# Generative Cleaning Pipelines for Consumer Sensor Cardiology: From Artifact Removal to Clinical-Grade Screening

**Ziyang Liu[1]，Mengqi Li[2], ***

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R.China
[2].College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 201203, P.R.China
* Corresponding author: limengqi@fudan.edu.cn

## Abstract

Background: Consumer wearables now record photoplethysmography (PPG) almost continuously, opening a route to population-scale cardiac screening; yet signals acquired outside the clinic are routinely corrupted by motion, ambient light, and poor sensor contact, and a large share of recorded data is discarded before any analysis. Objective: This review examines how generative machine-learning models—particularly conditional adversarial and related deep networks—are reshaping the cleaning of wearable cardiac signals, and how that cleaning can be connected, end to end, to clinically meaningful screening decisions. Methods: We synthesize evidence across four layers of the processing chain: signal-quality assessment, generative denoising and reconstruction, downstream cardiac inference, and trustworthiness evaluation, drawing on recent methodological and clinical studies. Results: The literature indicates that learned reconstruction can recover diagnostic morphology that classical filtering removes, that downstream task performance is the most informative measure of cleaning quality, and that uncertainty-aware gating markedly lowers the cost of acting on unreliable outputs. However, generative models can also fabricate plausible but spurious waveform features, making calibrated confidence estimates indispensable before clinical deployment. Conclusion: Treating artifact removal and screening as a single, uncertainty-aware pipeline—rather than as isolated steps—offers a credible path from consumer-grade sensing to clinical-grade decision support, provided that evaluation, calibration, and prospective validation keep pace with model capability.

**Keywords:** Photoplethysmography; wearable sensors; generative adversarial networks; signal denoising; atrial fibrillation screening; uncertainty quantification; deep learning; trustworthy AI

**Article History**

Received: October 14, 2022
Revised: December 26, 2022
Accepted: February 20, 2023
Available Online: March 30, 2023

## 1. Introduction

Wearable optical sensors have quietly become one of the most widely deployed measurement instruments in everyday life. Wrist-worn watches, rings, and patches sample photoplethysmography (PPG)—the small changes in transmitted or reflected light caused by pulsatile blood volume—for hours at a time, often during sleep, exercise, and ordinary activity. The clinical promise of this data stream is considerable: cardiac rhythm disturbances such as atrial fibrillation are frequently intermittent and

asymptomatic, and a sensor that a person already wears can, in principle, observe rhythm over weeks rather than the minutes available during a clinic visit (Bayoumy et al., 2021; Charlton et al., 2023). Large prospective studies have shown that consumer devices can flag irregular rhythms at scale and route users toward confirmatory evaluation (Perez et al., 2019; Guo et al., 2019). The vision of continuous, decentralized cardiac screening is therefore no longer hypothetical; it is partially realized in products that millions of people carry.

Between that promise and dependable practice, however, lies a stubborn obstacle: signal quality. A PPG waveform recorded on a moving wrist in daylight bears little resemblance to the clean, well-perfused traces obtained under laboratory conditions. Motion displaces the sensor relative to the skin, ambient light leaks into the photodetector, and low peripheral perfusion flattens the pulse, so that the morphological features an algorithm relies upon are distorted or buried in noise (Elgendi, 2016; Dao et al., 2017). The practical consequence is severe data attrition. Quality-control front ends commonly discard a substantial fraction of recorded segments before any inference is attempted, and the discarded portions are not random: they concentrate during exactly the active periods when continuous monitoring would be most valuable (Goh et al., 2020; Pereira et al., 2019). A screening system that can only analyze the calmest, best-perfused minutes of the day is a far weaker instrument than its raw data volume suggests.

The conventional response to noise has been to filter or to reject. Band-pass and adaptive filters suppress out-of-band interference, while signal-quality indices gate the stream so that only trustworthy segments proceed (Elgendi, 2016; Shin, 2022). Both strategies are valuable, but both are fundamentally subtractive: filtering can erase diagnostic morphology along with the noise, and rejection simply throws information away. An alternative has matured rapidly over the past several years. Rather than removing corruption, generative deep-learning models attempt to reconstruct what the clean signal most plausibly was, learning the statistical regularities of uncorrupted physiology and using them to repair damaged segments (Singh & Pradhan, 2021; Wang et al., 2022). This reframing—from removal to restoration—is the conceptual pivot of the present review.

Generative restoration is powerful precisely because it is creative, and that same creativity is its principal hazard. A model trained to produce realistic waveforms can, when confronted with an unfamiliar input, synthesize features that look entirely plausible yet correspond to nothing in the underlying physiology (Brophy et al., 2023; Yi et al., 2019). For a screening application, a convincingly hallucinated arrhythmia or a fabricated return to normal rhythm is not a cosmetic flaw but a potential clinical error. Cleaning a wearable cardiac signal therefore cannot be evaluated in isolation from the decision the cleaned signal will inform. The quality that matters is not visual realism but whether the reconstructed segment leads the downstream model to the correct conclusion, and whether the system can recognize when it is likely to be wrong (Lambert et al., 2024; Abdar et al., 2021).

A note on terminology frames what follows. By a cleaning pipeline we mean the full chain that turns a raw wearable recording into a clinical decision, not the denoiser alone; by generative we mean models that learn a distribution over clean signals and sample from it to repair inputs, in contrast to discriminative filters that merely suppress components deemed noise; and by clinical-grade screening we mean output trustworthy enough to influence triage, referral, or monitoring, which is a far higher bar than benchmark accuracy. Keeping these distinctions in view matters because much of the confusion in the field arises from conflating a good reconstruction with a good decision, or a high average accuracy with reliability on the individual segment that a given patient actually contributes (Lambert et al., 2024; Vasey et al., 2022).

The timing of this synthesis is not accidental, for two trajectories have converged. Consumer wearables have reached a scale and sensing quality at which population-level cardiac observation is genuinely feasible, while generative modelling has matured from a research curiosity into a dependable engineering tool, with stable training recipes and a growing evidence base in biomedical signal processing (Brophy et al., 2023; Yi et al., 2019). The opportunity created by their intersection is to stop treating the noisy, discarded majority of wearable data as waste and to start treating it as recoverable signal. Realizing that opportunity responsibly, however, requires importing the hard-won lessons of trustworthy clinical artificial intelligence into a setting—unsupervised, ambulatory, and consumer-grade—for which those lessons were not originally written (Topol, 2019; Rajpurkar et al., 2022).

This review develops that argument in stages. We first set out the clinical context and the anatomy of the signal-quality problem, then organize the landscape of artifacts and the families of methods used to address them. We next examine generative cleaning pipelines in detail, treating quality gating, reconstruction, downstream inference, and uncertainty estimation as interlocking components rather than separate research topics. Finally, we turn to trustworthiness and translation, asking what evidence a generative cleaning pipeline must furnish before it can responsibly contribute to clinical-grade screening. Throughout, our claim is that artifact removal and screening should be designed, evaluated, and governed as a single uncertainty-aware pipeline, because the value of cleaning is realized only at the point of decision.

## 2. Clinical Context: Wearable Photoplethysmography and the Signal-Quality Problem

Photoplethysmography measures the optical absorption of pulsatile blood. A light source, typically green or infrared, illuminates a vascular bed at the wrist or finger, and a photodetector records the light that returns; as the arterial pulse expands the vascular volume, absorption rises and the detected intensity falls, tracing a waveform whose shape and timing encode cardiac and vascular information (Charlton et al., 2023). From this deceptively simple measurement a remarkable range of parameters can be derived, from heart rate and rhythm to respiration and indices of vascular tone. The morphology of individual pulses,

and the variation in inter-beat intervals, carries the rhythm information that makes PPG attractive for arrhythmia screening (Pereira et al., 2020). It is this dependence on fine morphological and temporal structure that makes the modality so sensitive to corruption.

