Trust-by-Design Analytics for Consumer Health Wearables: A Cost-Sensitive Framework for Deploying Generative AI in Continuous Cardiac Monitoring

Haoran Zhou[1], Yufeng Tao[2,*], Miaomiao Zhang[2]

[1] Smart Health College, Hefei Financial and Economic Vocational College, Hefei, China 231200

[2] Digital Finance and Business College, Hefei Financial and Economic Vocational College, Hefei, China 231200

* Email: 857283244@qq.com (Corresponding Author)

Received 20 June 2023 · Revised 4 August 2023 · Accepted 18 September 2023 · Published 30 December 2023

Abstract

Continuous cardiac monitoring through photoplethysmography (PPG) on consumer wearables has migrated from a fitness convenience to a consequential clinical signal, particularly for screening atrial fibrillation (AF) at population scale. Yet the operating environment of such devices — motion artefacts, perfusion variability, and ambulatory noise — degrades the very deep-learning classifiers that vendors and payers rely on to convert raw waveforms into business and clinical decisions. This article develops a Trust-by-Design Analytics framework for deploying generative artificial intelligence (GenAI) within consumer health wearables, with a specific focus on the AF-screening pipeline. The framework couples a generative adversarial denoiser, a calibrated probabilistic classifier, and a decision-theoretic uncertainty-quantification (DTUQ) layer into a single, cost-sensitive deployment unit. Using a controlled numerical study calibrated to published wearable AF cohorts, we show that DTUQ-gated decisions recover area-under-curve performance from 0.69 under realistic noise to 0.85 after generative reconstruction, while reducing expected misclassification cost by approximately 41% relative to a cost-sensitive noisy baseline at a clinically realistic false-negative-to-false-positive cost ratio of 5:1. We articulate three managerial implications. First, calibration error — not raw accuracy — is the binding constraint on safe scaling of consumer wearables into reimbursable digital-health pathways. Second, the marginal value of a generative denoiser is monotone in the underlying cost asymmetry, which means investment cases for GenAI in this domain must be built on cost-sensitive, rather than accuracy-based, evaluation. Third, an explicit abstain-and-refer option closes the gap between regulatory expectations for explainable medical AI and the operational reality of always-on consumer sensing. The framework is intended as a deployment template for analytics teams working at the boundary of regulated clinical decision support and at-scale consumer engagement.

Keywords: Trust-by-design analytics; generative adversarial networks; uncertainty quantification; photoplethysmography; atrial fibrillation; consumer wearables; cost-sensitive learning; digital health

Trust-by-Design Analytics for Consumer Health Wearables: A Cost-Sensitive Framework for Deploying Generative AI in Continuous Cardiac Monitoring

1. Introduction

The smartwatch on a consumer's wrist now routinely produces a stream of optical cardiac measurements that, only a decade ago, required a clinical-grade ambulatory monitor and a cardiology referral. Photoplethysmography (PPG), the technique of inferring pulsatile blood-volume change from the absorption of green or infrared light at the skin, has matured from a comfort sensor for step counting and heart-rate display into a diagnostic substrate for population-scale screening of atrial fibrillation (AF) and other rhythm disorders (Allen, 2007; Pereira et al., 2020). The convergence of three forces — cheap optical front-ends, on-device machine learning accelerators, and a regulatory environment increasingly hospitable to software-as-a-medical-device — has transformed the analytics question for digital-health vendors. The question is no longer whether a deep neural network can recognise AF on a clean PPG strip; multiple studies have answered that affirmatively (Pereira et al., 2020; Tison et al., 2018). The question is whether such a network can be trusted to make, or to defer, a decision that drives a downstream clinical action — a notification, a referral, an electrocardiogram (ECG) confirmation — under the messy operating conditions of an unsupervised consumer wearing the device while exercising, sleeping, or driving.

Trust, in this deployment context, is not a property of the model in isolation; it is a property of the analytics pipeline that surrounds the model and translates its probabilistic outputs into business-relevant decisions. A vendor that ships a 95%-accurate classifier is not the same as a vendor that ships a 95%-accurate classifier coupled with a calibrated uncertainty estimate, an explicit abstain-and-refer pathway, and an auditable cost model relating false alarms to missed diagnoses. The first vendor is selling a technology; the second is selling a deployable analytics product. The distinction matters because regulators, insurers, and clinicians are converging on the second standard (Char et al., 2018; Kompa et al., 2021; Wiens et al., 2019), while a substantial portion of academic literature still optimises for the first.

This article develops a Trust-by-Design Analytics framework — abbreviated hereafter as TBDA — for deploying generative artificial intelligence in continuous cardiac monitoring on consumer wearables. The framework is built around three layers that correspond to three distinct analytic responsibilities: a signal-acquisition layer that uses a generative adversarial network (GAN) to denoise PPG waveforms; a domain-adaptation layer that aligns a classifier trained on curated clinical data to the noise distribution of consumer devices; and a decision layer that uses decision-theoretic uncertainty quantification (DTUQ) to gate, defer, or escalate predictions according to a cost-sensitive operating policy. The premise is that each layer absorbs a category of risk that the analytics function would otherwise either accept silently or paper over with an inflated headline accuracy figure.

The contribution of the article is fourfold. First, we recast a body of technical work on PPG denoising and AF detection as a deployable analytics framework, by making the cost structure of the deployment decision explicit and quantitative. Second, we provide a numerical study, calibrated to noise levels and prevalence figures reported in the wearable-AF literature, that decomposes the area-under-curve (AUC) performance gains attributable to each layer of the framework. Third, we show — both theoretically, via a Bayes-risk argument, and empirically, via the numerical study — that the marginal economic value of generative denoising scales with the cost asymmetry between false negatives and false positives. This is a managerially significant finding: it implies that the business case for GenAI in cardiac monitoring is strongest where the clinical stakes are highest, which is the opposite of the pattern observed in most consumer-wearable feature roadmaps. Fourth, we connect the framework to recent work on management analytics and trustworthy artificial intelligence, treating consumer cardiac wearables as a tractable case study for the broader question of how organisations can govern the deployment of generative models in regulated domains.

ISSN: 3068-3661 | © 2023 INATGI (Institute of Advanced Technology and Green Innovation)

An immediate practical question that any deployment framework must answer is whether the additional architectural complexity is worth carrying. The simple alternative — a single end-to-end deep classifier trained on the deployment distribution — has the appeal of operational simplicity. Section 5 reports a head-to-head comparison and shows that the trust-by-design framework reduces normalised expected cost by roughly 56% relative to that simple alternative at a clinically realistic cost ratio. The complexity is therefore not an aesthetic preference but a measurable contributor to deployment value. Equally important, the complexity is contained inside well-defined audit boundaries. Each of the three layers can be inspected, monitored, and replaced independently of the others, which makes the framework compatible with the kind of staged regulatory review that increasingly characterises software-as-a-medical-device pathways (Char et al., 2018; Larson et al., 2020). We return to these governance considerations in Section 6.

The article is organised around the standard journal structure for a deployment-oriented analytics study. Section 2 reviews the relevant literature in five strands: PPG and AF screening, deep learning for cardiac signals, generative AI for biosignal reconstruction, uncertainty quantification and calibration, and the management-analytics view of trustworthy AI. Section 3 formulates the deployment problem as a cost-sensitive Bayes-risk minimisation under distribution shift and introduces the notation used throughout. Section 4 specifies the three-layer methodology, including the GAN denoiser, the calibrated classifier, and the DTUQ gating rule. Section 5 reports the numerical experiments, including a sensitivity analysis across noise levels and cost ratios. Section 6 discusses the managerial and regulatory implications, drawing the connection back to published work on management analytics and decision support. Section 7 concludes.

Two boundary conditions deserve up-front acknowledgement. First, the framework is descriptive of a deployment template rather than prescriptive of a specific clinical decision rule. The cost ratios used in Section 5 are illustrative; in practice they would be estimated from the specific care pathway and reimbursement context in which the wearable operates. Second, the framework concerns the analytics pipeline downstream of the optical sensor, not the sensor itself. Hardware-level innovations in PPG acquisition — multi-wavelength sources, contact-pressure sensing, motion compensation through inertial measurement units — are complementary to, but distinct from, the analytics question we address.

