attention is expected to model. Lightweight classification must be used together with practical denoising, signal quality control, or data augmentation strategies so that the system can remain reliable in real clinical settings [11,12].

2.4. Distillation, dataset bias, and pediatric specificity

Compression on its own does not address the problem of generalization. Distilled student models often inherit the statistical bias present in their teachers and training datasets [9]. This is important in congenital screening because neonatal and infant heart sounds are different from adult recordings in rate, spectral distribution, and signal quality. If most of the public training data are drawn from broader heart-sound abnormality detection tasks, a compact model may achieve good overall performance but still underperform on the specific pediatric phenotypes that matter most for CHD referral. Recent reviews have pointed out that dataset composition, label quality, and demographic balance remain major limitations in heart-sound AI [11,13].

Xu et al. responded to this problem by constructing a pediatric CHD heart-sound dataset and designing a classification pipeline for that particular clinical target [6].

Recent reviews have indicated that dataset composition, label quality, and demographic balance continue to be major limitations in heart-sound AI [11,13]. Xu et al. tackled this issue by creating a pediatric CHD heart-sound dataset and designing a classification pipeline for that particular clinical purpose [6].

2.5. Multi-modal fusion for enhanced specificity

Although this review is centered on PCG-based screening, multimodal approaches still matter because ECG or other physiological signals can help make timing clearer and improve specificity. Recent reviews suggest that combining multiple modalities is a promising direction in heart-sound AI and in congenital applications more generally [11,13]. For example, ECG-based temporal alignment can make systolic and diastolic events easier to identify and may reduce ambiguity when the acoustic signal quality is poor. At the same time, multimodal fusion brings additional deployment costs. Extra sensors add hardware complexity, synchronized acquisition is not always possible in low-resource environments, and fusion layers can also increase inference latency. For point-of-care screening devices, the most practical future direction may be adaptive systems that use PCG alone as the default and incorporate ECG only when it is available and reliable [11,13].

2.6. Performance snapshot

Table 1 summarizes the representative strategies discussed in this review. The studies rely on different datasets, tasks, and evaluation protocols, so the table should be considered as a qualitative comparison and not a strict leaderboard.

DOI: 10.54254/2755-2721/2026.GU33150

Table 1. Representative strategies for embedded CHD screening

Strategy	Source	Main idea	Reported strength	Main caveat
Hybrid convolution-Transformer	[5]	1D convolution front end with Transformer encoder for joint local-global modeling	Strong benchmark performance across multiple heart-sound datasets	Noise robustness on edge recordings is still uncertain
Knowledge distillation/compression	[9]	Teacher-student compression for lighter heart-sound classifiers	Reduces model size and supports resource-constrained deployment	May transfer teacher bias or discard subtle cues
Pediatric CHD-specific modeling	[6]	Dataset and classifier built around pediatric congenital screening	Directly targets the intended clinical population	Needs broader validation beyond the collected cohort
Explainable attention / interpretable representations	[10], [14]	Uses attention or interpretable representations to increase model transparency	Improves prospects for clinician trust and error analysis	Interpretations remain indirect and are not always phase-aligned

3. Challenges and future directions

3.1. Acoustic heterogeneity and device shift

One major translational problem is the mismatch of the curated training data with the real recordings taken in busy clinics or in the community. Across the different sites, the sensor response, place of the sensor, quality of the placement, ambient noise, and the movement of the patient all can be different [11,12]. Lightweight models are especially prone to this because they have less unused capacity to cope with the nuisance variation. Future embedded systems will thus need not only efficient backbones, but also robust front-end preprocessing, confidence estimation, and clear quality-control mechanisms.

3.2. Clinician-centric interpretability

Attention maps are sometimes treated as an explanation. But not all attention patterns reflect physiologically meaningful events. Ren et al. and Wang et al. both show that explainability can be achieved in heart-sound analysis. But the current work is still not complete [10,14]. Interpretability is important in CHD screening because a system that simply gives a risk score is unlikely to be trusted in clinical practice. The next step is to bring the model explanation closer to recognizable cardiac phase, murmur timing, and signal regions that the clinician can interpret more easily.

3.3. Fair evaluation for embedded screening

Some papers report binary normal-abnormal classification, whereas others deal more directly with pediatric CHD recognition. It is still difficult to compare reported performance across studies because datasets, class definitions, preprocessing choices, and evaluation metrics differ substantially [1,11]. Embedded evaluation creates another layer of complexity because studies often report memory footprint, energy use, and latency inconsistently. More realistic benchmarking protocols

DOI: 10.54254/2755-2721/2026.GU33150

should therefore assess both diagnostic performance and deployment-related metrics, while also including noisy pediatric recordings captured on lower-cost hardware.