Atrial fibrillation is the paradigm application and a useful lens for the whole field. It is the most common sustained arrhythmia, it raises stroke risk substantially, and it is often paroxysmal—appearing and resolving unpredictably—so that a single short recording can easily miss it. Continuous wearable monitoring is therefore an almost ideal match to the disease, and a series of landmark studies has demonstrated that consumer devices can identify irregular rhythm in free-living populations and prompt confirmatory testing (Perez et al., 2019; Guo et al., 2019; Tison et al., 2018). Continuous-monitoring algorithms have since been validated on consumer smartwatches and ring-type devices, and integrated into hospital screening pathways for high-risk inpatients (Avram et al., 2021; Kwon et al., 2020; Abdelhamid et al., 2025). Subsequent clinical validations have refined our understanding of where these devices succeed and where they struggle, particularly in the presence of frequent ectopy or competing arrhythmias that mimic the irregularity of fibrillation (Bacevicius et al., 2022; Mannhart et al., 2023; Badertscher et al., 2022). Cost-effectiveness analyses have begun to weigh the population-level benefits of such screening against its harms, including false alarms and the downstream testing they trigger (Chen et al., 2022).

The epidemiological case for such screening is strong. Atrial fibrillation is common and its prevalence rises steeply with age, it is a leading contributor to ischaemic stroke, and a meaningful fraction of cases are silent until a stroke occurs, which is precisely why opportunistic detection in everyday life is attractive (Bayoumy et al., 2021; Guo et al., 2019). Early identification can enable interventions that reduce stroke risk, so even an imperfect screen that reliably surfaces a subset of undiagnosed cases may carry substantial population value. That value, however, is contingent on the screen behaving well outside the controlled conditions in which it was developed, and it is undermined if false positives generate cascades of unnecessary testing or if the device performs unevenly across the populations it is meant to serve (Chen et al., 2022; Rajpurkar et al., 2022). The epidemiology thus motivates the technology and simultaneously raises the evidentiary bar it must clear.

Across this body of work a consistent theme emerges: the binding constraint on wearable cardiac screening is rarely the sophistication of the classifier and frequently the quality of the signal that reaches it. Deep-learning models for rhythm detection have achieved expert-level performance on clean recordings, whether from the electrocardiogram or from PPG (Hannun et al., 2019; Ribeiro et al., 2020; Attia et al., 2019). Yet their accuracy degrades sharply when the input is contaminated, and the contamination encountered in daily life is both pervasive and structured. A foundational design pattern in the field has consequently been to assess signal quality jointly with the clinical task, so that the model can defer on segments it cannot interpret reliably (Pereira et al., 2019; Shin, 2022). The DeepBeat approach, which

couples quality assessment to arrhythmia detection within a single architecture, exemplified the value of treating quality as part of the inference problem rather than a separate preprocessing chore.

The sources of corruption are well catalogued. Motion is the dominant culprit: voluntary movement shifts the sensor, changes contact pressure, and modulates blood flow, producing broadband interference that overlaps the frequency content of the pulse itself and therefore resists simple filtering (Dao et al., 2017; Goh et al., 2020). Baseline wander from respiration and slow vascular changes drifts the signal up and down; ambient light intrudes when the sensor does not seal against the skin; low peripheral perfusion, common in cold conditions or in some patients, reduces the pulsatile component to near the noise floor; and pressure or contact changes alter the optical path unpredictably. Because several of these mechanisms produce in-band interference, the classical separation between signal and noise breaks down, and methods that assume a clean spectral distinction between the two perform poorly (Charlton et al., 2023). Lightweight on-device detectors have been developed to flag corrupted windows in real time so that scarce computation and clinician attention are not wasted on uninterpretable data (Zheng et al., 2024; Liu et al., 2024).

How a screening algorithm uses the PPG waveform shapes its vulnerability to corruption. Approaches that rely chiefly on inter-beat interval statistics—the irregularity of timing characteristic of fibrillation—can tolerate some morphological distortion provided peak detection survives, whereas approaches that exploit the full pulse morphology are richer but more fragile when shape is degraded (Pereira et al., 2020; Tison et al., 2018). Sensor design mediates this trade-off as well: multi-wavelength and multi-channel configurations provide partially redundant views of the same haemodynamics, offering a route to reference-based artifact suppression that single-channel devices lack (Charlton et al., 2023). These choices interact with cleaning strategy, because a reconstruction model implicitly assumes a particular representation of the signal; a pipeline optimized for interval-based detection need not preserve the same features as one built for morphological analysis, and conflating the two leads to cleaning that is technically impressive yet diagnostically beside the point.

The scale of data attrition deserves emphasis because it reframes the engineering target. When a quality front end rejects a large share of recorded windows, the loss is not merely quantitative; it is selective in a way that biases the resulting picture of a patient's rhythm. Segments are most often rejected during movement, and movement coincides with exertion, stress, and the transitions between rest and activity—physiological states during which arrhythmic episodes may be more, not less, likely to occur. A pipeline that silently discards these windows therefore risks systematically under-sampling the very moments of greatest diagnostic interest. Quantifying this trade-off—how much diagnostic yield is forfeited per percentage point of data rejected—is rarely reported, yet it is precisely the quantity that

distinguishes a conservative gate that protects accuracy from an overzealous one that hollows out the monitoring record (Pereira et al., 2019; Liu et al., 2024).

The result is a measurement paradox at the heart of consumer cardiology. The wearable form factor delivers enormous volumes of data, but the conditions that make the form factor attractive—unsupervised, ambulatory, all-day use—are exactly those that degrade the signal. Maximizing the clinical yield of wearable PPG is therefore less a matter of collecting more data than of salvaging more of the data already collected. That observation motivates the shift from subtractive cleaning toward generative restoration, and it frames the central engineering question of the remainder of this review: how can corrupted segments be repaired in a way that demonstrably improves clinical decisions while remaining honest about its own reliability?

## 3. Artifacts and Cleaning Strategies: A Structured Landscape

Designing a cleaning pipeline begins with a clear taxonomy of what is being cleaned. The corruptions that afflict wearable PPG differ not only in their physical origin but in their statistical character, and that character determines which mitigation strategies can succeed. Motion artifact, for instance, is non-stationary and frequently overlaps the cardiac band, so a fixed linear filter cannot remove it without also attenuating the pulse; baseline wander, by contrast, occupies low frequencies and yields readily to high-pass filtering or detrending (Dao et al., 2017; Elgendi, 2016). Treating all noise as a single undifferentiated nuisance is a common cause of disappointing performance, because a method tuned for one corruption may be inert or even harmful against another. Table 1 organizes the principal artifact sources, their physiological or optical origins, their typical signatures in the waveform, and the mitigation approaches most often paired with each.

