From Black Box to Accountable Sensor: A Reliability-Aware Generative AI Framework for Public-Facing Wearable Health Devices

Nadia Farhana Rahman[1], Amirul Hakim Ismail[2], Kavitha Nair[3,*]

[1] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia

[2] Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia

[3] Faculty of Electronics and Computer Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

* Email: *kavitha.nair@utem.edu.my* (Corresponding Author)

Abstract

Consumer wearables increasingly combine biosignal sensing, generative signal enhancement, and automated health interpretation. This convergence creates a new socio-technical problem: a device that appears to be a simple sensor may actually contain a generative model that edits, completes, or denoises the signal before a downstream classifier or large language interface turns it into advice. When the generated signal is reliable, such adaptation may improve public access to early health warnings. When it is unreliable, the same mechanism may transform a noisy input into a confident but misleading output. This article develops a Reliability-Aware Generative AI (RA-GAI) framework for public-facing wearable health devices. The framework combines signal quality assessment, generative adaptation, decision-theoretic uncertainty, action gating, and an accountability layer that records why a device produced, withheld, or escalated a health interpretation. Using a photoplethysmography-oriented wearable scenario, the article presents a transparent reliability simulation that compares raw inference, ungated generative enhancement, uncertainty-gated enhancement, and the full accountability framework across noise, domain shift, and deployment-risk conditions. The results indicate that generative adaptation can improve the apparent performance of health inference, but its public safety value depends on whether uncertainty is translated into action-specific controls. The proposed framework improves accepted-window accuracy, reduces unsafe automatic alerts, and clarifies when a device should return a user-facing caution rather than a diagnosis-like claim. The contribution is twofold: technically, the article converts uncertainty from a model-internal statistic into an operational reliability signal; socially, it defines accountable sensing as a governance requirement for wearable technologies that intervene directly in everyday health decisions.

Keywords: Generative AI; Wearable Health Devices; Photoplethysmography; Uncertainty Quantification; Accountability; Public-Facing Sensors; Digital Health Governance

Article History

Received: February 15, 2024

Revised: March 20, 2024

Accepted: May 28, 2024

Available Online: June 30, 2024

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

From Black Box to Accountable Sensor: A Reliability-Aware Generative AI Framework for Public-Facing Wearable Health Devices

1. Introduction

Wearable health devices have moved from specialised clinical accessories into the ordinary routines of public life. A wristband can count sleep events, a watch can warn about an irregular pulse, and a patch can stream physiology during exercise or recovery. This shift changes the social status of sensing. The sensor is no longer only a measurement component; it becomes an everyday health intermediary that converts noisy bodily traces into advice, reassurance, or alarm. The public value of this transition is considerable, because continuous monitoring can widen access to early warning and self-management. Yet the same transition creates a fragile dependency on automated interpretation in contexts where users may lack clinical training, device placement may be poor, and the decision environment may be emotionally charged (Piwek et al., 2016; Pevnick et al., 2018; Meskó et al., 2017; Mishra et al., 2020).

The next generation of wearables is likely to contain not only discriminative classifiers, but also generative models that transform signals before interpretation. Generative AI can denoise photoplethysmography (PPG), complete missing segments, align signals from a new device to the training domain of an existing classifier, and produce user-facing explanations in natural language. These functions are attractive because wearable data are messy by design: motion artifacts, loose straps, ambient light interference, skin optical differences, and low-cost sensors create frequent domain gaps between laboratory training data and public deployment data. The technical literature shows that physiological signals and ambulatory cardiac data can support powerful inference when models are trained and evaluated carefully (Allen, 2007; Tamura et al., 2014; Hannun et al., 2019; Perez et al., 2019).

The risk is that generative adaptation may make a device look more reliable than it is. A black-box classifier may already be difficult to audit, but a black-box sensor that edits the signal before classification is harder still. The user sees a single outcome, while the true pipeline includes sensing, signal-quality checking, generative transformation, risk scoring, thresholding, language generation, and interface design. If any layer fails, the resulting advice may be confidently wrong. This concern is not theoretical. Medical machine learning systems can rely on shortcuts, behave differently under dataset shift, and produce clinically meaningful errors even when average performance seems acceptable (Oakden-Rayner et al., 2020; DeGrave et al., 2021; Roberts et al., 2021).

