

Exploratory Five Class Heart Sound Classification Using a One Dimensional Convolutional Neural Network

Tathagata Mitra
Bengaluru, India

Abstract—Automated phonocardiogram classification requires evaluation that reflects class imbalance and recording heterogeneity. We evaluated a one-dimensional convolutional neural network (CNN) for five recording labels: artifact, extra heart sounds, murmur, normal, and extrasystole. The collection contained 585 recordings, comprising 124 previously held recordings and 461 labeled recordings from PASCAL Challenge Dataset B. The author reports clinic acquisition of the previously held collection; supporting acquisition records remain unavailable. Filename-derived groups and joint source–class stratification produced training, validation, and test partitions of 374, 94, and 117 recordings. Recordings were resampled to 2 kHz, processed using a fixed filtering pipeline, and represented by five-second windows. Training used inverse-frequency weighted cross-entropy, and validation macro F1 determined checkpoint selection. The selected model achieved 35.04% test accuracy and a macro F1 of 0.489. A majority-class baseline achieved 59.83% accuracy and a macro F1 of 0.150. Normal-class recall was 14.29%, with 38 of 70 normal recordings classified as extrasystole. These results identify substantial errors despite detection of several minority-class examples. Unverified patient independence, source–class confounding, and small minority-class test samples limit interpretation. The experiment provides a reproducible baseline and error analysis for further methodological comparison; it does not demonstrate readiness for clinical screening or cardiac risk prediction.

Index Terms—Class imbalance, convolutional neural network, error analysis, heart sound classification, phonocardiogram, reproducibility.

I. INTRODUCTION

Phonocardiograms (PCGs) represent acoustic recordings of cardiac activity and have supported research on automated heart-sound interpretation. Public collections have enabled evaluation across

different recording environments, while also exposing variation in data quality and label definitions [1]. A convolutional neural network (CNN) can learn representations from these recordings. A recording classifier can address a narrower task than clinical diagnosis: it predicts an annotated sound category rather than establishing a patient's underlying disease or future risk.

The distinction matters when several datasets are combined. PASCAL Dataset A uses normal, murmur, extra-heart-sound, and artifact labels, whereas Dataset B uses normal, murmur, and extrasystole labels [2]. Treating the two extra-sound categories as interchangeable would change the original labeling task. This study therefore retains them as separate outputs and asks how a waveform CNN behaves on the resulting five-class collection.

We report a single exploratory experiment with traceable recording-level partitions, checkpoint selection based on validation macro F1, and an explicit analysis of errors. The contribution is a documented baseline and an account of its limitations, rather than a new network architecture or a claim of superiority to published systems. The principal question is whether the selected classifier can distinguish the five supplied labels under a reproducible internal evaluation protocol. The experiment does not isolate the effect of adding Dataset B because a matched original-data-only comparison was not performed.

II. RELATED WORK

Liu et al. described the open heart-sound database supporting the PhysioNet/Computing in Cardiology 2016 challenge, providing a reference for evaluating algorithms on heterogeneous recordings [1]. Potes et al. combined a feature-based classifier with a CNN for normal/abnormal classification [3]. Their work

motivates comparison with both learned and engineered representations, but its binary task and evaluation protocol differ from those used here.

Zhang et al. investigated segmental CNNs for noisy recordings, with segments obtained through a heart-sound segmentation procedure [4]. Noman et al. studied waveform and time–frequency CNNs and combined their class probabilities [5]. These approaches motivate future comparisons of input representation and segmentation; neither was reproduced as a baseline in the present experiment.

The CirCor DigiScope study provides detailed murmur annotations and recordings from multiple auscultation locations [6]. More recent work has investigated attention-based models and explanation methods for PCG interpretation [7]. These studies suggest directions for richer annotation and model analysis. They do not establish that such methods would improve the current dataset without a controlled comparison. Published performance values are not ranked against the present results because class definitions, populations, partitions, and metrics differ. This section is a focused background review, not a systematic review.