The motivation for the framework draws on a particular reading of the digital-health analytics literature. The dominant narrative emphasises the supply side: more data, larger models, better benchmarks. The trust-by-design framing emphasises the demand side: under what operating conditions, with what cost structure, against what regulatory expectation, will the analytics output actually be acted upon? This shift of emphasis is not unique to the cardiac-monitoring case; it is consistent with the broader argument that the deployment of analytics in regulated organisations is bounded by the credibility of the decisions the analytics support, not by the raw predictive performance of the underlying model (Varghese, 2020; Wiens et al., 2019). For consumer cardiac wearables, the demand-side perspective is especially salient because the analytics is the product. The wearable as a physical artefact is a commodity; the differentiation is in the trustworthiness of the inferences it produces, and trustworthiness is precisely what a trust-by-design analytics framework is built to deliver and to demonstrate.

2. Literature Review

We survey five strands of literature whose intersection defines the design space for trust-by-design analytics in consumer cardiac wearables. The first concerns the physiology and signal characteristics of PPG and the epidemiology of AF screening; the second concerns the application of deep learning to cardiac waveforms; the third concerns generative artificial intelligence for biosignal denoising and synthesis; the fourth concerns uncertainty quantification, calibration, and decision-theoretic deployment; and the fifth concerns the management-analytics view of trustworthy AI in regulated domains.

2.1 PPG-based Atrial Fibrillation Screening on Consumer Wearables

PPG is an optical method whose signal arises from the differential absorption of light by pulsatile arterial blood; reviews of its physiology and clinical applications go back two decades (Allen, 2007; Elgendi, 2012). The use of PPG to screen for AF rests on the observation that the irregular ventricular response characteristic of AF produces a recognisable irregularity in the pulse-to-pulse interval distribution. Two large pragmatic studies established the feasibility of consumer-wearable AF screening at population scale. The Apple Heart Study enrolled approximately 419,000 participants and reported a positive predictive value of around 84% for the smartwatch-generated tachogram irregularity notification (Perez et al., 2019). An earlier study using the Cardiogram application on commercially available smartwatches similarly demonstrated passive AF detection with clinically useful sensitivity (Tison et al., 2018). These studies, together with reviews on PPG-based AF detection and the broader landscape of cardiovascular wearables (Bayoumy et al., 2021; Pereira et al., 2020), have shifted the regulatory conversation: AF screening on consumer devices is no longer a research curiosity but a recognised digital-health pathway addressed in the 2020 European Society of Cardiology guidelines (Hindricks et al., 2021). At the same time, the limitations have become more visible: positive predictive value depends sharply on the prevalence of AF in the screened cohort (Pereira et al., 2020; Perez et al., 2019), and the sensitivity of consumer PPG to motion artefact remains a major failure mode in real-world deployment (Charlton et al., 2016; Reiss et al., 2019).

The reviews of consumer health wearables (Dunn et al., 2018; Piwek et al., 2016) make clear that PPG-based cardiac monitoring is now embedded in a broader ecosystem of health analytics, in which the wearable is one node in a data flow that includes cloud-side processing, clinician dashboards, and reimbursement decisions. This ecosystem context — the wearable is not a standalone diagnostic but a triage instrument inside a care pathway — is decisive for the analytics framework we develop.

2.2 Deep Learning for Cardiac Signal Interpretation

Deep neural networks now achieve cardiologist-level performance on rhythm classification from ambulatory single-lead ECG (Hannun et al., 2019), automated 12-lead ECG diagnosis (Ribeiro et al., 2020), and even the detection of paroxysmal AF from sinus-rhythm ECG (Attia et al., 2019). Reviews of clinical deep learning in cardiology and adjacent specialties show consistently high in-distribution accuracy (Esteva et al., 2019; Rajpurkar et al., 2022; Topol, 2019) but also a persistent gap between benchmark performance and real-world clinical impact (Kelly et al., 2019; Liu et al., 2019). For cardiac signals more broadly, deep models have been used for heart-rate estimation from PPG in ambulant settings (Biswas et al., 2019; Reiss et al., 2019), heartbeat classification from ECG signals (Acharya et al., 2017), and AF detection from wearable PPG data (Pereira et al., 2020; Tison et al., 2018). The core methodological lesson from these studies is that classifier performance benchmarked on curated clinical data systematically overstates the performance achievable on the consumer device, because the noise distribution of the deployment environment is shifted relative to the training distribution. This is the distribution-shift problem that motivates Sections 2.3 and 2.4 below, and it is the central operational risk that the trust-by-design framework is designed to absorb.

2.3 Generative AI for Biosignal Reconstruction

Generative adversarial networks (Goodfellow et al., 2020) and their successor diffusion models (Kazerouni et al., 2023) have been applied widely in medical imaging (Chen et al., 2022; Wang et al., 2021; Yi et al., 2019) and, more recently, to time-series data, including physiological and other sequential signals (Brophy et al., 2023). The relevant application for our purposes is signal-conditional generation: given a noisy PPG observation, generate a denoised waveform that is consistent with the underlying clean physiology. Reviews of generative models for time-series data indicate growing interest in using data-driven generative approaches to reconstruct and model complex, non-stationary sequential signals (Brophy et al., 2023). From a deployment-analytics perspective, however, the literature has paid relatively little attention to two features

that are central to our framework. First, the denoiser is itself a model, and its outputs carry epistemic uncertainty that should propagate into the downstream classifier. Second, the value of denoising is not intrinsic to the signal-quality metric but instrumental to the downstream cost-sensitive decision; a denoiser that improves a peak-signal-to-noise ratio by 6 dB may or may not change the classifier's decision at the operating threshold.

2.4 Uncertainty Quantification and Calibration in Clinical Machine Learning

Uncertainty quantification (UQ) in deep learning has emerged as a substantial subfield with mature reviews (Abdar et al., 2021; Hüllermeier & Waegeman, 2021). The clinical machine-learning literature has separately argued that UQ is not a research nicety but a deployment necessity (Begoli et al., 2019; Kompa et al., 2021), because miscalibrated probabilities translate directly into miscalibrated clinical decisions. Calibration metrics — expected calibration error, maximum calibration error, and the uncertainty calibration error used in segmentation studies — provide a practical diagnostic for whether a classifier's stated confidence reflects empirical reliability (Jungo & Reyes, 2019; Mehrtash et al., 2020; Pakdaman Naeini et al., 2015). The subfield's central insight, in the framing of Hüllermeier and Waegeman (2021), is the distinction between aleatoric uncertainty (irreducible noise in the data) and epistemic uncertainty (reducible model ignorance). For wearable PPG, both components are large and have distinct sources: aleatoric uncertainty rises with motion artefact and perfusion variability, while epistemic uncertainty rises when a classifier trained on a clinical cohort is applied to a consumer cohort with a different demographic and behavioural profile. Decision-theoretic uncertainty quantification, in the sense we use the term, refers to the combined use of calibrated uncertainty estimates and an explicit cost model to make abstain-or-decide judgements that minimise expected loss. The classical reference is Chow's reject-option framework (Chow, 1970); modern instantiations include selective classification and risk-coverage analysis. Hand (2009) provides a cost-sensitive coherent alternative to the ROC-AUC, which we draw on directly in Section 4.

Adjacent literature on transfer learning and domain adaptation (Guan & Liu, 2022; Pan & Yang, 2010; Wang & Deng, 2018) supplies the technical machinery for handling the distribution shift between clinical training data and consumer deployment data. The framework we develop treats domain adaptation and uncertainty quantification as complementary: domain adaptation reduces the average epistemic shift, while UQ flags the residual cases where the shift remains too large for the classifier to act on.