The central question addressed in this article is therefore not whether generative AI can improve wearable health inference in a narrow benchmark. It is whether a public-facing wearable can be designed as an accountable sensor, where uncertainty, signal quality, generative transformation, and user-facing action are explicitly connected. This framing differs from conventional model evaluation. Accuracy, AUC, and F1 score are useful but insufficient when a device must decide whether to display a warning, request a retest, recommend professional review, or refuse to make a claim. A reliability-aware device must attach uncertainty to actions and harms, not merely to probabilities (Begoli et al., 2019; Kompa et al., 2021; Abdar et al., 2021).

This article proposes a Reliability-Aware Generative AI (RA-GAI) framework for public-facing wearable health devices. The framework builds on decision-theoretic uncertainty, medical AI governance, wearable biosignal engineering, and socio-technical accountability. It treats generative signal adaptation as an intervention inside the sensing pipeline and requires every health interpretation to pass through an action selector that weighs reliability, expected utility, and user harm. This design is especially important for consumer interfaces, where a message may influence anxiety, care-seeking, self-treatment, or delayed consultation (Char et al., 2018; Vayena et al., 2018; Kerasidou, 2020).

The article makes three contributions. First, it converts the idea of uncertainty-aware generative adaptation into a broader accountable-sensor architecture suitable for public-facing wearables. Second, it introduces a decision-

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

oriented reliability evaluation that distinguishes technical improvement from safe user-facing action. Third, it presents a transparent data-analysis demonstration showing how reliability thresholds alter coverage, accuracy, calibration, unsafe alert rate, and expected utility. The aim is not to claim clinical validation, but to show how wearable GenAI systems should be evaluated before they are allowed to shape everyday health decisions.

2. Related Work

Wearable health research has demonstrated that continuously collected signals can support clinically relevant inference, particularly in cardiovascular monitoring. PPG is attractive because it can be embedded in wrist devices and other low-burden form factors. The signal is related to blood volume variation and can encode pulse timing, morphology, and rhythm irregularity. Prior work on PPG analysis highlights both the promise and the fragility of the modality: artifacts, sensor placement, perfusion, and algorithm choice strongly affect interpretation (Allen, 2007; Elgendi, 2012; Schäfer and Vagedes, 2013). Consumer-device validation studies also show that heart-rate estimates can be accurate in some settings but degraded by activity, device design, and user variation (Shcherbina et al., 2017; Nelson and Allen, 2019; Bent et al., 2020).

Deep learning has increased the performance ceiling for biosignal classification. Ambulatory electrocardiogram models can reach expert-level arrhythmia classification under appropriate validation, and AI-enabled ECG systems have shown that latent risk can sometimes be detected before overt events are visible to clinicians (Hannun et al., 2019; Attia et al., 2019; Liu et al., 2019). Large-scale smartwatch studies further demonstrate that consumer devices can identify possible atrial fibrillation at population scale, although the pathway from alert to diagnosis requires confirmatory assessment and careful workflow design (Perez et al., 2019). These findings support the idea that wearables can extend detection, but they also show why public-facing interpretation must remain a screening function rather than a diagnosis-like promise.

Generative models add another layer to this landscape. In wearable pipelines, generative AI may denoise PPG, reconstruct missing intervals, simulate physiological states for testing, or translate a low-quality signal representation into one that resembles a training domain. Foundation models and large language models extend this logic by producing explanations and interacting with users. Generalist medical AI has the potential to integrate heterogeneous data, while language models can encode clinical knowledge and support health communication (Moor et al., 2023; Singhal et al., 2023; Thirunavukarasu et al., 2023). Yet generative models are also associated with hallucination, overconfidence, and the production of plausible but unsupported content, especially when evaluation focuses on fluency or average benchmark scores rather than grounded utility (Bender et al., 2021).