Demir et al. studied convolutional representations for heart-sound classification, including spectrogram-derived deep features with a support vector machine [8]. Hasan and Karim investigated mel-frequency cepstral coefficient (MFCC) inputs with a lightweight CNN for five-class classification [9]. The latter study describes a 585-recording collection with the same aggregate class counts as the present collection. Matching counts do not establish audio identity or an equivalent evaluation protocol, and its reported performance is not treated as a directly comparable benchmark.

Karthikeya et al. compared engineered acoustic features and learned representations using XGBoost and neural classifiers [10]. This provides a rationale for an engineered-feature baseline alongside the current waveform model. Reviews by Partovi et al. [11] and Ren et al. [12] provide broader context for representation learning and evaluation in heart-sound analysis. Sharma's mobile-platform study [13] is a preprint relevant to implementation design; it is not evidence of clinical validation of the present system.

None of these models was reproduced in the reported experiment.

III. MATERIALS AND METHODS

A. Data sources and provenance

The original collection contained 124 WAV recordings: 40 artifact, 19 extra-heart-sound, 34 murmur, and 31 normal recordings. These were mono, 16-bit, 44.1 kHz files with durations of approximately 0.94–9.00 seconds. The author reports that the collection was recorded by local clinics; independently documented collection procedures and patient linkage were not available for this analysis. Its class counts exactly match the labeled PASCAL Dataset A counts [2]. This correspondence is a provenance concern rather than proof of file identity. The collection is therefore termed the *original collection*, without claiming prospective or independent clinical acquisition. Author-supplied provenance clarification. The author confirms that the original recordings were acquired by three local clinics in Bengaluru, Karnataka, India. This account is based on the author's report; supporting acquisition records and patient linkage were not available for this analysis. The author confirms that 124 original recordings were supplied and analyzed.

H03 Original collection. The confirmed clinic-recording count is 124.

H04 Collection period. According to the author, collection in Bengaluru, Karnataka, India lasted approximately 8–21 months. Exact start and end dates were unavailable; this range describes collection duration, not recording age.

H05 Acquisition metadata. Recording device and anatomical auscultation site information were unavailable.

An additional 461 labeled recordings were downloaded from the original PASCAL publisher's website [2] on 28 September 2026: 320 normal, 95 murmur, and 46 extrasystole recordings. Noisy subgroups were retained under their supplied normal or murmur labels. The normal set includes 120 recordings in its noisy subgroup. Original archive paths and source URLs were preserved in a manifest. Extrasystole remained a separate category from extra heart sounds. Table I reports the combined inventory and realized partition sizes.

Table I. Recording counts by source and partition. O denotes the original collection and B denotes PASCAL Dataset B. These are recording counts, not patient counts.

Class	O	B	Total	Train	Validation	Test
Artifact	40	0	40	26	6	8
Extra heart sounds	19	0	19	12	4	3
Murmur	34	95	129	82	21	26
Normal	31	320	351	225	56	70
Extrasystole	0	46	46	29	7	10
Total	124	461	585	374	94	117

B. Labels and ethical documentation

Folder labels defined the reference categories. No independent relabeling, diagnostic adjudication, or inter-rater agreement measurement was performed in the present run. These labels should not be interpreted as verified echocardiographic diagnoses or confirmed S3/S4 subtypes. Age, sex, unique patient counts, and clinical outcomes were not available for the original collection. Identifiable patient information is not required to complete this manuscript; unavailable descriptors should be stated as unavailable. H06 Labeling source. According to the custodian, doctors at the source clinics assigned the original collection's labels. Their specialties, number, and labeling criteria have not been documented in this draft.

H07 Label review. Review or adjudication details were unavailable.

H08 Ethics documentation. Formal ethics approval or exemption documentation was not available to the author; approval or exemption status could not be verified.

H09 Research-use permission. The author reports permission to use the recordings for research, excluding use of patient names, gender, or other patient details.

H09 Consent status. Only research-use permission was provided to the author. Patient consent status is unknown, and no consent waiver is asserted.

C. Integrity checks and partitioning

File checksums and decoded-audio hashes were computed, and all 585 recordings were included. No exact decoded-audio duplicates were detected by the implemented sample-rate-aware comparison. This procedure does not exclude resampled copies, near-duplicates, or repeated recordings of one individual.