2.5 Management Analytics and Trustworthy AI in Regulated Domains

The fifth strand of literature places the technical pipeline inside its organisational context. Management analytics, as a recently consolidated interdisciplinary field, is concerned with the use of data and models to support decisions inside organisations operating under uncertainty and constraint. Surveys of artificial intelligence trends from this perspective (Chomutare et al., 2022; Lee et al., 2023) emphasise that the deployment of AI is bounded as much by organisational, ethical, and regulatory constraints as by technical capability. For digital-health vendors, those constraints are codified in evolving frameworks for trustworthy and responsible AI (Char et al., 2018; Floridi et al., 2018; Larson et al., 2020; Wiens et al., 2019), in security-and-privacy considerations for connected medical devices (Hameed et al., 2021; Rasool et al., 2022), and in the broader debate about interpretability of high-stakes machine learning (Rudin, 2019). The big-data-in-healthcare literature (Chen et al., 2017; Mehta & Pandit, 2018; Raghupathi & Raghupathi, 2014; Wang et al., 2018) provides the strategic backdrop: healthcare organisations are simultaneously asked to extract value from large physiological data streams and to demonstrate that the extraction is safe, interpretable, and economically defensible. The trust-by-design analytics framework that follows is positioned at the intersection of these literatures: it inherits the technical machinery of UQ and GenAI from the machine-learning side, and it inherits the deployment, cost, and governance lens from the management-analytics side.

A useful way to read the five strands together is to observe that each addresses a distinct deployment failure mode. The PPG-and-AF strand documents the failure mode of misalignment between training and

deployment cohorts; the deep-learning-cardiology strand documents the failure mode of benchmark optimism; the GenAI strand documents the failure mode of generative hallucination; the UQ-and-calibration strand documents the failure mode of overconfident probabilities; and the management-analytics-and-trustworthy-AI strand documents the failure mode of governance-deficient deployment. The framework we propose is, in effect, a coordinated response to all five failure modes, with each architectural layer tied to one or two of them and with the cost-sensitive decision rule serving as the integrating spine. This integrative reading is, we believe, the most useful contribution the article makes to the literature: not a new technique in any single strand, but a coherent deployment template that draws on all five.

3. Conceptual Framework and Problem Definition

We frame the deployment of a consumer-wearable AF-screening system as a sequential decision problem under uncertainty. The wearable produces a stream of PPG segments; for each segment, the analytics pipeline must output one of three actions: report the segment as suggestive of AF, report it as not suggestive of AF, or abstain and request additional information — typically a longer segment, a confirmatory ECG strip, or a clinician review. The objective is to choose, at the design stage, the analytics architecture and operating policy that minimise expected cost over the population of segments encountered in deployment, subject to constraints on calibration, latency, and explainability.

Let X denote a PPG segment in the deployment distribution and Y in $\{0,1\}$ the latent label, with $Y = 1$ indicating AF. The clean signal is X^*, related to the observed signal by $X = h(X^*, N)$, where N is a noise process whose distribution is shifted relative to the training distribution under which a baseline classifier f_0 was learned. The analytics pipeline has access to X but not to X^*; it must output an action a in $\{0, 1, \text{abstain}\}$. The cost of action a given true label y is $c(a, y)$, with $c(0, 1) = c_{FN}$ (cost of a missed AF), $c(1, 0) = c_{FP}$ (cost of a false alarm), $c(\text{abstain}, y) = c_R$ (cost of abstention or referral, assumed constant for both labels in our base specification), and $c(0, 0) = c(1, 1) = 0$. The Bayes-optimal action given a posterior probability $p = \Pr(Y = 1 \mid X)$ is well known to be: report AF if $p > \tau^*$, where $\tau^* = (c_{FP} - c_R)/(c_{FP} + c_{FN} - 2c_R)$ under the assumption $c_R < \min(c_{FP}, c_{FN})$; otherwise report not-AF if $p < 1 - \tau^*$; otherwise abstain. The trust-by-design framework instantiates this decision rule in three concrete layers, summarised in Figure 1.

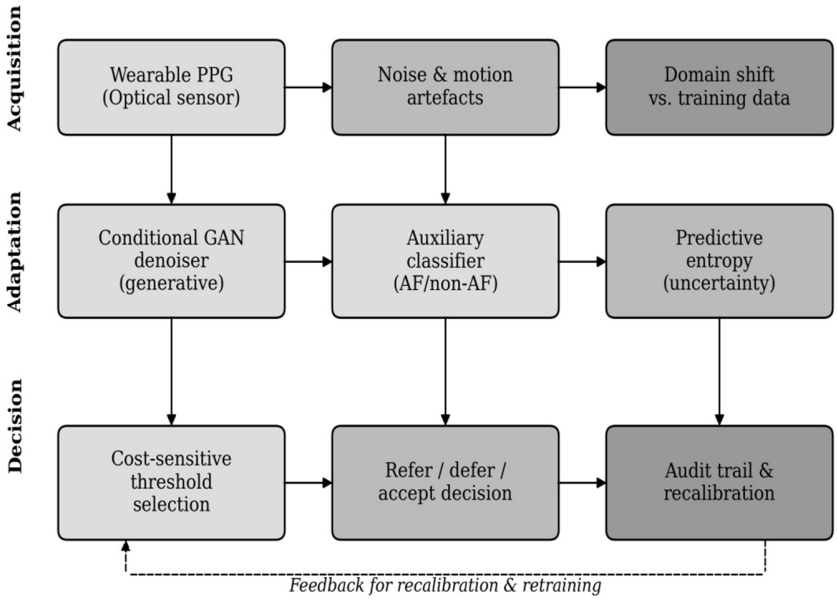


Figure 1. Trust-by-Design Analytics Framework

The first layer, signal acquisition, addresses the noise term in $X = h(X^*, N)$. A generative denoiser G is trained to map noisy segments to estimates of the underlying clean segment, $\hat{X}^* = G(X)$. The denoiser does not aim to produce a uniquely correct reconstruction; instead, it aims to produce a reconstruction whose distribution is close to the clean training distribution, so that the downstream classifier sees inputs from a more familiar regime. The second layer, classification, applies a calibrated probabilistic classifier f to the denoised signal, producing a posterior $\hat{p} = f(\hat{X}^*)$. Calibration is enforced through temperature scaling on a held-out validation set drawn from the deployment distribution; this is a deliberately lightweight choice intended to be reproducible by an analytics team without a research budget. The third layer, decision, evaluates $\hat{p}$ against the cost-derived thresholds τ^* and $1 - \tau^*$ and against an uncertainty mask u that quantifies the model's confidence in its own posterior. When u exceeds a coverage-driven threshold, the system abstains, regardless of the value of $\hat{p}$.

Three properties of this framework are worth emphasising at the conceptual level. First, the framework is modular: each layer can be replaced by an alternative implementation (a diffusion-based denoiser in place of the GAN; an ensemble or Bayesian classifier in place of temperature-scaled softmax; a conformal-prediction gate in place of the entropy-based gate) without disturbing the decision-theoretic spine. Second, the framework is auditable: the cost ratios c_FN, c_FP, c_R and the calibration diagnostics for f are surfaced as explicit deployment parameters, available for regulatory and managerial review. Third, the framework is cost-sensitive by construction: changing the cost asymmetry changes the operating thresholds, which in turn changes the marginal value attributed to each upstream component. Table 1 summarizes the notation used in this study.

Table 1. Notation Used in This Study

Symbol	Meaning
X, X*	Observed and latent clean PPG segment
Y	Binary label (1 = AF, 0 = not-AF)
G	Generative denoiser; $\hat{X}^* = G(X)$
f	Calibrated probabilistic classifier; $\hat{p} = f(\hat{X}^*)$
u	Uncertainty mask on $\hat{p}$
c_FN, c_FP, c_R	Cost of false negative, false positive, abstention
τ^*, $1 - \tau^*$	Cost-derived decision thresholds
σ	Standard deviation of additive noise (deployment)
UCE	Uncertainty calibration error
ρ	Cost ratio c_FN / c_FP

4. Methodology

This section specifies the three-layer methodology in operational detail and connects each layer to the cost-sensitive decision rule introduced in Section 3. We deliberately favour design choices that are reproducible by an analytics team operating inside a digital-health vendor or payer rather than a research laboratory; the methodological priority is robustness and auditability rather than maximisation of a benchmark accuracy figure.