Uncertainty quantification offers a partial solution, but only if it is treated as operational evidence rather than decorative confidence. Surveys of deep uncertainty distinguish aleatoric uncertainty, epistemic uncertainty, calibration, ensembles, Bayesian approximations, and distributional robustness (Gawlikowski et al., 2023; Abdar et al., 2021). In health care, however, communicating uncertainty is not simply a mathematical task. The user and clinician need to know whether the uncertainty should change the action. A low confidence output might require repetition of a measurement, a clinician review recommendation, or an explicit abstention. A device that displays an uncertain result as a polished sentence may be less safe than a simpler device that refuses to answer under poor signal conditions (Begoli et al., 2019; Kompa et al., 2021).

Explainability research is similarly useful but incomplete. Local explanation methods can identify features that influence a classifier, and surveys of black-box explanation methods provide a large toolbox for algorithmic transparency (Ribeiro et al., 2016; Guidotti et al., 2018). At the same time, scholars warn that explanation can create a false sense of understanding when a model is still unstable, biased, or not clinically validated (Lipton, 2018; Ghassemi et al., 2021). For high-stakes decisions, some argue that interpretable models should be preferred when possible rather than using explanations to justify opaque models (Rudin, 2019). In the wearable GenAI setting, an explanation must be tied to signal reliability, model version, and action gate; otherwise, it may simply

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

rationalize an unsafe output.

Governance research provides the broader lens. Responsible medical AI requires prospective evaluation, clear intended use, bias monitoring, privacy protection, clinician involvement, and ongoing post-deployment review (Kelly et al., 2019; Rajkomar et al., 2019; Reddy et al., 2020). Ethical AI guidelines and governance principles have proliferated, but principles alone are insufficient without concrete implementation pathways (Jobin et al., 2019; Mittelstadt, 2019). For public-facing wearables, accountability must be embedded in the device pipeline itself. This article therefore combines technical reliability metrics with a governance layer that can be audited by manufacturers, regulators, clinicians, and users.

Research on AI, cyber-physical systems, and IoT also helps situate the problem. Wearable health devices are cyber-physical systems because they link physical measurement, embedded computation, networked data flows, and user-facing decisions. Reviews of Industry 4.0 and cyber-physical systems emphasize that sensor intelligence must be evaluated at the level of integrated systems, not isolated algorithms (Lu, 2017). More general AI surveys and IoT cybersecurity research further show that connected intelligent devices require security, interoperability, and lifecycle governance alongside model performance (Lu, 2019; Lu and Xu, 2019; Xu et al., 2021). This perspective is central to accountable sensing because a reliable health output depends on secure data collection, trustworthy model updates, and accountable system design.

3. From Black Box Model to Accountable Sensor

A black-box model is opaque because its internal reasoning is difficult to inspect. An accountable sensor is different: it is a socio-technical system that makes visible the conditions under which a health interpretation was produced. The accountable sensor records the quality of the input, the transformation applied to the input, the uncertainty associated with inference, the action selected, the wording presented to the user, and the governance evidence retained for later review. This moves accountability upstream from the final prediction to the entire sensing chain.

The distinction matters because public users often encounter health wearables as objects of trust. A polished interface can turn a weak signal into a strong impression of medical authority. If a device says that an irregular rhythm was detected, the user may infer that the measurement was good, the model was robust, and the claim was clinically meaningful. These assumptions may be false. The device may have measured a loose strap, translated a noisy PPG segment through a generative denoiser, and produced a classifier probability that is poorly calibrated for that context. Accountable sensing requires the device to disclose the limitations of the pipeline in a useful form.

The RA-GAI framework begins from a simple principle: generative adaptation should never be treated as invisible preprocessing. It is a model-based intervention that can improve or degrade downstream inference. In a public-facing device, that intervention should be logged, measured, and linked to action. Table 1 translates the core technical risks into accountability requirements. The table also shows why this article uses the term accountable sensor rather than trustworthy model. The unit of analysis is the device-user decision environment.