The partition was created before optimization. Five-fold stratified group splitting with seed 42 supplied the first held-out fold as the test set. Another five-fold split, with seed 43, supplied the first validation fold from the remaining development recordings. Stratification used the joint source and class label. Only these selected folds were evaluated; this was not five-fold cross-validation.

For Dataset B, filenames were grouped by the first identifier after removing the local noise prefix. In the original collection, files sharing the same complete filename stem were grouped together. These rules kept matching filename-derived groups in one partition, but the groups were not verified patient identifiers. Thus, the test set was held out at the implemented group level, and patient independence cannot be asserted. The resulting split was 374 training, 94 validation, and 117 test recordings. All crops inherited their recording's partition.

D. Preprocessing and window construction

The audio channels were averaged to mono and resampled to 2 kHz using polyphase resampling. The existing preprocessing routine was applied to each full recording independently. It removed the mean, applied a 21-sample median filter, and combined zero-phase Butterworth band-pass outputs from 20–200 Hz and 200–500 Hz using weights 0.8 and 0.2. The respective filter design orders were six and four. Notch filters at 50 and 60 Hz used a quality factor of 30.

A Hilbert-envelope magnitude was median-smoothed over 11 samples and centered. The signal and centered envelope were combined with weights 0.7 and 0.3. Nonzero signals were divided by three times their root-mean-square amplitude, clipped to $[-1, 1]$, and smoothed with a five-sample, third-order Savitzky–Golay filter. This inherited pipeline was not optimized through an ablation study.

Training sampled one random five-second crop per recording per epoch; shorter recordings were padded with zeros on the right. Independently, gain scaling drawn uniformly from 0.8 to 1.2 and Gaussian noise with standard deviation 0.005 were each applied with probability 0.5. No pitch shifting or synthetic additional recordings were used. Evaluation used deterministic five-second windows and an end-anchored final window. The final window could overlap its predecessor. Softmax probabilities were averaged over windows to produce one prediction per recording; clips were not counted as independent test cases.

E. Network and optimization

The model used four one-dimensional convolution blocks with 32, 64, 128, and 256 output channels. Kernel sizes were 50, 25, 10, and 5, respectively, with stride two and padding 25, 12, 5, and 2. Each block used batch normalization and ReLU activation. The first three blocks used max pooling with kernel size four; the fourth used adaptive average pooling to one value per channel. Block dropout probabilities were 0.2, 0.2, 0.3, and 0.3. Fully connected layers mapped 256 features to 128, 64, and five outputs, with ReLU and dropout of 0.5 and 0.3 after the two hidden layers. The output layer returned logits.

Weights were initialized from scratch. The historical four-class checkpoint was not reused because its exact training partition was unavailable and might overlap the new evaluation recordings. Adam used an initial learning rate of 0.001, weight decay of 0.0001, and batch size 16. Cross-entropy class weights were

$$w_c = N_{\text{train}} / (5n_c) \quad (1)$$

where N_{train} is the total number of training recordings and n_c is the number of training recordings in class c . These weights were computed from training labels only.

The learning rate was halved after a validation macro-F1 plateau with scheduler patience five. The maximum duration was 60 epochs, with early stopping

Table II. Challenge count notation, adapted from the official scoring documentation. These are symbolic definitions, not results of this study.

Reference category	Normal	Uncertain	Abnormal
Normal, clean	Nn_1	Nq_1	Na_1
Normal, noisy	Nn_2	Nq_2	Na_2
Abnormal, clean	An_1	Aq_1	Aa_1
Abnormal, noisy	An_2	Aq_2	Aa_2

after 15 epochs without an improvement exceeding 10^{-8} in validation macro F1. Each improved checkpoint was serialized immediately. Python, NumPy, and PyTorch seeds were fixed to 42; deterministic PyTorch operations and four CPU threads were used. The recorded environment included Python 3.12.10 and PyTorch 2.14.0, with full package versions retained alongside the run. Reproduction artifacts include source snapshots, file hashes, exact split assignments, model weights, and predictions.