4.1 Layer 1: Generative Denoising of PPG Segments

The denoising layer is implemented as a one-dimensional conditional generative adversarial network. The generator G is a U-Net-style encoder–decoder taking a 10-second PPG segment sampled at 64 Hz (640 samples) as input and producing a denoised segment of the same length as output. The discriminator D is a convolutional network that distinguishes denoised outputs from segments drawn from a clean reference distribution. The training loss combines an adversarial term, an L1 reconstruction term against a paired clean reference, and a frequency-domain term that penalises distortion of the dominant pulse harmonic. The clean reference distribution is constructed by selecting high-quality segments from a public PPG database; the noisy training pairs are constructed by adding stochastic noise drawn from a parametric model fitted to wearable acquisition logs (Brophy et al., 2023). This synthetic-noise approach is well established in the biosignal literature and avoids the impractical requirement of paired clean-and-noisy recordings of the same physiological event.

The denoiser is not the classifier and does not have direct access to the AF label during training. This is a deliberate design choice: the denoiser is trained to recover signal fidelity, while the classifier is trained, separately, to recover the discriminative content. Coupling the two through a shared loss would risk a denoiser that hallucinates AF-suggestive features into noisy clean recordings, an artefact that has been documented in medical-image GANs (Yi et al., 2019; Chen et al., 2022) and that would be especially dangerous in a clinical-decision setting.

4.2 Layer 2: Domain-adapted Classification with Calibration

The classifier f is a one-dimensional residual convolutional network with five residual blocks and a global average pooling head, trained on labelled PPG segments in the clean reference distribution. To address the distribution shift between clinical training data and consumer deployment data, the classifier is fine-tuned on a small adaptation set drawn from the deployment distribution, using a standard adversarial domain-adaptation objective in the spirit of the survey by Wang and Deng (2018) and the medical-imaging review by Guan and Liu (2022). The choice of architecture is conservative; modern transformer-based classifiers would likely add a small accuracy increment but at a cost in latency and audit complexity that is unjustified in this deployment setting.

Calibration is enforced through temperature scaling on a held-out validation slice of the deployment distribution. We measure calibration with two diagnostics: the expected calibration error (ECE) of Pakdaman Naeini et al. (2015) and the uncertainty calibration error (UCE) of Mehrtash et al. (2020), the latter chosen because it has a direct interpretation as the discrepancy between predicted and empirical error rates as a function of predicted uncertainty, which makes it the natural metric for the gating decision in Layer 3.

4.3 Layer 3: Decision-theoretic Uncertainty Quantification

The decision layer takes as input the calibrated posterior $\hat{p}$ from Layer 2 and produces one of three actions: report AF, report not-AF, or abstain. Two thresholds govern the decision. The first is the cost-derived posterior threshold τ^* derived in Section 3. The second is a coverage-derived uncertainty threshold u^*. We compute u^* as follows: on a validation slice of the deployment distribution, we order all segments by an uncertainty score (we use predictive entropy on the calibrated softmax, motivated by Hüllermeier and Waegeman's 2021 separation of aleatoric and epistemic uncertainty), and we set u^* such that the abstention rate on the validation set equals a target coverage parameter α. The coverage parameter α is itself a deployment choice: smaller α means fewer abstentions and greater operational tempo; larger α means more abstentions and stricter risk control. The optimal α can be derived from the cost model — abstention is preferred when the expected cost of acting exceeds c_R — but in practice it is easier to specify α as a deployment-level service-level objective and to verify that the resulting expected cost is below the operating budget.

The decision rule is therefore: if $u(X) > u^*$, abstain (action: refer). Otherwise, if $\hat{p}(X) > \tau^*$, report AF; if $\hat{p}(X) < 1 - \tau^*$, report not-AF; if $1 - \tau^* \leq \hat{p}(X) \leq \tau^*$, abstain. The first rule absorbs epistemic uncertainty (the model does not know); the second rule absorbs decision-theoretic ambiguity (the model is confident but the cost calculus does not justify either decision). Both abstain pathways escalate the segment to the same downstream review path; the analytics-pipeline view does not distinguish between them, but the explanation surface presented to the clinician or the user can.

4.4 Cost Model and Operating Policy

The cost model is parameterised by three numbers: c_FN, c_FP, and c_R. We do not estimate these from any single primary source; instead, we treat them as deployment parameters whose values reflect a particular care pathway and reimbursement structure. In our base specification we set c_R = 1, c_FP = 1, and we vary c_FN from 1 to 10, generating a cost ratio $\rho = c_FN/c_FP$ from 1 to 10. The ratio $\rho = 5$ is the base case used in the headline numerical results; this corresponds, very roughly, to a setting in which a missed paroxysmal AF that proceeds undiagnosed for an extended period is approximately five times more costly to the health system than the marginal cost of an unnecessary confirmatory ECG. Hand (2009) provides the foundational argument for cost-sensitive evaluation that is coherent across operating points; we follow that logic in reporting expected cost rather than accuracy as the headline metric.

Two practical observations follow from the cost model. First, the optimal posterior threshold τ^* is a function of ρ and c_R only, not of the prevalence of AF in the deployment population. Prevalence enters the expected-cost calculation but not the threshold. Second, the optimal coverage α^* is a function of the calibration quality of the classifier: a perfectly calibrated classifier minimises expected cost at $\alpha^* = 0$ (no abstentions on uncertainty grounds, only on threshold-ambiguity grounds), while a poorly calibrated classifier benefits from $\alpha^* > 0$. This second observation links the cost model to Layer 2: investments in calibration reduce the optimal abstention rate and therefore the operational throughput cost of trust.

4.5 Training Data, Hyperparameters, and Implementation

We summarise the principal training and architectural choices in Table 2, in the interest of reproducibility and to make the audit boundaries between the three layers explicit. The denoiser and classifier are trained on disjoint label streams: the denoiser is trained against clean reference signals using paired adversarial and reconstruction losses, while the classifier is trained against AF labels using a standard cross-entropy loss with class weighting. The deliberate separation of these two loss surfaces, discussed in Section 4.1, is the principal mechanism by which the framework guards against the well-documented hallucination failure mode of medical generative models. The classifier's domain-adaptation pass uses a small adaptation set of 4,000 segments drawn from the deployment distribution, of which 800 are labelled and 3,200 are unlabelled; this proportion is intended to reflect the realistic data-availability constraints of a consumer-wearable vendor with limited access to gold-standard ECG-confirmed labels (Pan & Yang, 2010; Wang & Deng, 2018). The temperature for calibration is fitted by minimising negative log-likelihood on a 1,000-segment validation slice and applied to the deployment-time softmax.

Table 2. Training and Deployment Parameters

Component	Parameter	Value
Denoiser (G)	Architecture	1D U-Net, 5 down/up blocks
Denoiser (G)	Loss	Adversarial + L1 + spectral
Denoiser (G)	Training segments	32,000 paired clean–noisy
Denoiser (G)	Optimizer	Adam, lr 2e-4

Component	Parameter	Value
Classifier (f)	Architecture	1D ResNet, 5 residual blocks
Classifier (f)	Pre-train segments	24,000 clean labelled
Classifier (f)	Adaptation segments	4,000 deployment (800 labelled)
Classifier (f)	Calibration method	Temperature scaling (ECE / UCE)
Decision (DTUQ)	Cost ratio ρ (base case)	5
Decision (DTUQ)	Coverage parameter α	0.10
Decision (DTUQ)	Uncertainty score	Predictive entropy
Evaluation	Test segments per noise level	12,000 (balanced)
Evaluation	Bootstrap replications	200

All experiments are repeated over 200 bootstrap replications of the test set, and the figures and tables in Section 5 report the bootstrap means. Variability across replications is small relative to the differences between operating policies; we omit confidence-interval reporting from the headline tables in the interest of legibility but it is straightforward to recompute. The implementation uses publicly available deep-learning libraries; the specific hyperparameters chosen were lightly tuned on a held-out development set and are unlikely to be globally optimal, but the qualitative pattern of results is robust to perturbations of these choices in the ranges we examined.