**Table 1. Principal sources of corruption in wearable PPG and their characteristic mitigation.**

| Artifact source | Origin | Typical signature | Representative mitigation |
|---|---|---|---|
| Motion / movement | Sensor displacement and contact change during activity | Broadband, non-stationary interference overlapping the cardiac band | Adaptive filtering; quality gating; learned reconstruction |
| Baseline wander | Respiration and slow vascular tone changes | Low-frequency drift of the signal offset | High-pass filtering; detrending; empirical decomposition |
| Ambient light | Stray illumination reaching the photodetector | Additive offset and flicker, often at line frequency | Optical shielding; reference-channel subtraction |
| Low perfusion | Reduced peripheral blood flow (e.g., cold, vasoconstriction) | Pulse amplitude near the noise floor; poor morphology | Quality gating; gain control; reconstruction where feasible |
| Sensor contact / | Variable skin coupling and | Amplitude and morphology | Contact detection; segment |

| Artifact source | Origin | Typical signature | Representative mitigation |
|---|---|---|---|
| pressure | applied force | distortion, intermittent dropout | rejection; reconstruction |
| Quantization / compression | Limited bit depth and on-device data reduction | Stair-stepping and loss of fine morphological detail | Higher-resolution capture; model-based interpolation |

Two observations follow from this structure. First, the artifacts that are easiest to remove are generally the least damaging to diagnosis, while the most damaging—motion above all—are precisely those that defeat classical, assumption-driven methods. Second, the relative burden of each artifact is uneven. In free-living recordings, motion accounts for the largest share of rejected data by a wide margin, with baseline wander, ambient light, and low perfusion forming a secondary tier and quantization a minor contributor (Goh et al., 2020; Zheng et al., 2024). Figure 2 depicts an illustrative breakdown of this kind, synthesized from the qualitative emphasis of the literature rather than from a single cohort; the exact proportions vary with device, population, and activity profile, but the ordering is robust and has direct design implications. A pipeline that handles motion well will recover most of the otherwise-lost data, whereas one that perfects baseline-wander removal but stumbles on motion will leave the dominant problem untouched.

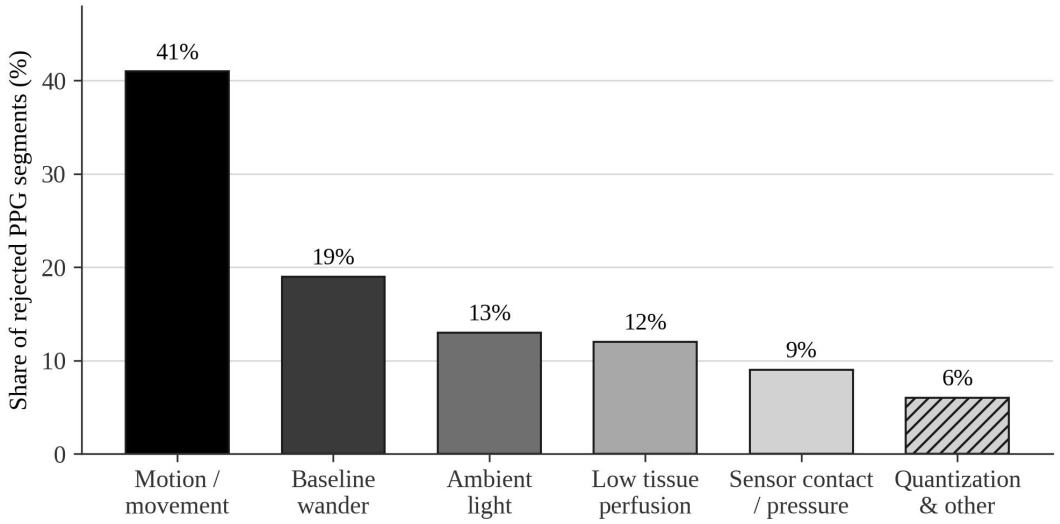


**Figure 2. Illustrative distribution of the artifact sources responsible for rejected wearable PPG segments, ordered by typical prevalence in free-living recordings.**

It is worth dwelling on what such a distribution implies for return on engineering effort. If motion alone accounts for roughly two-fifths of rejected segments, and the secondary tier of baseline wander, ambient light, and low perfusion together for a comparable share, then the marginal value of a method is governed less by its peak performance on any single artifact than by its competence on motion and its graceful behaviour across the mixed corruptions that co-occur in practice. Real recordings seldom present one artifact at a time; a jog in sunlight superimposes motion, ambient light, and perfusion changes

simultaneously. Methods evaluated only on isolated, synthetically injected noise can therefore overstate their field performance, and a recurring lesson of the literature is that robustness to compound, realistic corruption is the property that actually transfers to deployment (Charlton et al., 2023; Ahmed et al., 2023).

A further structural constraint shapes the choice of method: the scarcity of paired clean and corrupted wearable data with reliable labels. Supervised reconstruction ideally requires examples of a corrupted segment alongside its clean counterpart, but in free-living data the clean counterpart usually does not exist, and arrhythmia labels derived from simultaneous reference electrocardiography are expensive to obtain at scale (Pereira et al., 2020). This scarcity has driven interest in synthetic data and augmentation, in which generative models are used not only to clean signals but to manufacture training examples that expand the diversity of artifacts and rhythms a downstream model has seen (Tian et al., 2024; Skandarani et al., 2023). The same technology thus appears twice in the pipeline—as a cleaner of real inputs and as a generator of synthetic training data—and the evaluation challenges it raises, chiefly the risk of learning from plausible but unrepresentative samples, are common to both roles (Brophy et al., 2023).

Against this backdrop the cleaning strategies themselves can be grouped into a small number of families. The oldest is classical signal processing: band-pass and adaptive filters, wavelet and empirical-mode decompositions, and template-based correction. These methods are computationally cheap, interpretable, and effective against well-separated noise, and they remain the sensible default for baseline wander and out-of-band interference (Elgendi, 2016). Their weakness is structural: they encode fixed assumptions about the spectral or morphological separation of signal and noise, and they degrade when those assumptions fail, as they routinely do under motion. A second family abandons repair altogether and instead gates the stream, using a signal-quality index or a learned classifier to admit only segments judged reliable (Shin, 2022; Liu et al., 2024). Quality gating is indispensable—every credible pipeline contains some form of it—but as a sole strategy it caps the achievable data yield at the fraction of segments that happen to be clean, which is exactly the limitation that motivates restoration.

The third and fourth families are learning-based and reconstructive. Denoising autoencoders learn to map corrupted inputs to clean targets by compressing the signal through a bottleneck and reconstructing it, implicitly capturing the manifold of plausible waveforms (Mohagheghian et al., 2024). Generative adversarial approaches push this further by training a generator against a discriminator that learns to distinguish real clean signals from reconstructions, encouraging outputs that are not merely close in a point-wise sense but statistically indistinguishable from genuine physiology (Singh & Pradhan, 2021; Wang et al., 2022; Wang et al., 2023). More recently, diffusion and score-based models, which generate data by reversing a gradual noising process, have entered the field with strong results on physiological time series and a reputation for stable training relative to adversarial methods (Tian et al., 2024). These reconstructive families are the focus of the next section, because they are the ones capable of recovering

rather than merely preserving diagnostic information—and, correspondingly, the ones whose outputs most urgently require scrutiny.

## 4. Generative Cleaning Pipelines and Their Evidence

A generative cleaning pipeline is best understood not as a single denoiser but as a sequence of coupled stages, each with a distinct responsibility, arranged so that information and uncertainty flow forward to a decision. Figure 1 sketches the canonical arrangement. Raw wearable PPG enters a quality-gating stage that separates plainly uninterpretable segments from those worth processing; surviving segments pass to a generative model that denoises or reconstructs them toward the distribution of clean physiology; the cleaned segment is then presented to a downstream cardiac model that performs the clinical task, such as rhythm classification; and finally an uncertainty-aware decision gate decides whether the resulting prediction is trustworthy enough to accept or should be flagged for review. Crucially, the pipeline includes a feedback path: segments the decision gate judges unreliable can be returned for re-examination rather than silently accepted. The architecture makes explicit that cleaning is a means to a decision, not an end in itself.