F. Evaluation

Validation macro F1 was the model-selection criterion. For each class, precision, recall, and F1 were calculated using one-versus-rest counts. Macro F1 was the unweighted mean of the five class F1 scores, with zero assigned to undefined values. Accuracy and the full confusion matrix were also reported. A baseline always predicted normal, the most frequent training class. The held-out test recordings were evaluated after checkpoint selection. No calibration analysis, significance test, repeated-seed analysis, or confidence interval estimation was performed. Source-specific accuracy was descriptive because the source subsets had different class distributions.

Specificity was subsequently calculated from the saved test confusion matrix as $TN/(TN+FP)$, where TN denotes true negatives and FP denotes false positives, for each class against all other classes. Recall is equivalent to sensitivity in this one-versus-rest definition. Adding specificity required no retraining or change to the selected checkpoint.

G. PhysioNet 2016 scoring framework for future evaluation

The PhysioNet/Computing in Cardiology 2016 framework permits normal (-1), uncertain (0), and abnormal (+1) outputs [14]. Table II defines its counts: uppercase letters indicate reference status, lowercase letters indicate output, and $k=1$ or $k=2$ identifies clean or noisy recordings, respectively.

Thus, Nn_k , Nq_k , and Na_k count normal-reference recordings; An_k , Aq_k , and Aa_k count abnormal-reference recordings. Weights wa_k and wn_k are the clean/noisy proportions within abnormal and normal reference classes in the complete challenge test set. Counts in the following equations concern the evaluated subset. Modified sensitivity and specificity are

$$Se = \frac{wa_1 Aa_1}{(Aa_1 + Aq_1 + An_1) + wa_2 (Aa_2 + Aq_2) / (Aa_2 + Aq_2 + An_2)} \quad (2)$$

$$Sp = \frac{wn_1 Nn_1}{(Na_1 + Nq_1 + Nn_1) + wn_2 (Nn_2 + Nq_2) / (Na_2 + Nq_2 + Nn_2)} \quad (3)$$

$$Score = (Se + Sp) / 2 \quad (4)$$

Uncertain predictions receive credit for noisy recordings but not clean recordings. This quality-aware score differs from our five-class macro F1 and one-versus-rest specificity.

No challenge score is calculated here: the present experiment lacks a validated normal/abnormal

reference mapping, complete clean/noisy annotations, and an uncertainty output. The artifact class is not automatically equivalent to uncertainty. Applying this framework would require a separate prespecified evaluation; it would not establish an official challenge result.

IV. RESULTS

Training ended after 35 epochs; epoch 20 supplied the selected checkpoint. Reloading that checkpoint reproduced the selected validation macro F1. Validation accuracy was 42.55% and macro F1 was 0.560. On the 117 test recordings, 41 predictions were correct, giving 35.04% accuracy and macro F1 of 0.489. The majority-class baseline had higher accuracy (59.83%) but lower macro F1 (0.150), as shown in Table III. Thus, the CNN performed better by macro F1 but worse by overall accuracy.

Table III. Recording-level performance. Model selection used validation macro F1.

Method	Partition	Accuracy	Macro F1
CNN	Validation (n=94)	42.55%	0.560
CNN	Test (n=117)	35.04%	0.489
Always normal	Test (n=117)	59.83%	0.150

Table IV presents class-level performance. Although all eight artifact and all three extra-heart-sound examples were identified, extra-heart-sound precision was only 37.50%, reflecting false-positive

assignments from other classes. Extrasystole precision was 11.32%, despite 60.00% recall. These results do not establish reliable recognition of either class in a new population.

Table IV. Test performance by class. Support is the number of actual recordings.

Class	Support	Precision	Recall	Specificity	F1
Artifact	8	100.00%	100.00%	100.00%	1.000
Extra heart sounds	3	37.50%	100.00%	95.61%	0.545
Murmur	26	42.42%	53.85%	79.12%	0.475
Normal	70	66.67%	14.29%	89.36%	0.235
Extrasystole	10	11.32%	60.00%	56.07%	0.190

The largest error involved normal recordings: only 10 of 70 were correctly identified. Of the remaining 60, 38 were classified as extrasystole, 18 as murmur, and four as extra heart sounds. The confusion matrix in Table V shows that these errors dominate the test

result. Accuracy was 17/24 (70.83%) in the original collection and 24/93 (25.81%) in Dataset B. This difference is descriptive and cannot be attributed solely to source effects because the label composition differs.

