This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Preprocessing as Prior: Rethinking Self-Supervised Representation Learning in Cardiac Sound Classification

Mona Aman, Godbright Nixon Uiso, Brandone Fonya, John Bosco Gachomba Thuo, Maurine Wanjiku Gatimu, and Carine Mukamakuza

Carnegie Mellon University Africa, Kigali, Rwanda
{amona, guisso, bfonya, jthuo, mgatimu, cmukamak}@andrew.cmu.edu

Abstract. Self-supervised learning (SSL) is promoted as a way to reduce dependence on labeled data in medical machine learning, yet reported gains are rarely attributed to a specific pipeline component; in phonocardiogram (PCG) classification, preprocessing priors, learned representations, and task difficulty are seldom isolated. We audit when SSL adds measurable value with a controlled 2×2 factorial design crossing physiologically grounded Butterworth filterbank inputs against raw waveforms, and SSL pre-training against a randomly initialized frozen encoder, all sharing a matched Mamba backbone and evaluated with frozen linear probes. With the filterbank in place, SSL exceeds a seeded random frozen encoder on PhysioNet 2016 by only a small margin (AUROC ≈ 0.930 vs. 0.903; +0.027, disjoint bootstrap CIs), and low-label behaviour is similarly close. Removing the filterbank widens the SSL advantage on raw waveforms (+0.053 AUROC), and on a harder murmur grading task filterbank SSL models exceed the random baseline (n=34 patients, exploratory). We further test whether physiology-informed SSL confers an additional advantage using CycleSSL, an objective that pairs distinct cardiac cycles from the same patient and site; CycleSSL performs comparably to generic SimCLR rather than exceeding it. We propose random-encoder controls, preprocessing ablations, and difficulty-stratified evaluation as a general attribution framework for clinical SSL claims, with direct implications for deployment in low-resource screening settings.

Keywords: Self-supervised learning · Phonocardiogram · Attribution · Cardiac auscultation · State-space models.

1 Introduction

Cardiac auscultation remains a frontline tool for cardiovascular screening, particularly where echocardiography is scarce [9]. Phonocardiograms (PCG) digitize this signal and offer a low-cost path to scalable screening for valvular and structural heart disease, but routine use is limited by the need for expert interpretation and by the small size of reliably labeled corpora [10, 2]. Self-supervised learning (SSL) is an attractive response: it promises to convert unlabeled recordings

into transferable representations, reducing the labeled-data burden on downstream tasks.

The implicit assumption in much of this literature is that gains on downstream benchmarks reflect what the SSL encoder *learned* [12]. In practice, end-to-end PCG pipelines combine several strong inductive biases: bandpass or filterbank preprocessing tuned to known acoustic structure, segmentation into beats or windows, and architectural choices that already encode useful invariances [10]. Each can carry a meaningful share of the discriminative signal (indeed, even untrained networks can act as strong priors [11]), so headline numbers on coarse tasks such as normal-vs-abnormal classification do not, on their own, indicate whether SSL or the preprocessing prior is doing the work. If preprocessing dominates, reported SSL value may be overestimated when pipelines are ported without it; if instead SSL contributes most when preprocessing is weak or tasks are fine-grained, easy benchmarks *underestimate* it. These distinctions are not academic for African and other low-resource screening programmes, the motivating setting for this work: there the preprocessing priors that inflate benchmark scores are often unavailable at deployment, so whether a gain comes from SSL or from a handcrafted prior determines whether it survives the move from curated data to the clinic.

We therefore ask a more precise question: when does SSL add measurable value, and when are gains instead explained by preprocessing or task difficulty? We answer this with a controlled audit (Fig. 1) disentangling three drivers of downstream performance: preprocessing priors (filterbank vs. raw waveform), representation learning (SSL vs. random frozen encoder), and task difficulty (coarse vs. murmur grading). All conditions share a matched Mamba backbone [1] with frozen linear probes, isolating representation quality from fine-tuning effects. A natural hypothesis is that self-supervised objectives informed by domain structure should outperform generic augmentation-based pairing. To test this directly, we introduce CycleSSL, a contrastive objective that constructs positive pairs from distinct cardiac cycles of the same patient and auscultation site, using expert per-beat segmentation in CirCor DigiScope [2]. CycleSSL allows us to ask whether physiologically grounded positive-pair design confers a measurable advantage in this setting. We find that with the filterbank in place SSL gives only a small gain over a seeded random frozen encoder (+0.027 AUROC, disjoint CIs); removing preprocessing widens the gain on raw waveforms (+0.053); and on the harder CirCor grading task (n=34, exploratory) filterbank SSL beats the random baseline, while CycleSSL remains comparable to SimCLR throughout.