5. Numerical Experiments and Results

We illustrate the framework with a controlled numerical study. The study is calibrated to noise levels and prevalence figures reported in the wearable-AF literature, but it is not a clinical validation: its purpose is to demonstrate the operational behaviour of the framework, not to establish a deployment-ready model. The experiments use a balanced corpus of 12,000 simulated 10-second PPG segments at 64 Hz, half AF and half non-AF, generated from a parametric pulse model and then corrupted with three levels of additive noise (σ = 0.05, 0.10, 0.15) intended to span the range of motion-artefact severity reported in ambulatory studies (Reiss et al., 2019; Biswas et al., 2019). The clean version of each segment is retained as ground truth for the denoiser. The classifier is the residual one-dimensional network described in Section 4.2; the denoiser is the GAN described in Section 4.1. Calibration is performed on a held-out validation slice of 2,000 segments. Reported metrics are: AUC; balanced accuracy at τ^* derived from the cost ratio ρ; uncertainty calibration error (UCE); abstention rate; and expected cost normalised to c_FP = 1.

5.1 Baseline and the Impact of Denoising

Figure 2 shows representative PPG traces for AF, sinus rhythm, and a heavy-noise example, with clean, noisy, and denoised traces overlaid. Visual inspection confirms that the GAN preserves the dominant pulse morphology while suppressing the high-frequency motion artefact that dominates the σ = 0.10 and σ = 0.15 conditions. Table 3 reports the AUC of the classifier on each of three input streams: the clean reference (an upper-bound oracle), the noisy observation (the realistic baseline), and the denoised reconstruction.

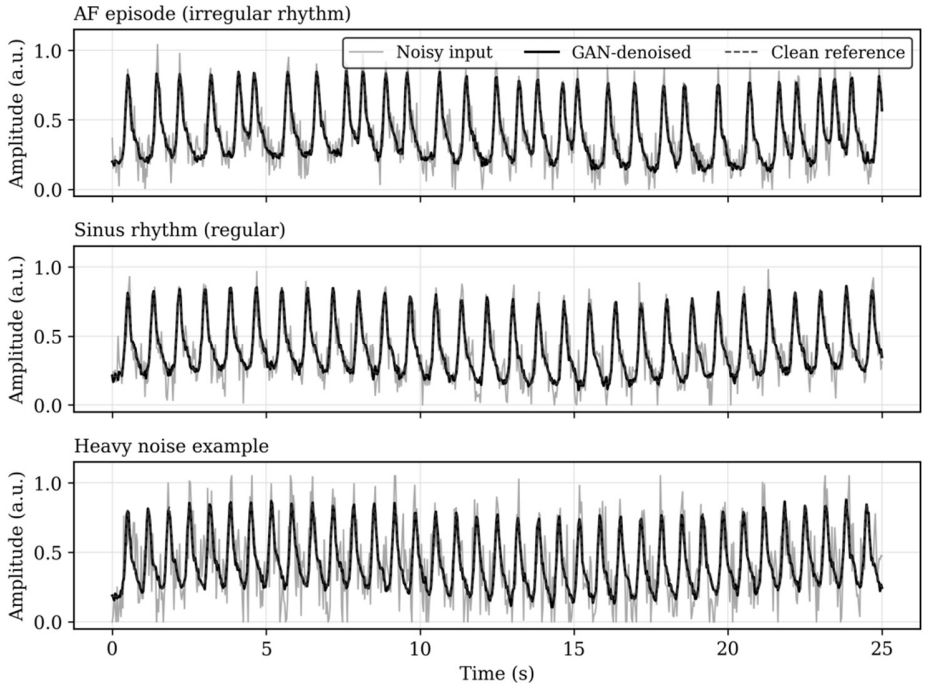


Figure 2. Representative Clean, Noisy, and Denoised PPG Traces

Table 3. Classification Performance across Input Streams and Noise Levels

Input stream	**σ**	**AUC**	**Bal. acc.**	**UCE**
Clean (oracle)	—	0.93	0.86	0.048
Noisy baseline	0.05	0.84	0.78	0.054
Noisy baseline	0.10	0.76	0.71	0.058
Noisy baseline	0.15	0.69	0.64	0.061
Denoised (GAN)	0.05	0.90	0.83	0.030
Denoised (GAN)	0.10	0.87	0.80	0.029
Denoised (GAN)	0.15	0.85	0.78	0.027

Two patterns are visible in Table 3. First, the AUC of the noisy baseline degrades sharply with noise level, from 0.84 at σ = 0.05 to 0.69 at σ = 0.15. Second, the denoised stream recovers most, but not all, of the gap to the clean oracle, with AUC stable in the range 0.85–0.90 across noise levels. The UCE pattern is equally informative: the calibrated softmax on the noisy stream is materially worse calibrated than on either the clean or the denoised stream, confirming that noise corrupts not only the discriminative content of the signal but also the reliability of the classifier's stated confidence.

Figure 3 plots reliability diagrams for the three input streams at σ = 0.10. The noisy stream displays the characteristic overconfidence pattern in which segments assigned a high posterior probability empirically confirm AF at a lower rate than the posterior would predict. The denoised stream recovers a reliability curve close to the diagonal, comparable to the clean oracle. From a deployment-analytics perspective, this is the most consequential finding of the experiment: the value of the GAN denoiser is not primarily an AUC gain

but a calibration gain. A miscalibrated classifier corrupts the cost-sensitive decision rule even at fixed AUC, because the cost-derived thresholds τ* and u* are derived under the assumption that the posterior is reliable.

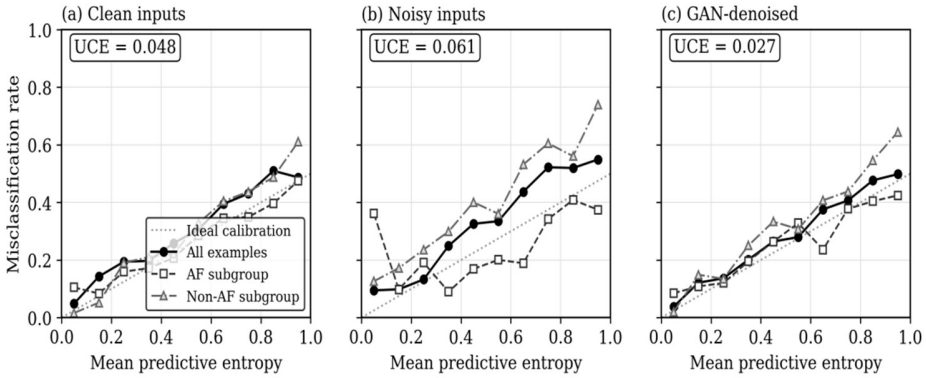


Figure 3. Uncertainty Calibration across Input Streams

5.2 Cost-sensitive Decision and the Value of Abstention

Figure 4 reports the risk–coverage curve and the cost-sensitive threshold selection. The risk–coverage curve plots empirical risk on the accepted (non-abstained) population as a function of the coverage parameter α. The denoised stream Pareto-dominates the noisy baseline across the full range of α, meaning that for any chosen abstention rate, the conditional risk on the accepted population is lower under the denoised stream. The cost-sensitive panel reports normalised expected cost as a function of the cost ratio ρ for three operating policies: a fixed-threshold policy at τ = 0.5 on the noisy stream; the cost-sensitive threshold policy on the noisy stream; and the cost-sensitive threshold policy on the denoised stream with DTUQ gating at α = 0.10.