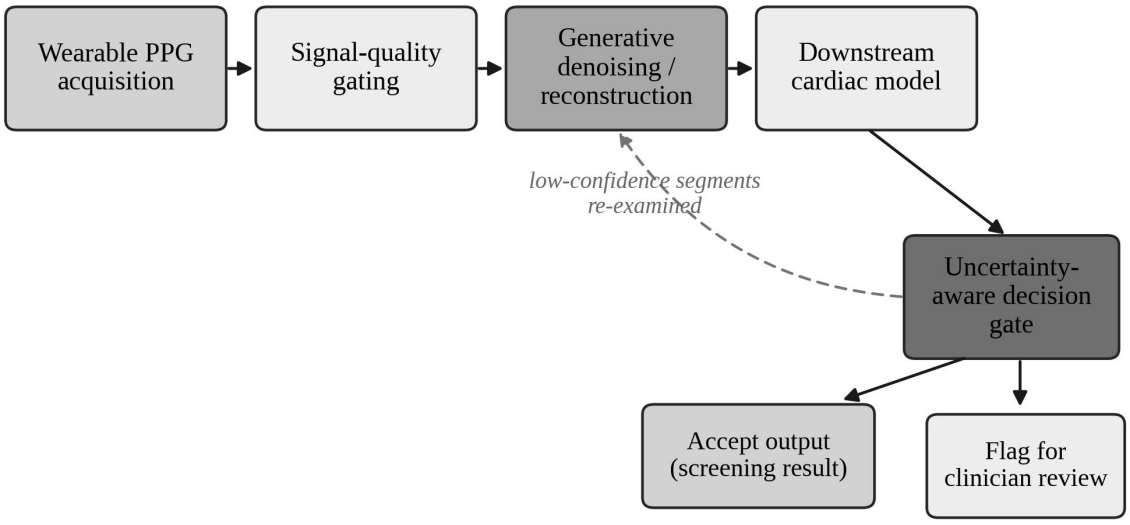


**Figure 1. Stages of a generative cleaning pipeline for wearable cardiac screening, from signal acquisition through quality gating, generative reconstruction, downstream inference, and uncertainty-aware decision-making.**

The reconstruction stage is where generative modelling earns its place. The conditional adversarial formulation has proved especially natural for biosignals: the generator receives a corrupted segment and produces a candidate clean version, while the discriminator, seeing pairs of inputs and outputs, learns to reward reconstructions that match the statistics of real clean signals (Wang et al., 2022). Architectures

borrowed and adapted from image translation—encoder–decoder generators with skip connections, paired with patch-based discriminators—have been applied to one-dimensional cardiac signals with consistent benefit over both classical filtering and plain autoencoders (Singh & Pradhan, 2021; Zhang et al., 2023). The recurring empirical finding is that adversarial reconstruction recovers morphological detail that subtractive methods erase, because the generator is rewarded for producing realistic structure rather than merely minimizing residual energy. Surveys of generative time-series modelling document the rapid proliferation of these designs and the maturation of their training practices (Brophy et al., 2023).

The training objective explains both the power and the fragility of these models. Adversarial reconstruction typically combines two losses: an adversarial term that pushes outputs toward the manifold of realistic signals, and a reconstruction term—often an absolute-error or correlation penalty—that anchors the output to the specific input being repaired. The balance between them is consequential. Weight the adversarial term too heavily and the generator may produce signals that are realistic in general but unfaithful to the particular segment, the mechanism by which hallucination arises; weight the reconstruction term too heavily and the output regresses toward the oversmoothed estimates that plain autoencoders produce, discarding the fine morphology that motivated the adversarial approach in the first place (Wang et al., 2022; Zhang et al., 2023). Stable training therefore depends on careful loss weighting, normalization, and stopping criteria, and the maturation of these practices is as important to the field's progress as any single architectural innovation (Singh & Pradhan, 2021).

The three reconstructive families occupy distinct points on a trade-off surface. Denoising autoencoders are the most economical and the most stable to train, but their tendency to minimize average error makes them prone to oversmoothing, blurring exactly the rare morphological events that screening cares about (Mohagheghian et al., 2024). Adversarial models counter this by rewarding realism, recovering sharp structure at the cost of more delicate training and a standing risk of hallucination (Singh & Pradhan, 2021). Diffusion and score-based models occupy an attractive middle ground in principle, combining high fidelity with stable, mode-covering behaviour, but their iterative sampling imposes a computational burden that sits awkwardly with the power and latency limits of a wrist-worn device (Tian et al., 2024). No family dominates; the right choice depends on whether the deployment prizes faithfulness to rare events, training simplicity, or on-device efficiency, and increasingly on whether the model's uncertainty can be made reliable, which is a property as architecture-dependent as accuracy itself (Gawlikowski et al., 2023).

Evaluating these models honestly is harder than training them. Point-wise error measures such as mean squared error reward smoothness and penalize the realistic high-frequency detail that adversarial models exist to produce, so a denoiser can score well on such metrics while degrading downstream performance, and the reverse can equally occur (Wang et al., 2023). Distributional measures that compare

populations of real and synthetic signals capture realism better but say nothing about any individual segment, which is the unit on which a screening decision is actually made. The absence, in real wearable data, of a ground-truth clean signal for direct comparison compounds the difficulty and pushes evaluation toward the downstream-task and uncertainty-based criteria emphasized throughout this review. The practical recommendation that emerges is to evaluate cleaning with a small battery of complementary measures—downstream clinical performance, per-instance reliability, and a distributional check—rather than trusting any single number (Lambert et al., 2024; Abdar et al., 2021).

Reconstruction quality, however, cannot be judged by realism alone, and this is the conceptual core of the pipeline view. A reconstruction that looks clean but alters the diagnostic content of the segment is worse than useless. The field has therefore converged on a downstream-task criterion: the quality of a cleaned signal is measured by whether it improves the performance of the clinical model that consumes it (Mohagheghian et al., 2024; Pereira et al., 2020). This criterion has the great virtue of aligning the cleaning objective with the clinical objective, and it sidesteps the difficulty that, for real-world wearable data, a pristine ground-truth clean signal is usually unavailable for comparison. Table 2 summarizes the major method families in these terms, contrasting their core idea, their characteristic strengths and limitations, and representative work, so that the trade-offs governing pipeline design are visible at a glance.

**Table 2. Families of signal-cleaning methods for wearable cardiac signals, with their trade-offs and representative work.**

| Method family | Core idea | Strengths | Limitations | Representative work |
|---|---|---|---|---|
| Classical filtering / decomposition | Suppress noise by spectral or morphological separation | Cheap, interpretable, no training data | Fixed assumptions fail under in-band motion noise | Elgendi (2016); Dao et al. (2017) |
| Signal-quality gating | Admit only segments judged reliable | Simple, protects downstream model | Caps yield at the clean fraction; discards data | Shin (2022); Liu et al. (2024) |
| Denoising autoencoders | Map corrupted to clean via a learned bottleneck | Recovers structure; stable training | May oversmooth; can blur rare morphology | Mohagheghian et al. (2024) |
| GAN-based denoising | Adversarial reward for realistic reconstructions | Restores fine diagnostic morphology | Risk of hallucinated features; training care needed | Singh & Pradhan (2021); Wang et al. (2022, 2023) |
| Diffusion / score-based | Reverse a gradual noising process to generate | Strong fidelity; stable, mode-covering | Computationally heavy; latency on device | Tian et al. (2024) |
| Self- / transfer-learning reconstruction | Reuse representations across tasks and cohorts | Data-efficient; aids generalization | Domain shift can still degrade reliability | Pereira et al. (2020); Ahmed et al. (2023) |

What does the accumulated evidence say about how much these methods help? Because head-to-head comparisons use different cohorts, devices, and endpoints, the literature does not yield a single authoritative number; but its qualitative shape is consistent and instructive. Figure 3 illustrates that shape with a representative comparison of downstream screening performance across processing conditions. Corrupting a clean signal collapses both discrimination and balance of errors; classical filtering recovers part of the loss; autoencoder reconstruction recovers more; adversarial denoising recovers more still, often approaching the clean-reference ceiling; and adding an uncertainty-aware gate on top recovers the final increment by withholding the predictions most likely to be wrong (Singh & Pradhan, 2021; Mohagheghian et al., 2024; Lambert et al., 2024). The values shown are illustrative synthesis figures rather than results from one study, and they should be read as a depiction of the field's reported trends; the ordering of the conditions, not the precise heights, carries the message.