Problem in generative wearable AI	Accountable-sensor response	Why it matters for public-facing devices
Noisy or shifted biosignals are transformed by a generative model before inference.	Store signal quality state and generative model version for every output.	Users and auditors can distinguish sensor failure from model-induced transformation.
Generated waveforms may contain plausible but misleading artifacts.	Attach reliability score, uncertainty decomposition, and action gate.	A device can withhold or escalate high-risk interpretations rather than hiding uncertainty.
Standard accuracy metrics do not encode downstream harm.	Use decision-theoretic utility with harm-sensitive costs.	False reassurance, false alarm, and delayed care receive different costs.
User advice may look authoritative even when evidence is weak.	Require plain-language confidence statements and escalation prompts.	A wearable becomes a decision-support interface, not a black-box diagnostician.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

4. Reliability-Aware Generative AI Framework

The proposed framework is organised around five layers: sensing, quality assessment, generative adaptation, decision-theoretic reliability, and accountability. The sensing layer collects the raw signal and context metadata such as movement, contact quality, sampling irregularity, ambient light exposure, and device battery state. The quality layer converts these raw indicators into a signal quality score. The generative adaptation layer then decides whether denoising, completion, or domain alignment is warranted. The inference layer produces a health risk score. Finally, the decision layer translates reliability into action: display, request a retest, recommend professional review, or abstain.

Figure 1 summarises the architecture. The arrows are important because reliability is not a one-way calculation. A poor signal can prevent generative adaptation. A high-risk inference can trigger stricter uncertainty requirements. A user-facing explanation can be limited by the quality of the evidence trail. A governance dashboard can feed back into model updates and threshold changes. This loop reflects a public-facing reality: the device is not only predicting a state; it is deciding how to intervene in a user's interpretation of their body.

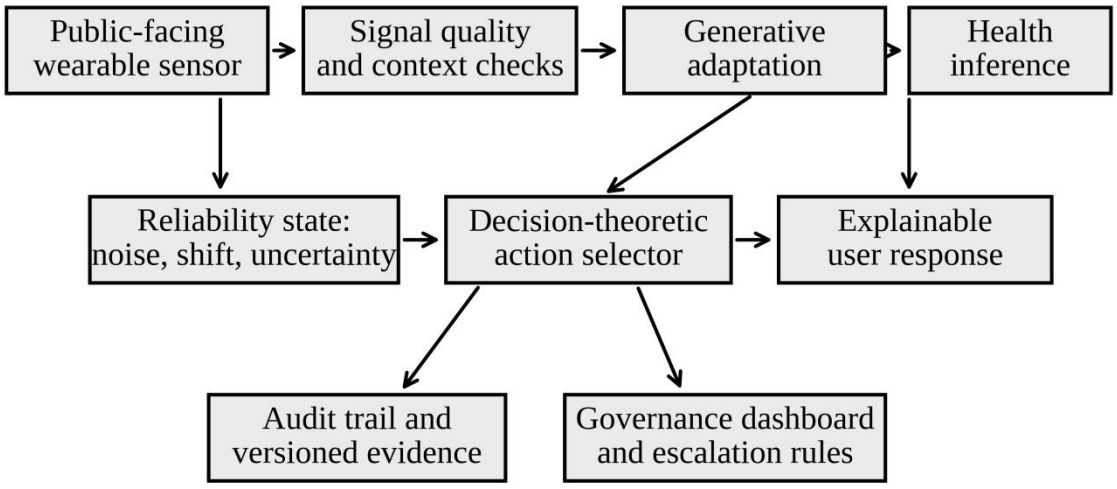

Figure 1. Reliability-aware accountable-sensor architecture for public-facing wearable health devices.

The architecture intentionally separates signal transformation from user-facing interpretation. In conventional pipelines, a denoiser and a classifier may be evaluated as a single black-box stack. In the accountable-sensor pipeline, each stage leaves evidence for later review. This structure is particularly important when the device uses a generative model to transform signals that users never see.

5. Decision-Theoretic Reliability Model

Reliability-aware sensing requires a formal link between prediction and action. Let x_i denote an observed wearable signal window and $g\theta(x_i)$ denote the output of a generative adaptation model. A downstream model $f\varphi$ produces a risk score $p_i = f\varphi(g\theta(x_i))$. The system also estimates uncertainty u_i and signal quality q_i. Instead of treating p_i as the final product, the framework selects an action a_i from a finite set: automatic display, retest request, clinician-review recommendation, or abstention. The best action is the one that maximises expected

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

utility under harm-sensitive costs.