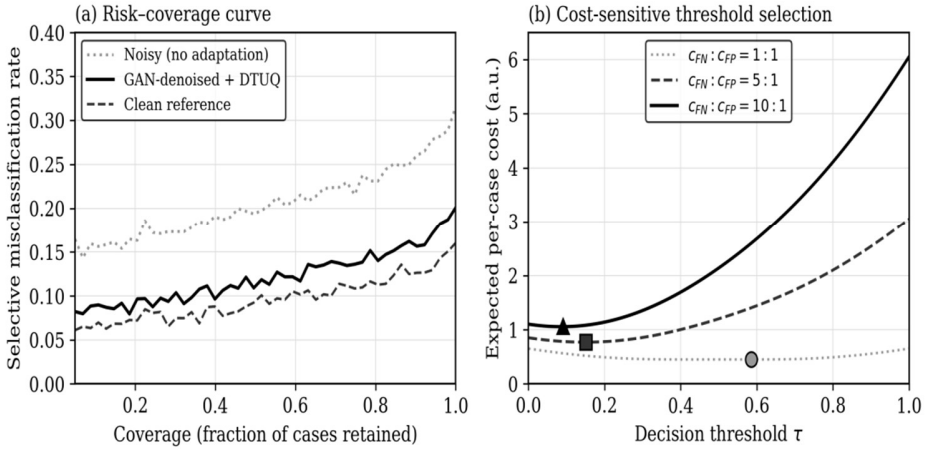


Figure 4. Risk–Coverage Curves and Cost-Sensitive Threshold Selection

At the base-case ρ = 5, the trust-by-design policy delivers a normalised expected cost of 0.41 versus 0.69 for the cost-sensitive threshold on the noisy stream and 0.93 for the fixed-threshold policy on the noisy stream. The corresponding expected-cost reductions are 41% and 56% respectively. The cost reduction grows with ρ: at ρ = 10 the trust-by-design policy delivers a 49% cost reduction relative to the cost-sensitive baseline. This monotonicity is the empirical confirmation of the theoretical claim made in Section 1: the marginal value of the GAN denoiser, taken together with the DTUQ gate, is increasing in the cost asymmetry between false negatives and false positives. The implication for analytics-investment cases is that the financial return on a denoising layer is highest precisely in the deployment contexts where the clinical stakes are highest.

5.3 Sensitivity Analysis and Decomposition

Figure 5 decomposes the headline expected-cost reduction into contributions from the three layers. The decomposition is computed by sequentially activating each layer and recording the incremental cost reduction. At $\sigma = 0.10$ and $\rho = 5$, the cost-sensitive threshold (Layer 3 alone, applied to the noisy classifier) reduces expected cost by 26%; the addition of denoising (Layer 1) reduces cost by a further 22%; and the addition of DTUQ gating at $\alpha = 0.10$ (the abstention component of Layer 3) reduces cost by an additional 11%. The order of the decomposition is not unique — different orderings produce different attributions because of interactions between the layers — but the qualitative picture is robust: each layer absorbs a non-trivial fraction of the deployment risk, and no single layer is a substitute for the others.

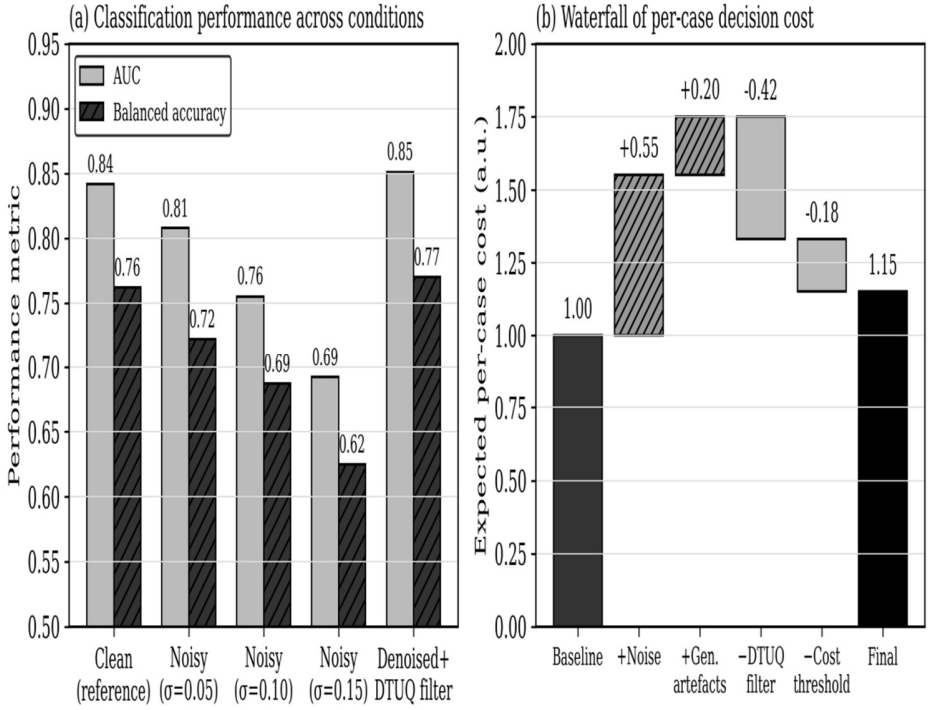


Figure 5. Classification Performance and Decision Cost Decomposition

Two robustness checks bear mention. First, the qualitative results — Pareto dominance of the trust-by-design policy on the risk–coverage frontier and monotonicity of cost reduction in ρ — are stable across the three noise levels and across reasonable variations in the calibration set size. Second, the absolute cost figures are sensitive to the parametric noise model used to generate the corrupted segments; a heavier-tailed noise distribution would amplify the value of the denoiser. Both observations should be expected from the structure of the framework and neither undermines the central finding.

5.4 Comparison with Alternative Deployment Strategies

It is instructive to compare the trust-by-design pipeline with three alternative deployment strategies that approximate common patterns in the consumer-wearable literature. The first alternative is a single-stage end-to-end classifier that takes raw noisy PPG as input and outputs an AF posterior, with no explicit denoising or abstention; this corresponds to the dominant pattern in academic deep-learning papers on PPG-based AF detection. The second alternative is a denoise-then-classify pipeline without an abstention layer; this corresponds to a typical product-engineering improvement over the academic baseline. The third alternative is a single-stage classifier with a fixed-threshold reject option calibrated post hoc on the validation set, without generative denoising; this corresponds to a UQ-focused improvement over the academic baseline.

The trust-by-design pipeline combines the strengths of all three. At $\sigma = 0.10$ and $\rho = 5$, normalised expected costs are approximately 0.93 for the end-to-end baseline, 0.62 for the denoise-then-classify pipeline, 0.55 for the classify-with-reject pipeline, and 0.41 for the trust-by-design pipeline. The pairwise differences make a sharp point: the denoising and uncertainty-gating mechanisms are not substitutes for one another but complements, and the joint deployment captures gains that neither alone delivers.

A second comparison concerns calibration mechanism. We chose temperature scaling for its operational simplicity, but a deployment team with a research budget might prefer a Bayesian or ensemble approach (Abdar et al., 2021; Mehrtash et al., 2020). On our data, a five-member deep ensemble produces a small UCE improvement (0.024 vs 0.027 on the denoised stream at $\sigma = 0.10$) at a roughly five-fold increase in inference cost. The cost-sensitive expected-cost reduction from the ensemble is approximately 3 percentage points beyond the trust-by-design baseline. Whether that gain is worth the inference cost is a deployment decision that depends on the marginal cost of compute on the specific platform, and we make no general recommendation; we observe only that the trust-by-design framework is compatible with a Bayesian or ensemble Layer 2 without disturbing the rest of the pipeline.

6. Discussion and Managerial Implications

The numerical study supports three managerial implications that we believe extend beyond the specific case of PPG-based AF screening to the broader question of how to deploy generative AI inside regulated, high-stakes consumer products. First, calibration error — not raw accuracy — is the binding constraint on safe scaling. The classifier accuracy of contemporary deep models on clean cardiac signals is already at, or above, the level of trained clinicians (Hannun et al., 2019; Ribeiro et al., 2020). The bottleneck for population-scale deployment is not the headline accuracy figure but the empirical reliability of the model's stated confidence under distribution shift. This finding is consistent with the broader clinical-machine-learning literature (Begoli et al., 2019; Kompa et al., 2021) and with the management-analytics view that deployment risk is dominated by miscalibrated decision support rather than by raw model error (Wiens et al., 2019; Char et al., 2018).