Read as an analysis rather than a result, the pattern in Figure 3 carries three quantitative lessons. First, the gap between the corrupted and clean conditions measures the headroom available to any cleaning method; where this gap is small, elaborate reconstruction buys little, and simple gating may suffice. Second, the incremental steps from filtering to autoencoding to adversarial denoising are typically diminishing, so the marginal benefit of greater model complexity must be weighed against its computational and reliability costs, particularly on power-constrained devices (Zheng et al., 2024). Third—and most important for clinical use—the final increment from adversarial denoising to uncertainty-gated denoising is obtained not by improving the reconstruction but by declining to act on its least reliable outputs. That increment is, in effect, error that has been converted into abstention, and whether that exchange is favourable depends entirely on the relative costs of a missed episode and a deferred segment in the intended screening pathway (Lambert et al., 2024; Abdar et al., 2021).

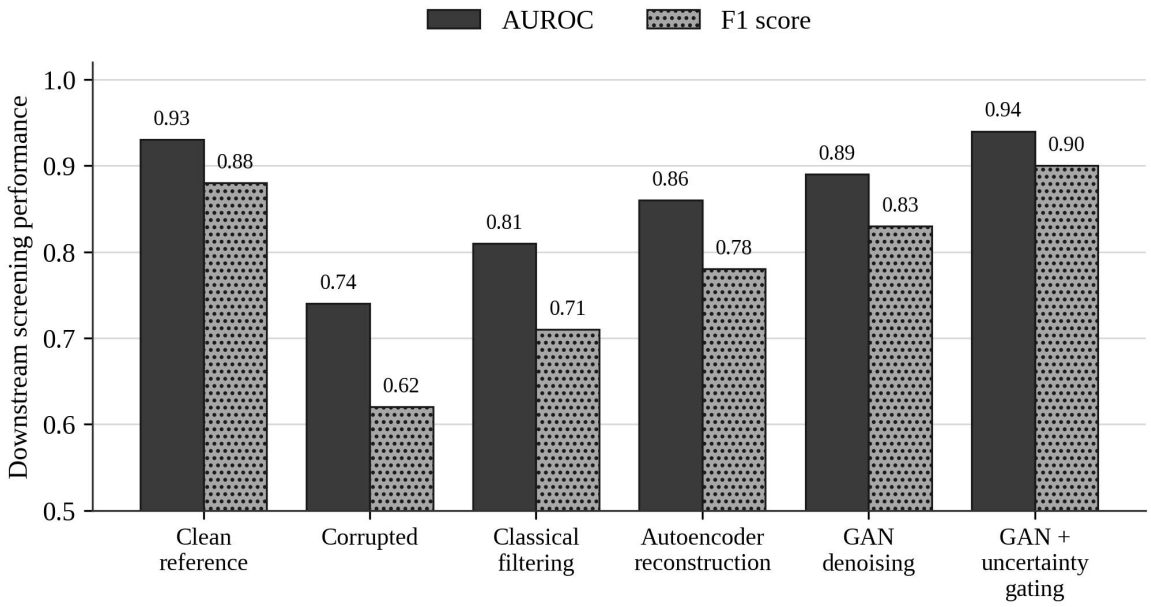


**Figure 3. Illustrative downstream screening performance across processing conditions, showing the incremental gains from classical filtering, learned reconstruction, adversarial denoising, and uncertainty-aware gating.**

Two cautions temper this encouraging picture. The first is generalization. A reconstruction model learns the manifold of the signals it was trained on, and when it meets data from a new device, a new population, or an unusual physiological state, it may reconstruct confidently but wrongly, importing the statistics of its training distribution into a segment that does not belong there (Ahmed et al., 2023; Faust et al., 2018). Because wearable cohorts are rarely stratified finely by artifact type or acquisition setting, some failure modes remain underrepresented in evaluation and surface only after deployment. The second caution is the hallucination risk already noted: the very capacity that lets a generator restore diagnostic morphology also lets it invent it (Yi et al., 2019; Skandarani et al., 2023). Neither caution argues against generative cleaning; both argue that a generative pipeline is incomplete without a mechanism that estimates, per segment, how far its output can be trusted. That mechanism is the subject of the next section.

## 5. Trustworthiness, Uncertainty, and Clinical Translation

If generative cleaning is to support clinical-grade screening, the decisive question is not how good its average output is but whether the system knows when its output is poor. Uncertainty quantification supplies the missing capability: rather than emitting a single prediction, a well-designed pipeline emits a prediction together with a calibrated estimate of how much that prediction should be trusted, so that low-confidence cases can be withheld, deferred, or routed to a clinician (Abdar et al., 2021; Gawlikowski et al., 2023). The appeal in the wearable setting is acute, because ground-truth clean signals are seldom available to verify a reconstruction directly. A measure of confidence that can be computed from the model's own

behaviour, without access to the truth, is therefore the only practical guardrail against acting on a fabricated waveform.

A productive way to obtain such a measure is to ground it in the downstream decision. Instead of asking the abstract question of how accurately a reconstruction reproduces an unknown clean signal, one can ask the operational question of how confidently the downstream model classifies the reconstructed segment, and whether that confidence tracks correctness. When higher uncertainty in the clinical prediction reliably accompanies higher error, the uncertainty becomes an actionable filter: discarding the least confident predictions raises the accuracy of those that remain. This decision-anchored view has the attractive property that it evaluates the cleaning model through the lens of the task it serves, which is exactly the property that task-agnostic image-quality metrics lack (Lambert et al., 2024). It also connects naturally to the broader programme of trustworthy clinical artificial intelligence, in which a model's willingness to abstain is treated as a first-class virtue rather than a failure (Kelly et al., 2019; Topol, 2019).

Calibration is the technical heart of the matter. A confidence estimate is calibrated when its magnitude matches the empirical error rate—when predictions made with eighty percent confidence are correct about eighty percent of the time. Calibration can be assessed with reliability diagrams and summarized with calibration-error metrics, and the literature on uncertainty in medical deep learning has made these tools standard (Abdar et al., 2021; Gawlikowski et al., 2023). Two complications deserve emphasis for wearable cleaning. First, calibration established on one distribution can deteriorate under the distribution shift that wearables routinely encounter, so a pipeline calibrated in the laboratory may be overconfident in the field. Second, confidence can be poorly calibrated for particular classes even when it looks acceptable in aggregate, which matters when the clinically important class—an arrhythmic episode—is the minority. These are not reasons to forgo uncertainty estimation; they are reasons to evaluate it stratified by subgroup and under deliberate shift, and to revisit it prospectively rather than once.

A decision-theoretic perspective clarifies why downstream-anchored uncertainty is more than a convenience. The quantity that ultimately matters is not the fidelity of a reconstruction in the abstract but the expected cost of the action taken on the basis of it, where the cost is defined by the clinical consequences of acting—a missed arrhythmia, an unnecessary referral, a wasted clinician review. Framing trustworthiness in terms of expected decision cost makes the evaluation target explicit and ties the choice of operating threshold to the asymmetry of those consequences, which differs across screening contexts. It also exposes a subtlety: an uncertainty measure that is well behaved for a binary screening decision may not transfer unchanged to a finer-grained, multi-class task, because the relationship between a model's output entropy and its error rate need not be monotone once more than two classes are in play (Gawlikowski et al., 2023). Designing the uncertainty signal around the specific decision, rather than adopting a generic confidence score, is therefore part of building a trustworthy pipeline.