The expected utility can be written as $U(a_i \mid x_i) = \Sigma_\gamma P(y \mid x_i)B(a_i, y) - C(a_i, y) - \lambda H(a_i, u_i, q_i)$, where B represents benefit, C represents direct cost, and H penalises harm from unreliable action. This formulation matters because false positives and false negatives are not symmetric in consumer health. A false alarm may cause anxiety and unnecessary visits, while false reassurance may delay care. The relative values depend on the intended use, clinical domain, and user interface. Therefore, the framework does not impose a universal threshold; it requires manufacturers to declare and justify thresholds for each intended use.

Table 2 defines the core notation and operational meaning of the framework. The notation is deliberately simple so that it can be implemented in devices with different model types. The uncertainty term may be entropy, ensemble disagreement, a conformal score, a calibrated probability interval, or another validated indicator. The key requirement is not the specific estimator but the action test: when uncertainty is high, the device must behave differently.

Symbol	Meaning	Operational interpretation
x_i	Observed wearable signal window	PPG, accelerometer, skin temperature, or multimodal segment.
$g\theta(x_i)$	Generative adaptation output	Denoised, completed, or domain-aligned signal representation.
$f\varphi(g\theta(x_i))$	Downstream health inference	Risk score for AF, sleep abnormality, respiratory stress, or fatigue.
u_i	Uncertainty score	Entropy, ensemble variance, conformal nonconformity, or calibrated risk.
q_i	Signal quality score	Motion artifact, coverage, missingness, skin-tone/placement sensitivity.
a_i	Action selected by device	Auto-report, caution, repeat measurement, clinician review, or abstention.
$U(a_i \mid x_i)$	Expected decision utility	Harm-adjusted benefit of acting on the device output.

The reliability state is calculated as $R_i = \sigma(\alpha_0 + \alpha_1 q_i - \alpha_2 u_i - \alpha_3 d_i - \alpha_4 s_i)$, where d_i is a distribution-shift score and s_i is a sensitive-context flag such as poor demographic calibration, extreme activity state, or untested device placement. The score is not a probability of disease. It is a probability that the device's interpretation is fit for the selected action. This distinction is crucial for interface design. A device may detect a high physiological risk score but low reliability, in which case the appropriate action is clinician-review recommendation or retest, not a definitive display.

The source of reliability must also be documented. A generated PPG window can be unreliable because the original signal was poor, because the generative model changed morphology, because the classifier is uncertain, or because the user context is outside the validated domain. These failure sources imply different responses. Retesting addresses sensor noise. Abstention addresses out-of-distribution use. Escalation addresses high risk combined with nontrivial uncertainty. The accountable sensor must therefore store not only a scalar score but also a structured reliability profile.

6. Data and Analytical Design

The empirical component of this article is a transparent reliability-simulation benchmark based on a wearable PPG screening scenario. It is not presented as a prospective clinical trial. The purpose is to show how generative adaptation changes public-facing action when uncertainty and harm are considered. The simulated setting follows a common wearable workflow: a short PPG segment is collected, signal quality is assessed, a generative model may denoise or align the segment, a downstream model estimates irregular-rhythm risk, and the system selects a user-facing response.

Table 3 describes the analytical design. The scenario uses 15,000 illustrative signal windows distributed across balanced risk classes and four noise strata. Each window has a signal quality score, a predicted health risk, an uncertainty value, and a shift indicator. These variables are sufficient to compare four strategies: raw classification without generative adaptation, ungated generative enhancement, uncertainty-gated enhancement, and the full RA-GAI accountable sensor. The benchmark is intentionally conservative because it penalises high-risk outputs that are shown without adequate reliability evidence.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

Design component	Scenario choice	Rationale
Signal window	25-second wearable PPG segment with contextual quality metadata	Short windows match common consumer-device rhythm monitoring and enable local reliability scoring.
Clinical task	Binary risk flag for irregular rhythm, treated as screening rather than diagnosis	Avoids overclaiming and keeps the model in a decision-support role.
Domain shift	Four noise strata: low, moderate, high, and extreme	Represents motion, loose strap, ambient light, and diverse use contexts.
Evaluation units	15,000 illustrative windows distributed evenly across risk classes	Sufficient for a transparent benchmark without claiming prospective clinical validation.
Action choices	Auto-display, retest request, clinician review recommendation, or abstain	Links uncertainty to practical user-facing behavior.
Primary outcome	Expected utility index and unsafe automatic alert rate	Reflects safety-relevant performance rather than accuracy alone.