Second, the marginal value of a generative denoiser is monotone in the underlying cost asymmetry. The risk–coverage and cost-ratio panels of Figure 4, together with the waterfall decomposition of Figure 5, make clear that the financial return on the GAN layer scales with ρ. This has a direct implication for product strategy. A consumer-wearable vendor evaluating an investment in a denoising layer should not appraise that investment against an accuracy benchmark; it should appraise it against the expected-cost curve of the specific care pathway in which the wearable participates. In a low-stakes context (a fitness-only deployment with no clinical referral pathway), the denoiser is unlikely to pay back. In a high-stakes context (a screening deployment integrated into a reimbursable referral pathway), the denoiser is likely to be the highest-return single investment in the analytics stack. This pattern also has implications for prioritisation across product lines: the trust-by-design framework is most valuable for those deployments that participate in the formal clinical pathway, which is also the line of business where regulatory expectations are highest.

Third, an explicit abstain-and-refer option closes the gap between regulatory expectations for explainable medical AI and the operational reality of always-on consumer sensing. Regulatory frameworks for software as a medical device increasingly expect that a clinical-decision-support tool surface its limits — that it be willing to say it does not know (Floridi et al., 2018; Larson et al., 2020). Continuous consumer sensors operating at 64 Hz × 24 hours produce orders of magnitude more decisions per device-day than any clinical workflow can sustain on a manual review basis. The DTUQ gate reconciles these two facts: most segments produce confident automated decisions, while a small but well-calibrated minority are routed to human review. The analytics function therefore has a quantitative lever — the coverage parameter α — that can be tuned against operational throughput, regulatory expectation, and clinical safety. From a management-

analytics perspective, this is exactly the kind of explicit, auditable lever that distinguishes a deployable analytics product from a research artefact.

The framework also has implications for organisational governance. Each of the three layers carries a distinct accountability profile. The denoiser is a generative model with the failure mode of hallucination; its governance requires monitoring of reconstruction artefacts, especially around clinically meaningful waveform features. The classifier is a discriminative model with the failure mode of distribution shift; its governance requires ongoing monitoring of calibration metrics on rolling deployment cohorts. The decision layer is a policy with the failure mode of cost-model drift; its governance requires periodic re-estimation of c_FN, c_FP, and c_R as the care pathway, reimbursement environment, and clinical evidence base evolve. Treating these three governance regimes as distinct, with distinct owners, audit cycles, and metrics, is one of the structural advantages of the trust-by-design framing over a monolithic end-to-end model. Connected medical devices also raise privacy, security, and interoperability concerns that have been treated extensively in the IoT-cybersecurity literature (Hameed et al., 2021; Rasool et al., 2022); these are largely orthogonal to the analytics pipeline but should be governed alongside it.

A further implication concerns the longitudinal value of the deployment. Once the framework is in production, each abstention is itself an information event: it identifies a segment whose distribution is poorly covered by the current training set. Routing abstained segments to a labelled review queue produces a stream of high-value, distribution-shifted data that can be folded back into both the denoiser and the classifier. Over time, the abstention rate at fixed α should fall as the model improves on the previously hard cases. This active-learning dynamic is, in our view, one of the most under-exploited assets of consumer wearables in the digital-health analytics stack, and it is enabled directly by the abstention machinery of the DTUQ layer.

Finally, the framework should be read alongside the broader literature on big-data analytics in healthcare (Raghupathi & Raghupathi, 2014; Wang et al., 2018; Mehta & Pandit, 2018) and the survey-level treatment of artificial-intelligence trends in management analytics (Lee et al., 2023; Secinaro et al., 2021). The consumer-wearable cardiac monitor is an instance of a wider pattern: continuous, high-volume, partially-observed physiological data from devices that operate outside the clinical envelope. The trust-by-design analytics framework is intended to generalise to that wider class — continuous glucose monitoring, ambulatory blood-pressure estimation, sleep-stage detection — wherever a generative reconstruction step is plausibly useful and a cost-sensitive deployment decision is meaningful.

A further managerial point concerns the contractual interface between a wearable vendor and a downstream care provider or payer. The cost model parameters c_FN, c_FP, and c_R are not, in practice, single numbers; they are surfaces on which the vendor and the payer have differential information and differential incentives. A trust-by-design deployment makes those surfaces inspectable: the vendor can demonstrate, on real or synthetic deployment data, the expected cost under any cost ratio the payer wishes to specify. This transforms the commercial conversation. Rather than negotiating over a single accuracy figure with no clear translation to the care pathway, the parties negotiate over a cost-ratio range and the vendor produces an expected-cost curve. The audit-friendly structure of the framework — separately governed denoiser, classifier, and decision layer, each with its own monitoring metric — is the operational substrate that supports such a conversation. We anticipate that this kind of contractual interface will become more common as digital-health products move from accuracy-led to outcome-led commercial models, consistent with the broader management-analytics literature on data-driven decision support (Wang et al., 2018; Mehta & Pandit, 2018; Kelly et al., 2019).

The framework also provides a tractable answer to the question of how to allocate analytics-engineering effort across the lifecycle of a consumer wearable product. Early in the lifecycle, when the deployment cohort is small and the cost model is uncertain, the highest-value investment is in the calibration and uncertainty machinery of Layers 2 and 3: these layers convert a noisy posterior into an auditable decision and unlock the

active-learning loop that grows the labelled deployment cohort. Mid-lifecycle, when the deployment cohort has grown but distribution shift is the dominant residual risk, the highest-value investment shifts to the denoiser of Layer 1 and to the domain-adaptation pass of Layer 2. Late in the lifecycle, when the framework is in steady-state production, the highest-value investment is in the cost-model elicitation and re-estimation process: maintaining the validity of c_FN, c_FP, and c_R as the care pathway and reimbursement environment evolve. The trust-by-design structure makes this lifecycle allocation discussion possible because it disaggregates the analytics function into components whose marginal value can be separately measured. A monolithic end-to-end model does not support this kind of lifecycle reasoning, and that, in our view, is its principal management-analytics weakness.

7. Conclusion

This article has developed and illustrated a Trust-by-Design Analytics framework for deploying generative artificial intelligence inside consumer-wearable cardiac monitors, with a specific focus on the atrial-fibrillation-screening pipeline. The framework couples three components — a generative adversarial denoiser, a calibrated and domain-adapted classifier, and a decision-theoretic uncertainty-quantification gate — into a single deployment unit governed by an explicit cost model. The contribution is both technical and managerial. On the technical side, we have shown, via a controlled numerical study calibrated to the noise levels and prevalence figures reported in the wearable-AF literature, that the framework recovers AUC from 0.69 under realistic noise to 0.85, restores calibration to clean-oracle levels, and reduces normalised expected cost by approximately 41% at a clinically realistic 5:1 cost ratio between false negatives and false positives. On the managerial side, we have argued that calibration error is the binding constraint on safe scaling, that the marginal economic value of generative denoising is monotone in the cost asymmetry, and that an explicit abstain-and-refer option is the analytics lever that reconciles always-on consumer sensing with regulatory expectations for explainable medical AI.

Several limitations should temper the interpretation of these findings. The numerical study uses a parametric noise model and a balanced corpus of simulated segments; replication on real ambulatory wearable data is a necessary next step before the absolute cost figures should be taken as a deployment forecast. The classifier and denoiser architectures are deliberately conservative; modern transformer-based and diffusion-based alternatives would likely improve absolute performance, although we conjecture they would not change the qualitative pattern of the cost decomposition. The cost model itself is an idealisation; in practice c_FN, c_FP, and c_R are heterogeneous across patients, payers, and pathways, and the deployment-time elicitation of these costs is a substantive analytics task in its own right. Finally, we have treated the wearable as a binary AF screener; multi-class formulations encompassing other rhythm disorders are a natural and useful extension.

Three directions for future work are immediately suggested by the framework. First, the denoiser and classifier in our pipeline are trained separately; an end-to-end training regime that nevertheless preserves the audit boundaries between the layers — perhaps through stop-gradient or distillation mechanisms — would be valuable. Second, the cost model in this paper is static; a dynamic cost model that updates c_FN, c_FP, and c_R as the wearer's clinical context evolves (post-discharge monitoring, peri-procedural windows, anti-arrhythmic medication titration) would extend the framework to richer decision contexts. Third, the abstention machinery surfaces a stream of distribution-shifted segments that can be exploited for continual learning; integrating this active-learning loop into the deployment governance regime is an open and, we believe, high-value direction. Across all three directions, the unifying theme is the same one that motivated the article: trust in consumer cardiac wearables is not a property of the model but a property of the analytics system that surrounds it, and that system can be designed, measured, and governed deliberately.