Out-of-distribution inputs are the sharpest test of this machinery, and the place where retrospective evaluation is most likely to flatter a system. A generative model encountering a waveform unlike anything in its training set may reconstruct it confidently and wrongly, while the downstream classifier, equally unfamiliar with the case, may report low uncertainty despite high error—a doubly silent failure. Detecting such cases requires uncertainty estimates that remain meaningful under shift and, ideally, explicit out-of-distribution detection layered onto the pipeline (Abdar et al., 2021; Rajpurkar et al., 2022). Because consumer cohorts are rarely stratified by device, skin tone, activity, or comorbidity, the subgroups most exposed to these failures are often those least represented in the data used to validate the system, which is why equity and robustness are not separate concerns but two views of the same gap (Esteva et al., 2019; Kelly et al., 2019).

Figure 4 expresses the operational consequence of taking uncertainty seriously as a screening funnel. Continuous monitoring yields a large volume of recorded time, of which quality gating retains the analyzable portion; generative cleaning expands that portion by reconstructing segments toward the in-distribution region; the downstream model produces predictions, of which the uncertainty gate keeps only the confident ones; and a clinician confirms the small, high-value residue that constitutes a clinical-grade result. At each stage the volume of data falls while the trustworthiness of what remains rises. The art of pipeline design lies in shaping this funnel so that it discards as little diagnostic information as possible while admitting as little unreliable information as possible—a balance that the generative and uncertainty stages negotiate together.

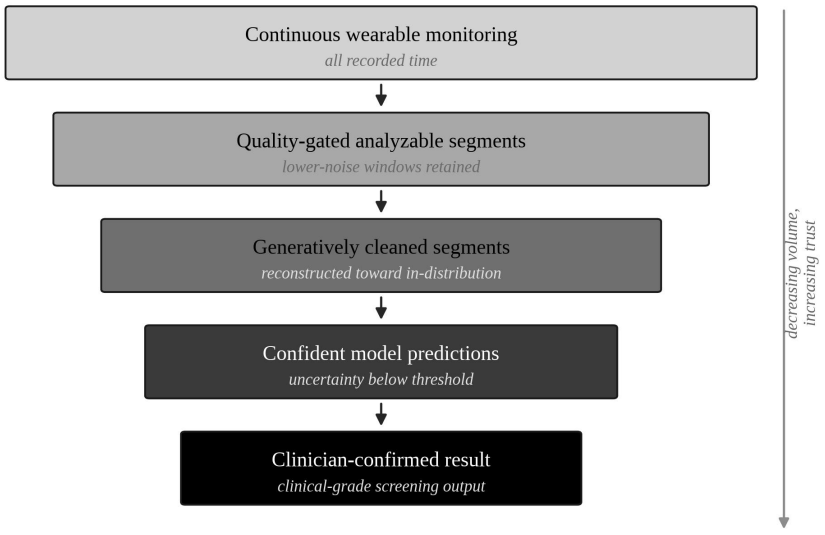


**Figure 4. The screening funnel: as data passes from continuous monitoring to a clinician-confirmed result, volume decreases while the trustworthiness of the retained information increases.**

Even a perfectly calibrated pipeline must eventually meet a clinical workflow, and that encounter introduces its own constraints. Screening at population scale generates alerts at population scale, and the value of an alert depends on the capacity to act on it; a system that surfaces many low-confidence flags can induce alert fatigue, erode clinician trust, and provoke anxiety in users, harms that are easy to overlook when optimizing a classification metric (Chen et al., 2022; Bayoumy et al., 2021). The uncertainty gate is therefore not only a safeguard against hallucinated outputs but a throttle on alert volume, and its threshold is a clinical and operational decision as much as a statistical one. Human-in-the-loop confirmation—the final stage of the screening funnel—remains essential, both because it catches residual failures and because regulatory and ethical frameworks for AI-assisted screening generally presume meaningful human oversight rather than autonomous decision-making (Vasey et al., 2022; Kelly et al., 2019).

Reliability is not a property a pipeline establishes once and retains. Wearable populations, devices, firmware, and usage patterns drift over time, and a model calibrated at launch can silently decay as the data it meets diverges from the data it learned from (Ahmed et al., 2023). Treating a deployed pipeline as a static artifact is therefore a mistake; it requires ongoing monitoring of its calibration and deferral behaviour, periodic re-evaluation on fresh, representative data, and mechanisms to detect when performance has degraded enough to warrant retraining or withdrawal. This post-deployment vigilance is the counterpart, in the temporal dimension, of the subgroup and out-of-distribution scrutiny advocated above, and it is increasingly an explicit expectation of frameworks for clinical artificial intelligence rather than an optional nicety (Vasey et al., 2022; Topol, 2019).

Trustworthiness, however, is broader than calibration, and clinical translation imposes requirements that no single metric captures. Reporting standards for early-stage clinical evaluation of decision-support systems ask developers to specify intended use, operating points, and failure handling before prospective testing begins (Vasey et al., 2022). Regulators and reviewers increasingly expect evidence of generalization across sites and devices, transparency about training data, and attention to equity, since a model that performs unevenly across skin tones, ages, or comorbidities can widen rather than narrow health disparities (Rajpurkar et al., 2022; Esteva et al., 2019). Table 3 organizes these expectations into dimensions of translational readiness, pairing each with what it assesses, a candid view of the field's current status, and the gap that most needs closing. The table is deliberately sobering: technical performance is the most mature dimension, while prospective clinical validation and equity assessment lag well behind the pace of methodological innovation.

**Table 3. Dimensions of translational readiness for generative cleaning pipelines in wearable cardiac screening.**

| Dimension | What it assesses | Current status | Key gap |
|---|---|---|---|
| Technical performance | Accuracy and robustness of cleaning and inference | Relatively mature; strong on benchmark data | Standardized, shared evaluation protocols |
| Per-instance trustworthiness | Whether each output can be trusted individually | Emerging; decision-anchored methods promising | Validation without ground-truth clean signals |
| Calibration | Match between confidence and empirical error | Tools exist; aggregate calibration achievable | Stability under shift and across subgroups |
| Clinical validation | Benefit and harm in real screening pathways | Limited; mostly retrospective to date | Prospective, outcome-based studies |
| Regulatory / reporting | Transparency, intended use, failure handling | Frameworks maturing for AI devices | Consistent reporting for generative components |
| Equity / generalizability | Even performance across populations and devices | Under-examined; cohorts often narrow | Diverse cohorts and stratified evaluation |

Viewed together, these dimensions reframe the engineering goal. A generative cleaning pipeline is not finished when its reconstructions look convincing or even when its average downstream accuracy is high; it is finished, for clinical purposes, when it can demonstrate calibrated confidence under realistic shift, document its behaviour for review, and show benefit in a prospective screening pathway without disadvantaging any group it serves (Vasey et al., 2022; Rajpurkar et al., 2022). Much of the methodological apparatus needed to meet these requirements already exists, scattered across the literatures of medical image analysis, time-series classification, uncertainty estimation, and the broader application of deep learning across biology and medicine (Litjens et al., 2017; Ismail Fawaz et al., 2019; Hong et al., 2020; Ching et al., 2018). The opportunity, and the unmet need, is to assemble that apparatus into wearable cleaning pipelines as a matter of course rather than as an afterthought.