Table V. Test confusion matrix. Rows are actual labels and columns are predicted labels. EHS means extra heart sounds and ES means extrasystole.

Actual / predicted	Artifact	EHS	Murmur	Normal	ES
Artifact	8	0	0	0	0
EHS	0	3	0	0	0

Actual / predicted	Artifact	EHS	Murmur	Normal	ES
Murmur	0	1	14	2	9
Normal	0	4	18	10	38
ES	0	0	1	3	6

V. DISCUSSION

A. Interpretation of the principal findings

This experiment demonstrates the limits of the selected waveform pipeline on the combined collection. The higher macro F1 relative to the majority baseline indicates recognition beyond the dominant class, but the low normal recall and low extrasystole precision prevent a favorable assessment of overall discrimination. A class-weighted objective may contribute to the distribution of predictions; without an unweighted comparison, it cannot be identified as the cause of the errors.

Several additional explanations merit testing. The inherited filtering and envelope mixture may alter discriminative temporal or spectral information, and a five-second crop may provide insufficient context for intermittent events. Differences in acquisition conditions and annotation conventions may also contribute. These are hypotheses, not measured causal effects. The present results support examining these factors through controlled validation experiments rather than attributing errors to any one component.

B. Data provenance and generalization

Source and class are partly confounded: artifact and extra heart sounds occur only in the original collection, while extrasystole occurs only in Dataset B. A model could therefore exploit source characteristics correlated with labels. The mixed-source test set does not measure performance at a wholly unseen clinic or device. Furthermore, the original collection's correspondence with PASCAL Dataset A counts must be resolved before claiming a newly acquired clinical cohort. Public-data provenance and clinical provenance require different descriptions; neither can be inferred solely from the organization of a local folder.

The absence of verified patient identifiers also limits the independence of the evaluation. Keeping filename-derived groups intact reduces a specific form of overlap but does not demonstrate that different groups belong to different people. Unique recording counts

should therefore not be presented as patient counts. Missing age, sex, and diagnostic descriptors further prevent demographic subgroup analysis or interpretation as disease detection.

C. Relationship to the historical model

A historical four-class run recorded 76.92% accuracy on 13 test recordings. That value is not a comparator for the current experiment: the class set, recordings, partition, windowing, and model-selection workflow differ. The historical code also retained a shallow copy of model weights when attempting to preserve its best checkpoint. The present workflow serializes checkpoints immediately, but this correction does not retrospectively validate the earlier model-selection result. No conclusion about the benefit or harm of additional data can be drawn from the two accuracy values.

D. Limitations and planned extensions

The main limitations are the single evaluated split, small minority-class test samples, incompletely documented original-data provenance, unverified patient independence, and lack of diagnostic reference standards. Only a majority-class baseline was evaluated; established feature-based and alternative neural models remain untested. Exact duplicate screening cannot exclude near-duplicates. Model probabilities were not calibrated, and there was no prospective validation, external-site evaluation, or longitudinal outcome analysis.

Future work should compare class weighting, simpler filtering, longer windows, and waveform versus time–frequency models using a predefined development protocol. Feature-based baselines and repeated grouped evaluation would help assess whether any improvement is robust. A new independent evaluation set or an appropriately nested protocol is needed for subsequent confirmatory claims, because the current test results are now available to the investigators. Clinical expert review could clarify mislabeled or ambiguous recordings where permitted. These

extensions have not yet been completed and their outcomes are not part of the present results.

In particular, planned comparisons include MFCC and log-mel inputs to a two-dimensional CNN, an XGBoost model using engineered acoustic features, and augmentation enabled versus disabled. All comparisons should use compatible label definitions and the same development partitions. The literature motivates these experiments but does not establish which representation will perform best on the present collection.