3.4. Toward clinically viable systems

The shift from laboratory prototypes to clinically useful tools will require a broader validation strategy. Reviews in both heart-sound AI and congenital cardiology continue to stress the need for prospective studies, better external validation, and closer alignment between model design and real clinical workflows [11,13]. In practice, this means evaluating in unseen hospitals, recording the failure modes clearly, and designing the outputs to help with triage, not to overstate the certainty. If the researchers develop lightweight Transformers with these constraints in mind, they may become practical decision-support tools for early CHD screening, instead of staying as models that are only effective in the benchmark setting.

4. Conclusion

Lightweight Transformer research has created a practical path toward heart-sound-based embedded CHD screening. Sparse attention, convolution-Transformer hybrids, quantization, and knowledge distillation all help make global sequence modeling more feasible on low-power devices. However, deployment is not simply a compression issue. In congenital screening, success also requires relevant pediatric data, robustness to noisy recordings, interpretability, and realistic evaluation. In that sense, the most promising systems are probably the ones that combine efficient architectures with clear clinical framing, with age-aware datasets, clear decision support, and the metrics of the deployment reported together with the accuracy. If the progress continues in those directions, the lightweight Transformers may help to bring the AI-assisted auscultation from the research to the point of care as screening tools for the early detection of CHD.

References

[1] Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E. W., Syed, Z., Schmidt, S. E., Papadaniil, C. D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., ... Clifford, G. D. (2016). An open-access database for the evaluation of heart sound algorithms. Physiological Measurement, 37(12), 2181-2213.

[2] Springer, D. B., Tarassenko, L., & Clifford, G. D. (2016). Logistic regression-HSMM-based heart sound segmentation. IEEE Transactions on Biomedical Engineering, 63(4), 822-832.

[3] Rubin, J., Abreu, R., Ganguli, A., Nelaturi, S., Matei, I., & Sricharan, K. (2016). Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients. Computing in Cardiology, 43, 813-816.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

[5] Cheng, J., & Sun, K. (2023). Heart sound classification network based on convolution and transformer. Sensors, 23(19), 8168.

[6] Xu, W., Yu, K., Ye, J., Li, H., Chen, J., Yin, F., Xu, J., Zhu, J., Li, D., & Shu, Q. (2022). Automatic pediatric congenital heart disease classification based on heart sound signal. Artificial Intelligence in Medicine, 126, 102257.

[7] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1-28.

[8] Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In International Conference on Learning Representations.

[9] Song, Z., Zhu, L., Wang, Y., Sun, M., Qian, K., Hu, B., Yamamoto, Y., & Schuller, B. W. (2023). Cutting weights of deep learning models for heart sound classification: Introducing a knowledge distillation approach. In 2023 45th

DOI: 10.54254/2755-2721/2026.GU33150

Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 1-4). IEEE.

[10] Ren, Z., Qian, K., Dong, F., Dai, Z., Nejdl, W., Yamamoto, Y., & Schuller, B. W. (2022). Deep attention-based neural networks for explainable heart sound classification. Machine Learning with Applications, 9, 100322.

[11] Zhao, Q., Geng, S., Wang, B., Sun, Y., Nie, W., Bai, B., Yu, C., Zhang, F., Tang, G., Zhang, D., Zhou, Y., Liu, J., & Hong, S. (2024). Deep learning in heart sound analysis: From techniques to clinical applications. Health Data Science, 4, 0182.

[12] Chorba, J. S., Shapiro, A. M., Le, L., Maidens, J., Prince, J., Pham, S., et al. (2021). Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform. Journal of the American Heart Association, 10(9), e019905.

[13] Jone, P.-N., Gearhart, A., Lei, H., Xing, F., Nahar, J., Lopez-Jimenez, F., Diller, G.-P., Marelli, A., Wilson, L., Saidi, A., Cho, D., & Chang, A. C. (2022). Artificial intelligence in congenital heart disease: Current state and prospects. JACC: Advances, 1(5), 100153.

[14] Wang, Z., Qian, K., Liu, H., Hu, B., Schuller, B. W., & Yamamoto, Y. (2023). Exploring interpretable representations for heart sound abnormality detection. Biomedical Signal Processing and Control, 82, 104569.