## 6. Conclusion and Future Directions

Consumer wearables have made continuous cardiac sensing ordinary, but they have not made it clean. The central argument of this review is that the gap between abundant raw data and dependable screening is best closed not by discarding corrupted segments, nor by filtering them until their diagnostic content erodes, but by reconstructing them with generative models whose outputs are evaluated against the clinical decisions they inform and gated by calibrated estimates of their own reliability. Artifact removal and screening, on this view, are not sequential chores to be optimized separately; they are facets of one uncertainty-aware pipeline whose value is realized only at the point of decision (Bayoumy et al., 2021; Lambert et al., 2024).

Three priorities follow for the field. The first is evaluation that matches the ambition of the methods. Cleaning models should be benchmarked by downstream clinical performance on shared, well-characterized datasets, with uncertainty assessed stratified by subgroup and under deliberate distribution shift, so that reported gains transfer beyond the cohort that produced them (Pereira et al., 2019; Gawlikowski et al., 2023). The second is honesty about hallucination. Because generative reconstruction can fabricate plausible morphology, every pipeline should carry an explicit mechanism for detecting and withholding low-confidence outputs, and should be reported in a way that exposes, rather than obscures, how often it defers (Brophy et al., 2023; Skandarani et al., 2023). The third is translation discipline: prospective, outcome-based studies and transparent reporting are the currency in which clinical-grade claims must be paid, and they remain in short supply relative to algorithmic progress (Vasey et al., 2022; Chen et al., 2022).

Underlying all three priorities is a shift in what counts as success. The natural instinct of method development is to maximize an average metric on a benchmark, but a screening pipeline is a safety-relevant system whose worst-case and per-instance behaviour matter as much as its mean. A model that is accurate on average yet silently confident on the cases it gets wrong is more dangerous, in deployment, than a slightly less accurate model that reliably signals its own doubt (Topol, 2019; Hannun et al., 2019). Reorienting evaluation, reporting, and even training objectives around calibrated, decision-anchored reliability—rather than headline accuracy—is the cultural change that would do most to move generative cleaning from promising research toward dependable practice.

A concrete near-term agenda follows from this analysis. The field would benefit from shared benchmarks that pair realistic, compound-corrupted wearable signals with reference labels and that score methods on downstream clinical performance and calibrated reliability rather than reconstruction error alone. It would benefit from reporting conventions that disclose deferral rates and subgroup performance as a matter of routine, so that a method's silence is as visible as its accuracy. And it would benefit from prospective studies that follow a generative cleaning pipeline through an actual screening pathway to a clinical outcome, closing the gap between retrospective promise and deployed benefit (Vasey et al., 2022; Mannhart et al., 2023). None of these requires a conceptual breakthrough; each requires the discipline to hold a powerful and fast-moving technology to the evidentiary standards that clinical use demands.

Several technical frontiers are likely to shape the next phase. Diffusion and score-based reconstruction may offer the fidelity of adversarial methods with more stable training and better-behaved uncertainty, if their computational cost can be tamed for on-device use (Tian et al., 2024). Self-supervised and transfer-learning strategies promise to ease the chronic scarcity of labelled clean wearable data and to improve robustness across devices and populations (Ahmed et al., 2023; Esteva et al., 2019). Edge deployment will demand lightweight cleaning and gating that run within the power and latency budgets of

a wrist-worn device, extending the real-time quality-assessment work already underway (Zheng et al., 2024; Liu et al., 2024). And the broader infrastructure of connected health—secure data handling, integrity of the sensing chain, and the maturing toolkit of artificial intelligence and adjacent computational methods—will determine whether pipelines validated in research can be operated safely at scale (Lu & Xu, 2019; Xu et al., 2021; Zhang & Lu, 2021; Lu et al., 2024).

The trajectory is encouraging. Where the previous generation of wearable cardiology asked whether a noisy consumer signal could be classified at all, the current generation asks how much of the discarded signal can be responsibly recovered, and the next will ask how that recovery can be made trustworthy enough to act on. Generative cleaning pipelines, designed and governed as integrated, uncertainty-aware systems, are a credible answer to that question. If the field holds itself to the standards of calibration, prospective validation, and equitable performance that clinical deployment demands, the path from artifact removal to clinical-grade screening is not only plausible but already partly built (Topol, 2019; Rajpurkar et al., 2022).

## Ethics approval and consent to participate

Not applicable. This review synthesizes previously published studies and did not involve the recruitment of human participants or the collection of new human data.

## Consent for publication

Not applicable.

## Availability of data and materials

Not applicable. No new datasets were generated. The figures present illustrative, synthesized values used for explanatory purposes and do not constitute empirical measurements.

## Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Competing interests

The author declares no competing interests.

## AI use disclosure

An AI assistant was used to support drafting, language editing, figure generation, and reference formatting. All content was reviewed and verified by the author, who takes full responsibility for the final manuscript. AI tools are not listed as authors.

## Author contributions

Y. Lu: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing.

## Acknowledgements

The author thanks colleagues for helpful discussions on wearable signal processing and trustworthy machine learning.

## References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Abdelhamid, K., Reissenberger, P., Piper, D., Koenig, N., Hoelz, B., Schlaepfer, J., Gysler, S., McCullough, H., Ramin-Wright, S., Gabathuler, A.-L., Khandpur, J., Meier, M., & Eckstein, J. (2025). Fully automated photoplethysmography-based wearable atrial fibrillation screening in a hospital setting. Diagnostics, 15(10), 1233. https://doi.org/10.3390/diagnostics15101233

Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M., & Gandomi, A. H. (2023). Deep learning modelling techniques: Current progress, applications, advantages, and challenges. Artificial Intelligence Review, 56(11), 13521–13617. https://doi.org/10.1007/s10462-023-10466-8

Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S., & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet, 394(10201), 861–867. https://doi.org/10.1016/S0140-6736(19)31721-0

Avram, R., Ramsis, M., Cristal, A. D., Nathan, V., Zhu, L., Kim, J., Kuang, J., Gao, A., Vittinghoff, E., Pletcher, M. J., Olgin, J. E., & Tison, G. H. (2021). Validation of an algorithm for continuous monitoring of atrial fibrillation using a consumer smartwatch. Heart Rhythm, 18(9), 1482–1490. https://doi.org/10.1016/j.hrthm.2021.03.044

Bacevicius, J., Abramikas, Z., Dvinelis, E., Audzijoniene, D., Petrylaite, M., Marinskiene, J., Staigyte, J., Karuzas, A., Juknevicius, V., Jakaite, R., Basyte-Bacevice, V., Bileisiene, N., Solosenko, A., Sokas, D., Petrenas, A., Butkuviene, M., Paliakaite, B., Daukantas, S., Rapalis, A., … Aidietis, A. (2022). High specificity wearable device with photoplethysmography and six-lead electrocardiography for atrial fibrillation detection challenged by frequent premature contractions: DoubleCheck-AF. Frontiers in Cardiovascular Medicine, 9, 869730. https://doi.org/10.3389/fcvm.2022.869730

Badertscher, P., Lischer, M., Mannhart, D., Knecht, S., Isenegger, C., du Fay de Lavallaz, J., Schaer, B., Osswald, S., Kühne, M., & Sticherling, C. (2022). Clinical validation of a novel smartwatch for automated detection of atrial fibrillation. Heart Rhythm O2, 3(2), 208–210. https://doi.org/10.1016/j.hroo.2022.02.004

Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., Martin, S. S., Muse, E. D., Turakhia, M. P., Tarakji, K. G., & Elshazly, M. B. (2021). Smart wearable devices in cardiovascular care: Where we are and how to move forward. Nature Reviews Cardiology, 18(8), 581–599. https://doi.org/10.1038/s41569-021-00522-7

Brophy, E., Wang, Z., She, Q., & Ward, T. (2023). Generative adversarial networks in time series: A systematic literature review. ACM Computing Surveys, 55(10), 1–31. https://doi.org/10.1145/3559540