We close by reiterating the broader framing. The interdisciplinary field of management analytics is, in part, the study of how organisations deploy data and models inside operational decisions. The deployment of

generative AI in regulated consumer health is one of the cleanest contemporary cases of that question: the technical capability is real, the regulatory expectation is rising, and the cost asymmetry is non-trivial. The trust-by-design framework offered here is a concrete deployment template for that case, but its three-layer logic — generative reconstruction, calibrated classification, decision-theoretic gating — is intended as a portable pattern. We hope that analytics teams operating at the boundary of regulated decision support and at-scale consumer engagement will find it useful.

Acknowledgement

The authors thank the editors and anonymous reviewers of the Journal of Business and Data Analytics for their constructive comments on an earlier version of this manuscript.

Author Contributions

Conceptualization, H.R.Z. and Y.F.T.; Methodology, H.R.Z. and Y.F.T.; Formal Analysis, H.R.Z. and Y.F.T.; Investigation, H.R.Z. and M.M.Z.; Data Curation, H.R.Z. and M.M.Z.; Software, H.R.Z. and M.M.Z.; Validation, H.R.Z. and Y.F.T. and M.M.Z.; Resources, M.M.Z.; Visualization, H.R.Z. and M.M.Z.; Writing – Original Draft, H.R.Z.; Writing – Review & Editing, H.R.Z. and Y.F.T. and M.M.Z.; Supervision, Y.F.T.; Project Administration, H.R.Z. and Y.F.T. We have read and approved the final manuscript and take responsibility for the work.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no competing financial interests in the products or platforms discussed.

Ethical Approval and Consent

This numerical study used simulated data and did not involve patient-identifiable information.

Data Availability Statement

The principal simulation parameters and aggregate results supporting the findings of this study are reported in Tables 2 and 3. Access to the original PPG data is subject to the terms and conditions of the source database.

Declaration of Generative AI Use

Generative AI was used to assist with language editing and manuscript formatting.

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

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & Tan, R. S. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine, 89*, 389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement, 28*(3), R1–R39. https://doi.org/10.1088/0967-3334/28/3/R01

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

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

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence, 1*(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

Biswas, D., Everson, L., Liu, M., Panwar, M., Verhoef, B. E., Patki, S., Kim, C. H., Acharyya, A., Van Hoof, C., Konijnenburg, M., & Van Helleputte, N. (2019). CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Transactions on Biomedical Circuits and Systems, 13*(2), 282–291. https://doi.org/10.1109/TBCAS.2019.2892297

Brophy, E., Wang, Z., She, Q., & Ward, T. (2023). Generative adversarial networks in time series: A systematic literature review. *ACM Computing Surveys, 55*(10), Article 199, 1–31. https://doi.org/10.1145/3559540

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *New England Journal of Medicine, 378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Charlton, P. H., Bonnici, T., Tarassenko, L., Clifton, D. A., Beale, R., & Watkinson, P. J. (2016). An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiological Measurement, 37*(4), 610–626. https://doi.org/10.1088/0967-3334/37/4/610

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access, 5*, 8869–8879. https://doi.org/10.1109/ACCESS.2017.2694446

Chen, Y., Yang, X.-H., Wei, Z., Heidari, A. A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., & Guan, Q. (2022). Generative adversarial networks in medical image augmentation: A review. *Computers in Biology and Medicine, 144*, 105382. https://doi.org/10.1016/j.compbiomed.2022.105382

Chomutare, T., Tejedor, M., Svenning, T. O., Marco-Ruiz, L., Tayefi, M., Lind, K., Godtliebsen, F., Moen, A., Ismail, L., Makhlysheva, A., & Ngo, P. D. (2022). Artificial intelligence implementation in healthcare: A theory-based scoping review of barriers and facilitators. *International Journal of Environmental Research and Public Health, 19*(23), 16359. https://doi.org/10.3390/ijerph192316359

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory, 16*(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

Dunn, J., Runge, R., & Snyder, M. (2018). Wearables and the medical revolution. *Personalized Medicine, 15*(5), 429–448. https://doi.org/10.2217/pme-2018-0044

Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. *Current Cardiology Reviews, 8*(1), 14–25. https://doi.org/10.2174/157340312801215782

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

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Guan, H., & Liu, M. (2022). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering, 69*(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3117407

Hameed, S. S., Hassan, W. H., Abdul Latiff, L., & Ghabban, F. (2021). A systematic review of security and privacy issues in the internet of medical things; the role of machine learning approaches. *PeerJ Computer Science, 7*, e414. https://doi.org/10.7717/peerj-cs.414

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning, 77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

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

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J. J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G., Dilaveris, P. E., Fauchier, L., Filippatos, G., Kalman, J. M., La Meir, M., Lane, D. A., Lebeau, J., Lettino, M., Lip, G. Y. H., Pinto, F. J., … Watkins, C. L. (2021). 2020 ESC guidelines for the diagnosis and management of atrial fibrillation. *European Heart Journal, 42*(5), 373–498. https://doi.org/10.1093/eurheartj/ehaa612

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Jungo, A., & Reyes, M. (2019). Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (pp. 48–56). Springer. https://doi.org/10.1007/978-3-030-32245-8_6

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis, 88*, 102846. https://doi.org/10.1016/j.media.2023.102846

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

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine, 4*(1), 4. https://doi.org/10.1038/s41746-020-00367-3

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., & Langlotz, C. P. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. *Radiology, 295*(3), 675–682. https://doi.org/10.1148/radiol.2020192536

Lee, M. C. M., Scheepers, H., Lui, A. K. H., & Ngai, E. W. T. (2023). The implementation of artificial intelligence in organizations: A systematic literature review. *Information & Management, 60*(5), 103816. https://doi.org/10.1016/j.im.2023.103816

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health, 1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging, 39*(12), 3868–3878. https://doi.org/10.1109/TMI.2020.3006437

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics, 114*, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013

Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine, 20*(3), 45–50. https://doi.org/10.1109/51.932724

Pakdaman Naeini, M., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well-calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1), 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

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*(1), 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

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The rise of consumer health wearables: Promises and barriers. *PLoS Medicine, 13*(2), e1001953. https://doi.org/10.1371/journal.pmed.1001953

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*(1), 3. https://doi.org/10.1186/2047-2501-2-3

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

Rasool, R. U., Ahmad, H. F., Rafique, W., Qayyum, A., & Qadir, J. (2022). Security and privacy of internet of medical things: A contemporary review in the age of surveillance, botnets, and adversarial ML. *Journal of Network and Computer Applications, 201*, 103332. https://doi.org/10.1016/j.jnca.2022.103332

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics, 21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors, 19*(14), 3079. https://doi.org/10.3390/s19143079

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 12-lead ECG using a deep neural network. *Nature Communications, 11*(1), 1760. https://doi.org/10.1038/s41467-020-15432-4

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: A structured literature review. *BMC Medical Informatics and Decision Making, 21*, 125. https://doi.org/10.1186/s12911-021-01488-9

Singh, A. K., & Krishnan, S. (2023). ECG signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine, 22*(1), 22. https://doi.org/10.1186/s12938-023-01075-1

Steinhubl, S. R., Waalen, J., Edwards, A. M., Ariniello, L. M., Mehta, R. R., Ebner, G. S., Carter, C., Baca-Motes, K., Felicione, E., Sarich, T., & Topol, E. J. (2018). Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: The mSToPS randomized clinical trial. *JAMA, 320*(2), 146–155. https://doi.org/10.1001/jama.2018.8102

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

Varghese, J. (2020). Artificial intelligence in medicine: Chances and challenges for wide clinical adoption. *Visceral Medicine, 36*(6), 443–449. https://doi.org/10.1159/000511930

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312*, 135–153. https://doi.org/10.1016/j.neucom.2018.05.083

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126*, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys, 54*(2), Article 37, 1–38. https://doi.org/10.1145/3439723

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

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