VI. CONCLUSION

A five-class CNN evaluated on 585 heart-sound recordings achieved test macro F1 of 0.489 and accuracy of 35.04%. Its recognition of minority-class examples was offset by substantial confusion of normal recordings with extrasystole and murmur. The documented pipeline provides a reproducible exploratory baseline, while the observed errors and data limitations preclude claims of reliable screening, diagnosis, or early cardiac risk stratification. Subsequent work should prioritize provenance, independent evaluation, and controlled methodological comparisons.

DECLARATIONS

Ethics and consent. The author reports that the clinic recordings were supplied with permission for research use, subject to exclusion of patient names, gender, and other patient details. Formal ethics approval or exemption documentation was not available to the author. Only research-use permission was provided; patient consent status is unknown, and no consent waiver is asserted. The reported research-use permission is not presented as evidence of either determination.

H10 Statement status. The author-information fields are completed from the available information. Ethics documentation remains unavailable and patient consent status remains unknown; completion of these statements does not establish approval or consent.

Data availability. PASCAL Dataset B is available through the challenge publisher [2]. The original clinic recordings are subject to the research-use restrictions described below.

H11 Access restrictions. The reported permission covers research use and excludes patient names,

gender, and other patient details. Permission to redistribute the audio has not been established; no public release or access-on-request commitment is made.

Code availability. Training code, configuration, split manifests, package versions, checkpoints, and recording-level predictions were retained locally for the reported experiment.

H12 Code release. A public code repository is not currently available while the research is ongoing.

H13 Funding. This research received no external funding.

H14 Competing interests and motivation. The author reports no financial competing interests. The author describes a personal motivation to develop machine-learning methods that may ultimately contribute to reducing sudden cardiac risk for public benefit; no other competing interests are reported. This motivation does not constitute evidence that the present model predicts or prevents sudden cardiac events.

H15 Author contributions. Tathagata Mitra obtained the recordings, initiated the research, worked with the original model, reviewed model results, and wrote the research paper. AI assistance is described separately below.

AI assistance. An OpenAI assistant assisted with code preparation, execution of the exploratory training workflow, literature retrieval, and manuscript drafting. The authors must verify the analysis, references, and final text and take responsibility for the submitted work.

H17 Additional AI assistance. The author reports no AI assistance beyond that described above. The disclosure will be adapted to the selected venue's requirements once a venue is chosen.

APPENDIX: AUTHOR COMPLETION GUIDE

H01 Author: Tathagata Mitra.

H02 Location and contact: Bengaluru, India;

H02 Affiliation status: Institutional affiliation information is unavailable.

H02 Corresponding author: Tathagata Mitra; *Draft status:* Author-information fields have been completed, with unavailable information stated explicitly. The current results describe an exploratory experiment; the manuscript does not establish clinical utility. Remove this guide and the completion appendix before submission.

This appendix is an editing aid and should be removed before submission. The numeric results already entered are from the completed five-class run. Missing metadata should be filled from records, or explicitly declared unavailable. Do not create patient identities, approvals, labels, or results to complete a blank.

H01–H02 Authors

Tathagata Mitra, Bengaluru, India, and the supplied email have been entered. Institutional affiliation information is unavailable. Tathagata Mitra is the corresponding author at the supplied email.

H03–H05 Provenance

Clinic acquisition in Bengaluru, Karnataka, India is author-confirmed. The confirmed original count is 124 recordings. Collection lasted approximately 8–21 months according to the author. Exact dates, device, and anatomical sites are unavailable.

H06–H07 Annotation

Clinic doctors assigned the labels. Review or adjudication details were unavailable. Annotator specialties, number, and labeling criteria remain undocumented.

H08–H10 Ethics and consent

Research-use permission and restrictions on patient details are recorded. Ethics approval or exemption documentation is unavailable to the author; patient consent status is unknown. These limitations remain despite completion of the draft fields.

H11–H12 Availability

Research-use restrictions are stated; audio redistribution permission has not been established. A public code repository is not currently available.