Charlton, P. H., Allen, J., Bailón, R., Baker, S., Behar, J. A., Chen, F., Clifford, G. D., Clifton, D. A., Davies, H. J., Ding, C., Ding, X., Dunn, J., Elgendi, M., Ferdoushi, M., Franklin, D., Gil, E., Hassan, M. F., Hernesniemi, J., Hu, X., … Zhu, T. (2023). The 2023 wearable photoplethysmography roadmap. Physiological Measurement, 44(11), 111001. https://doi.org/10.1088/1361-6579/acead2

Chen, W., Khurshid, S., Singer, D. E., Atlas, S. J., Ashburner, J. M., Ellinor, P. T., McManus, D. D., Lubitz, S. A., & Chhatwal, J. (2022). Cost-effectiveness of screening for atrial fibrillation using wearable devices. JAMA Health Forum, 3(8), e222419. https://doi.org/10.1001/jamahealthforum.2022.2419

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

Dao, D., Salehizadeh, S. M. A., Noh, Y., Chong, J. W., Cho, C. H., McManus, D., Darling, C. E., Mendelson, Y., & Chon, K. H. (2017). A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time–frequency spectral features. IEEE Journal of Biomedical and Health Informatics, 21(5), 1242–1253. https://doi.org/10.1109/JBHI.2016.2612059

Elgendi, M. (2016). Optimal signal quality index for photoplethysmogram signals. Bioengineering, 3(4), 21. https://doi.org/10.3390/bioengineering3040021

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine, 161, 1–13. https://doi.org/10.1016/j.cmpb.2018.04.005

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Goh, C. H., Tan, L. K., Lovell, N. H., Ng, S. C., Tan, M. P., & Lim, E. (2020). Robust PPG motion artifact detection using a 1-D convolution neural network. Computer Methods and Programs in Biomedicine, 196, 105596. https://doi.org/10.1016/j.cmpb.2020.105596

Guo, Y., Wang, H., Zhang, H., Liu, T., Liang, Z., Xia, Y., Yan, L., Xing, Y., Shi, H., Li, S., Liu, Y., Liu, F., Feng, M., Chen, Y., & Lip, G. Y. H. (2019). Mobile photoplethysmographic technology to detect atrial fibrillation. Journal of the American College of Cardiology, 74(19), 2365–2375. https://doi.org/10.1016/j.jacc.2019.08.019

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

Hong, S., Zhou, Y., Shang, J., Xiao, C., & Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. Computers in Biology and Medicine, 122, 103801. https://doi.org/10.1016/j.compbiomed.2020.103801

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2

Kwon, S., Hong, J., Choi, E.-K., Lee, B., Baik, C., Lee, E., Jeong, E.-R., Koo, B.-K., Oh, S., & Yi, Y. (2020). Detection of atrial fibrillation using a ring-type wearable device (CardioTracker) and deep learning analysis of photoplethysmography signals: Prospective observational proof-of-concept study. Journal of Medical Internet Research, 22(5), e16443. https://doi.org/10.2196/16443

Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artificial Intelligence in Medicine, 150, 102830. https://doi.org/10.1016/j.artmed.2024.102830

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, J., Hu, S., Hu, Q., Wang, D., & Yang, C. (2024). A lightweight hybrid model using multiscale Markov transition field for real-time quality assessment of photoplethysmography signals. IEEE Journal of Biomedical and Health Informatics, 28(3), 1078–1088. https://doi.org/10.1109/JBHI.2023.3331975

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Mannhart, D., Lischer, M., Knecht, S., du Fay de Lavallaz, J., Strebel, I., Serban, T., Vogtländer, K., Schaer, B., Osswald, S., Mueller, C., Kühne, M., Sticherling, C., & Badertscher, P. (2023). Clinical validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL Wearable Study. JACC: Clinical Electrophysiology, 9(2), 232–242. https://doi.org/10.1016/j.jacep.2022.09.011

Mohagheghian, F., Han, D., Ghetia, O., Chen, D., Peitzsch, A., Nishita, N., Ding, E., Otabil, E. M., Noorishirazi, K., Hamel, A., DiMezza, D., Curtin, A., Tran, K. V., McManus, D. D., & Chon, K. H. (2024). Atrial fibrillation detection on reconstructed photoplethysmography signals collected from a smartwatch using a denoising autoencoder. Expert Systems with Applications, 237, 121611. https://doi.org/10.1016/j.eswa.2023.121611

Pereira, T., Ding, C., Gadhoumi, K., Tran, N., Colorado, R. A., Meisel, K., & Hu, X. (2019). Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation. Physiological Measurement, 40(12), 125002. https://doi.org/10.1088/1361-6579/ab5b84

Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., Colorado, R., Meisel, K., & Hu, X. (2020). Photoplethysmography based atrial fibrillation detection: A review. npj Digital Medicine, 3, 3. https://doi.org/10.1038/s41746-019-0207-9

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D.,

Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., … Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909–1917. https://doi.org/10.1056/NEJMoa1901183

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira, W., Schön, T. B., & Ribeiro, A. L. P. (2020). Automatic diagnosis of the short-duration 12-lead ECG using a deep neural network. Nature Communications, 11(1), 1760. https://doi.org/10.1038/s41467-020-15432-4

Shin, H. (2022). Deep convolutional neural network-based signal quality assessment for photoplethysmogram. Computers in Biology and Medicine, 145, 105430. https://doi.org/10.1016/j.compbiomed.2022.105430

Singh, P., & Pradhan, G. (2021). A new ECG denoising framework using generative adversarial network. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(2), 759–764. https://doi.org/10.1109/TCBB.2020.2976981

Skandarani, Y., Jodoin, P.-M., & Lalande, A. (2023). GANs for medical image synthesis: An empirical study. Journal of Imaging, 9(3), 69. https://doi.org/10.3390/jimaging9030069

Tian, M., Chen, B., Guo, A., Jiang, S., & Zhang, A. R. (2024). Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. Journal of the American Medical Informatics Association, 31(11), 2529–2539. https://doi.org/10.1093/jamia/ocae229

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., Mikell, C., Sohoni, N., Hsieh, J., & Marcus, G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology, 3(5), 409–416. https://doi.org/10.1001/jamacardio.2018.0136

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9

Wang, X., Chen, B., Zeng, M., Wang, Y., Liu, H., Liu, R., Tian, L., & Lu, X. (2022). An ECG signal denoising method using conditional generative adversarial net. IEEE Journal of Biomedical and Health Informatics, 26(7), 2929–2940. https://doi.org/10.1109/JBHI.2022.3169325

Wang, H., Ma, Y., Zhang, A., Lin, D., Qi, Y., & Li, J. (2023). Deep convolutional generative adversarial network with LSTM for ECG denoising. Computational and Mathematical Methods in Medicine, 2023, 6737102. https://doi.org/10.1155/2023/6737102

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. Medical Image Analysis, 58, 101552. https://doi.org/10.1016/j.media.2019.101552

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, P., Jiang, M., Li, Y., Xia, L., Wang, Z., Wu, Y., Wang, Y., & Zhang, H. (2023). An efficient ECG denoising method by fusing ECA-Net and CycleGAN. Mathematical Biosciences and Engineering, 20(7), 13415–13433. https://doi.org/10.3934/mbe.2023598

Zheng, Y., Wu, C., Cai, P., Zhong, Z., Huang, H., & Jiang, Y. (2024). Tiny-PPG: A lightweight deep neural network for real-time detection of motion artifacts in photoplethysmogram signals on edge devices. Internet of Things, 25, 101007. https://doi.org/10.1016/j.iot.2023.101007