Contributions. (1) A controlled 2×2 factorial audit isolating preprocessing from SSL representation learning. (2) A random-encoder baseline and handcrafted-feature floor establishing an attribution ladder. (3) The conditional-benefit finding: the SSL gain over a random encoder is small under a strong preprocessing prior and grows as the prior is weakened or the task is harder.

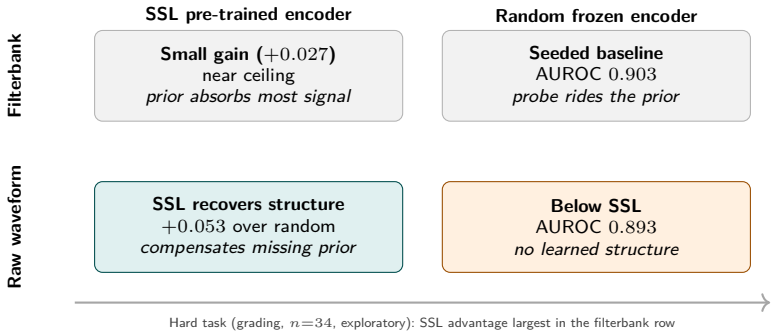

Fig. 1. The audit design. Rows vary the preprocessing prior; columns vary SSL vs. randomly frozen encoder, with a shared Mamba backbone and linear probe. The SSL gain is small under the filterbank, larger without it, and largest on the harder grading task.

2 Related Work

Self-supervised contrastive learning. Contrastive SSL pulls together augmented views of an input and pushes apart other samples, typically via the NT-Xent loss of SimCLR [3]; BYOL [4] and MoCo [5] reduce reliance on explicit negatives. Generic augmentations (cropping, masking, noise) give no guarantee that the induced invariances align with clinically meaningful physiological variation.

SSL for cardiac signals. In ECG, CLOCS [6] extends contrastive learning across temporal segments, leads, and patients, but still relies on segment-level or augmentation-driven correspondence rather than physiologically defined units. Subsequent ECG SSL work explores 12-lead and lead-agnostic pre-training [13, 14], but likewise defines correspondence at the segment or lead level. For PCG, SSL is comparatively underexplored: Listen2YourHeart [7] studies contrastive learning for heart sound analysis and shows strong sensitivity to augmentation design, but constructs positives from generic crops rather than the cardiac cycle.

The missing axis: attributing SSL gains. Largely absent is a controlled attribution of *why* reported PCG SSL pipelines work; random-encoder baselines, raw-input ablations, and difficulty-stratified evaluation are rarely reported together, in contrast to a broader vision literature that scrutinises what SSL representations actually capture [12]. We fill this gap with an audit that holds architecture and probe fixed, varies preprocessing and encoder training factorially, and stresses models on tasks of varying difficulty.

3 Methods

3.1 Audit design

The 2×2 factorial in Fig. 1 crosses preprocessing (filterbank vs. raw waveform) with encoder training (SSL vs. randomly initialized frozen encoder), holding the Mamba backbone fixed in all four cells so differences reflect preprocessing, training, or their interaction rather than architecture. Three axes layer on top: objective (CycleSSL vs. SimCLR, ±phase), architecture (Mamba vs. Transformer), and task difficulty (PhysioNet binary and Yaseen 5-class as easier; CirCor murmur grading as harder).