H13–H17 Declarations

No external funding, financial competing interests, or additional AI assistance are reported. Personal research motivation is disclosed, with no other competing interests reported. Author contributions are recorded: obtaining recordings, initiating research, working with the original model, reviewing results, and writing the paper. Acknowledgments have been omitted at the author's request.

Research that remains unperformed. Controlled original-only versus combined-data training; alternative model baselines; preprocessing and weighting ablations; repeated-seed or nested grouped evaluation; uncertainty estimation; expert error review; and external validation. These are described as future work, not left as empty results that might imply the experiments already occurred.

Submission preparation. This draft follows the supplied IJIRT template. Before submission, confirm the current venue requirements, verify each reference against the cited source, and have all authors review the conclusions. This manuscript is a starting draft for an exploratory study rather than a submission-ready clinical validation paper.

REFERENCES

- [1] C. Liu et al. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12), 2181–2213, 2016. <https://doi.org/10.1088/0967-3334/37/12/2181>.
- [2] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. Dataset and challenge documentation. <https://www.peterjbentley.com/heartchallenge/>. Accessed 28 September 2026.
- [3] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. *Computing in Cardiology*, vol. 43, 2016. <https://doi.org/10.22489/CinC.2016.182-399>.
- [4] Y. Zhang, S. Ayyar, L.-H. Chen, and E. J. Li. Segmental convolutional neural networks for detection of cardiac abnormality with noisy heart sound recordings. *arXiv preprint*, 1612.01943, 2016. <https://doi.org/10.48550/arXiv.1612.01943>.
- [5] F. Noman, C.-M. Ting, S.-H. Salleh, and H. Ombao. Short-segment heart sound classification using an ensemble of deep convolutional neural networks. *IEEE ICASSP*, 2019. <https://doi.org/10.1109/ICASSP.2019.8682668>. Author manuscript: <https://arxiv.org/abs/1810.11573>.
- [6] J. Oliveira et al. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6), 2524–2535, 2022. <https://doi.org/10.1109/JBHI.2021.3137048>.
- [7] B. Althaph and N. P. Challa. Explainable attention-based deep learning for classification and interpretation of heart murmurs using phonocardiograms. *Scientific Reports*, 15, 37991,

2025. <https://doi.org/10.1038/s41598-025-21971-x>.
- [8] F. Demir, A. Şengür, V. Bajaj, and K. Polat. Towards the classification of heart sounds based on convolutional deep neural network. *Health Information Science and Systems*, 7, 16, 2019. <https://doi.org/10.1007/s13755-019-0078-0>.
- [9] A. Hasan and M. Karim. Robust Heart Sound Analysis With MFCC and Light Weight Convolutional Neural Network. *IEEE Open Journal of Engineering in Medicine and Biology*, 6, 549–556, 2025. <https://doi.org/10.1109/OJEMB.2025.3615395>.
- [10] P. P. S. Karthikeya et al. Heart Sound Classification for Early Detection of Cardiovascular Diseases Using XGBoost and Engineered Acoustic Features. *Sensors*, 26(2), 630, 2026. <https://doi.org/10.3390/s26020630>.
- [11] E. Partovi, A. Babic, and A. Gharehbaghi. A review on deep learning methods for heart sound signal analysis. *Frontiers in Artificial Intelligence*, 7, 1434022, 2024. <https://doi.org/10.3389/frai.2024.1434022>.
- [12] Z. Ren, Y. Chang, T. T. Nguyen, Y. Tan, K. Qian, and B. W. Schuller. A Comprehensive Survey on Heart Sound Analysis in the Deep Learning Era. *IEEE Computational Intelligence Magazine*, 19(3), 42–57, 2024. <https://doi.org/10.1109/MCI.2024.3401309>.
- [13] P. Sharma. Building a Mobile Platform For CNN Based Heart Murmur Identification. *TechRxiv preprint*, version 2, 2023. Not peer reviewed. <https://doi.org/10.36227/techrxiv.23716155.v2>.
- [14] C. Liu et al. Classification of Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016, version 1.0.0. PhysioNet, 2016, Scoring section. <https://physionet.org/content/challenge-2016/1.0.0/>. Accessed 28 September 2026.