3.2 Signal preprocessing

All recordings are processed at 4000 Hz (CirCor native; PhysioNet upsampled from 2000 Hz; Yaseen downsampled from 8000 Hz), divided into 1.5 s sliding windows (6,000 samples, 50% hop), and z-score normalized per window. The *filterbank* condition applies a three-band fourth-order Butterworth design aligned with known heart-sound acoustics: Band 0 (10–200 Hz, S1/S2 fundamentals), Band 1 (200–450 Hz, murmur body), Band 2 (450–700 Hz, high-frequency murmur texture), stacked into $(T, 3)$; when phase annotations are available, a fourth channel encodes cardiac phase (unknown, S1, systole, S2, diastole $\rightarrow \{0, 0.25, 0.5, 0.75, 1.0\}$), giving $(T, 4)$. The *raw* condition bypasses the filterbank entirely, feeding the z-scored waveform with a zero phase channel as $(T, 2)$, isolating whether the encoder can recover useful frequency structure directly.

3.3 Encoder architectures

We use the Mamba state-space model [1] as primary encoder: four Mamba blocks (d_{model}=128) with residual connections, global mean pooling, and a two-layer projection head Linear(128→256) → GELU → Linear(256→128) → ℓ_2-norm (∼3.2M parameters; $\mathcal{O}(T)$ selective scan suits long recordings and near-real-time screening). Input channels are 4 (filterbank) or 2 (raw). The **random-encoder baseline** uses the identical architecture at random initialization with no training [11]; the gap between trained encoders and this baseline measures what representation learning contributes in each cell. We also evaluate a Transformer encoder (Conv1d patch embedding, kernel/stride 8, sinusoidal positions, 4 pre-norm layers, d_{model}=128, n_{head}=4, FFN dim 512) to check architecture-specificity; it matches Mamba on easier tasks, with Mamba retained as the primary encoder for deployment.

3.4 Self-supervised objectives

SimCLR [3] serves as the generic baseline. **CycleSSL** is the domain-informed test objective: using CirCor beat segmentation, each batch samples anchor–positive pairs from distinct cardiac cycles of the same patient and auscultation

site, with negatives from other patients at the same site. This isolates physiology-grounded pairing from generic augmentation while controlling for site. "+phase" variants add a small auxiliary phase-decoding loss ($0.2 \cdot \mathcal{L}_{\text{phase}}$). CycleSSL is thus a controlled hypothesis test for domain-informed SSL, not a headline method.

3.5 Augmentation and training

Views use Gaussian noise, time masking, amplitude scaling, and frequency-band dropout. All SSL models are pre-trained on the CirCor training split (90% of 941 patients) without downstream labels, using AdamW (lr 10^{-4}, batch size 128, τ=0.2, 100 epochs, seed 42).

4 Cohort Selection and Tasks

We evaluate on three benchmarks: **PhysioNet/CinC 2016** (binary, 3,240 recordings; 2,575 normal, 665 abnormal; stratified 80/20); **Yaseen** (5-class, 990 recordings: Normal, AS, MR, MS, MVP; stratified 80/20); and a **CirCor harder task**, 3-class murmur grading (Levine scale) on a held-out 20% patient split with a 34-patient test cohort, treated as exploratory throughout. PhysioNet and Yaseen lack patient identifiers, so recording- and patient-level splits coincide.

5 Results

5.1 Evaluation protocol

Encoder parameters are frozen after pre-training; a `StandardScaler` and logistic regression probe are fit on training embeddings only, with recording-level predictions from averaged window posteriors. The primary metric is recording-level one-vs-rest macro AUROC. Frozen probing is used for attribution: unlike fine-tuning, it keeps preprocessing, encoder geometry, and supervised adaptation separated. Fully supervised PhysioNet 2016 systems reached AUROC 0.86–0.94 [8]; our linear-probe results fall within this range.

5.2 Main result: an attribution ladder

A **handcrafted band-energy baseline** (per-window mean energy in each Butterworth band, three scalars, same logistic probe) reaches AUROC 0.828 on PhysioNet with no temporal modeling, confirming the filterbank alone is a strong prior. Adding a *random* Mamba encoder raises this to 0.903 (seeded): the temporal dynamics of energy across S1, systole, S2, and diastole carry major additional value even without training. SSL pre-training reaches 0.924–0.931 (Table 1), a small but measurable improvement over the random encoder (disjoint CIs; see below), since the prior and a generic sequential projection already solve most of the coarse task.

The full 2×2 interaction reinforces this: removing the filterbank leaves the random encoder near 0.893 but raises the SSL encoder to 0.946, so the SSL-over-random gain grows from +0.027 (filterbank) to +0.053 (raw): the objective is largely, though not entirely, redundant with a strong input pipeline on easy benchmarks. Bootstrapped 95% CIs (2,000 resamples) give CycleSSL+phase AUROC 0.930 [0.929, 0.961] and the seeded random filterbank encoder 0.903 [0.878, 0.928] (mean over seeds $\{0, 1, 2, 42\}$, ± 0.004, range 0.897–0.909); an earlier unseeded random run reached 0.935, above its own bootstrap interval and not reproduced across seeds. The two intervals are disjoint, so the filterbank-condition SSL gain, though small, is measurable rather than null. ECE is also lower for SSL than random (0.108 vs. 0.164), indicating better-calibrated probabilities.

Table 1. Central 2×2 audit plus a handcrafted reference. Filterbank+SSL is representative (SimCLR+phase); across objectives and both architectures, filterbank SSL spans 0.924–0.931 (PhysioNet) and 0.982–0.994 (Yaseen). The filterbank random-encoder value is the seeded mean over $\{0, 1, 2, 42\}$.

Condition	**PhysioNet**	**Yaseen**
Handcrafted band-energy features	0.828	–
Filterbank + SSL	0.931	0.987
Filterbank + Random encoder	0.903	0.987
Raw waveform + SSL (SimCLR)	**0.946**	1.000
Raw waveform + Random encoder	0.893	0.964

5.3 All models in the easy regime

Within the filterbank setting, all eight SSL variants (CycleSSL/SimCLR $\times$ phase, Mamba/Transformer) fall in a narrow band, 0.924–0.931 on PhysioNet and 0.982–0.994 on Yaseen: neither objective nor architecture is the primary driver, and CycleSSL matches SimCLR across tasks and inputs, so domain-informed positive-pair construction confers no measurable advantage over generic pairing here. A randomly initialized Transformer reaches 0.859 [0.831, 0.887] with the filterbank, above the handcrafted baseline and near-ceiling on Yaseen (0.993), confirming the pattern is not Mamba-specific.

5.4 Label efficiency in the easy regime

Varying the PhysioNet training-label fraction over $\{1, 2, 5, 10, 20, 50, 100\}\%$ and re-fitting the probe, filterbank+SSL and filterbank+random label-efficiency curves track each other closely (both ≈ 0.73 at 1%), with SSL modestly ahead at higher fractions in line with the small filterbank-condition gap. When the prior already makes the task nearly linearly separable, the probe extracts most of the usable signal regardless of SSL pre-training, so the large low-label SSL advantage expected on raw inputs is muted here.

5.5 Harder task: an exploratory SSL benefit on murmur grading

On the harder murmur grading task (Table 2), all encoders degrade relative to coarse classification, and the random encoder degrades most. This suggests that when the preprocessing prior is insufficient for the task, learned representations begin to matter. We note the small test cohort (n=34) and treat this result as exploratory. The grading cohort shares recording conditions with the CirCor pretraining corpus, while PhysioNet/Yaseen are out-of-distribution transfer, making the coarse-task result the more conservative test of the framework.

Table 2. Harder CirCor grading task (exploratory, n=34 test patients).

Model	**Grading**
CycleSSL + phase (Mamba, FB)	0.873
CycleSSL – phase (Mamba, FB)	0.868
SimCLR – phase (Mamba, FB)	0.839
SimCLR + phase (Mamba, FB)	0.818
Transformer – SimCLR (FB)	0.804
Random encoder (Mamba, FB)	0.802
Random encoder (Mamba, raw)	0.802
SimCLR (Mamba, raw)	0.780

6 Discussion

SSL's value on a downstream clinical task is not a property of the objective alone, but of the joint configuration of preprocessing, encoder, probe, and task difficulty. The unparameterized Butterworth filterbank carries a large share of the discriminative signal (even a randomly initialized encoder rides it to near-ceiling AUROC), so SSL adds only a small further increment, and comparisons among SSL methods on such benchmarks have little dynamic range. Removing the prior (raw streaming audio, different sensors) widens the SSL advantage, and difficulty-stratified evaluation is what makes representation differences visible, consistent with our grading results (small, n=34, exploratory). Under end-to-end fine-tuning on the coarse task, SSL and random encoders converge (0.960 vs. 0.956), consistent with the frozen-probe finding. Injecting physiological structure into the objective (CycleSSL) does not unlock gains that generic objectives miss; within the regime we test, the preprocessing prior dominates both.

Implications for low-resource and African deployment. Attribution matters most where deployment is hardest. In many African and other low-resource screening settings the conditions that make our filterbank prior effective (clean recordings, a fixed sensor, expert per-beat segmentation for the phase channel) cannot be assumed, and labeled corpora are even scarcer, so gains credited to SSL on

curated benchmarks may partly reflect a handcrafted prior that will not transfer intact to the field. Encouragingly, the same analysis shows SSL contributes most exactly when the prior is weak or absent (raw waveforms) and tasks are harder (grading), the realistic low-resource regime, so representation learning may matter more, not less, once pipelines leave the benchmark. Auditing which component carries the signal before deployment is thus a practical prerequisite for trustworthy screening in these settings.

Practical guidance. Before claiming SSL gains: (1) test random-encoder baselines under the same probe and pipeline; (2) ablate preprocessing priors; (3) evaluate across task difficulty, since near-ceiling nulls say little; (4) report uncertainty, especially on small cohorts.

7 Limitations

The grading test set has only 34 patients, so that result is exploratory and needs replication. Downstream evaluation uses frozen linear probes (an attribution-focused choice that does not characterize end-to-end fine-tuning), and SSL pretraining uses a single corpus (CirCor), leaving cross-corpus and cross-sensor generalization open; a single 3-band Butterworth filterbank is used, and alternatives may shift the preprocessing/representation balance. These caveats qualify but do not undermine conclusions that hold across objectives, architectures, and label fractions.

8 Conclusion

We presented a controlled audit of SSL for cardiac sound classification that disentangles preprocessing, representation learning, and task difficulty. Strong handcrafted preprocessing attenuates SSL gains on coarse benchmarks, while SSL becomes more useful once preprocessing is removed or the task grows harder (exploratory grading evidence), and physiology-informed CycleSSL matches generic SimCLR rather than exceeding it. Beyond PCG, the framework (random-encoder controls, preprocessing ablations, and difficulty-stratified evaluation) attributes downstream gains to the right pipeline component, a prerequisite for trustworthy deployment in low-resource settings.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752 (2023)
2. Oliveira, J., et al.: The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE J. Biomed. Health Inform. **26**(6), 2524–2535 (2022)

3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
4. Grill, J.B., et al.: Bootstrap your own latent: a new approach to self-supervised learning. In: NeurIPS. pp. 21271–21284 (2020)
5. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
6. Kiyasseh, D., Zhu, T., Clifton, D.A.: CLOCS: contrastive learning of cardiac signals across space, time, and patients. In: ICML. pp. 5606–5615 (2021)
7. Ballas, A., Papapanagiotou, V., Delopoulos, A., Diou, C.: Listen2YourHeart: a self-supervised approach for detecting murmur in heart-beat sounds. arXiv:2208.14845 (2022)
8. Clifford, G.D., et al.: Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016. In: Computing in Cardiology. pp. 609–612 (2016)
9. Marijon, E., Ou, P., Celermajer, D.S., et al.: Prevalence of rheumatic heart disease detected by echocardiographic screening. N. Engl. J. Med. **357**(5), 470–476 (2007)
10. Springer, D.B., Tarassenko, L., Clifford, G.D.: Logistic regression-HSMM-based heart sound segmentation. IEEE Trans. Biomed. Eng. **63**(4), 822–832 (2016)
11. Saxe, A.M., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: On random weights and unsupervised feature learning. In: ICML (2011)
12. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: CVPR (2019)
13. Mehari, T., Strodthoff, N.: Self-supervised representation learning from 12-lead ECG data. Comput. Biol. Med. **141**, 105114 (2022)
14. Oh, J., Chung, H., Kwon, J., Hong, D., Choi, E.: Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In: Conf. on Health, Inference, and Learning (CHIL). PMLR (2022)