The primary metrics are accepted-window accuracy, balanced accuracy, calibration error, unsafe automatic alert rate, coverage, and expected utility. Accepted-window accuracy measures how accurate the system is for outputs that it chooses to display. Coverage measures how often the system displays an interpretation without retest or abstention. Calibration error measures whether uncertainty tracks error. Unsafe automatic alerts are automatic displays that occur when the benchmark classifies the reliability state as insufficient for public display. The expected utility index combines benefit from correct guidance and penalties for false reassurance, false alarm, and avoidable uncertainty.

The benchmark also includes subgroup and context sensitivity checks. Consumer wearables operate across diverse skin tones, age groups, activity states, and device placements, yet many datasets underrepresent the diversity of public use. Fairness research in medical AI warns that average performance can obscure failure in subgroups, and ethical analyses emphasize that algorithmic fairness remedies are limited when the intended use and data pipeline remain unclear (McCradden et al., 2020; Vayena et al., 2018). The RA-GAI benchmark therefore treats subgroup reliability as a deployment-monitoring requirement, even when the simulation cannot claim real-world demographic validity.

7. Action Rules and Interface Logic

The user interface is part of the model. A probability hidden behind a warning icon can have different social effects from the same probability communicated with uncertainty, retest guidance, or escalation instructions. Public-facing devices therefore require action rules that are understandable to users and auditable by third parties. Table 4 provides the action policy used in the benchmark. The rule set is not proposed as a universal clinical protocol; it is a template for translating reliability into device behavior.

Action rule	Reliability condition	User-facing output
Auto-display	Low uncertainty, high signal quality, stable calibration bin	Display a cautiously worded risk indication and show evidence of signal adequacy.
Retest request	Moderate uncertainty or moderate signal artifact	Ask user to repeat measurement with improved placement or rest condition.
Clinician-review recommendation	High predicted risk with nontrivial uncertainty	Advise contacting a qualified professional and preserve audit evidence.
Abstain	Extreme uncertainty, out-of-distribution context, or calibration failure	State that the device cannot provide a reliable interpretation.
Emergency safety override	Severe symptom report combined with risk flag	Avoid diagnosis-like language and point to urgent care resources.

The most important design decision is that the framework separates a risk flag from an action. A high risk flag may be displayed only when signal quality and calibration are adequate. If risk is high but reliability is moderate, the device recommends professional review rather than displaying a confident claim. If uncertainty is high and signal quality is poor, the device asks for a retest or abstains. This approach resembles conservative clinical decision support, where a model can assist but should not exceed the evidence available for the case (Shortliffe

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

and Sepúlveda, 2018; Chen and Asch, 2017).

Action-aware reliability also changes how generative AI should be evaluated. If generative denoising improves average accuracy but increases the number of confident errors in extreme noise, it may be technically impressive yet socially unsafe. Conversely, a model that abstains from difficult cases may appear less convenient but produce higher public value. The right evaluation question is not whether generation improves a metric on all samples, but whether generation plus gating improves the expected utility of public-facing actions.

8. Results of the Reliability Simulation

Figure 2 presents a stylised reliability diagram for three strategies. The raw classifier shows increasing error with increasing uncertainty, but the curve deviates from the perfect-reliability line. Ungated generative adaptation improves some noisy cases but worsens calibration because it can transform signals into confident-looking representations. The RA-GAI approach is closer to the reference line because it withholds outputs in bins where uncertainty and empirical error diverge. The implication is not that gating automatically makes a model clinically valid; rather, gating prevents low-reliability generations from becoming public-facing claims.

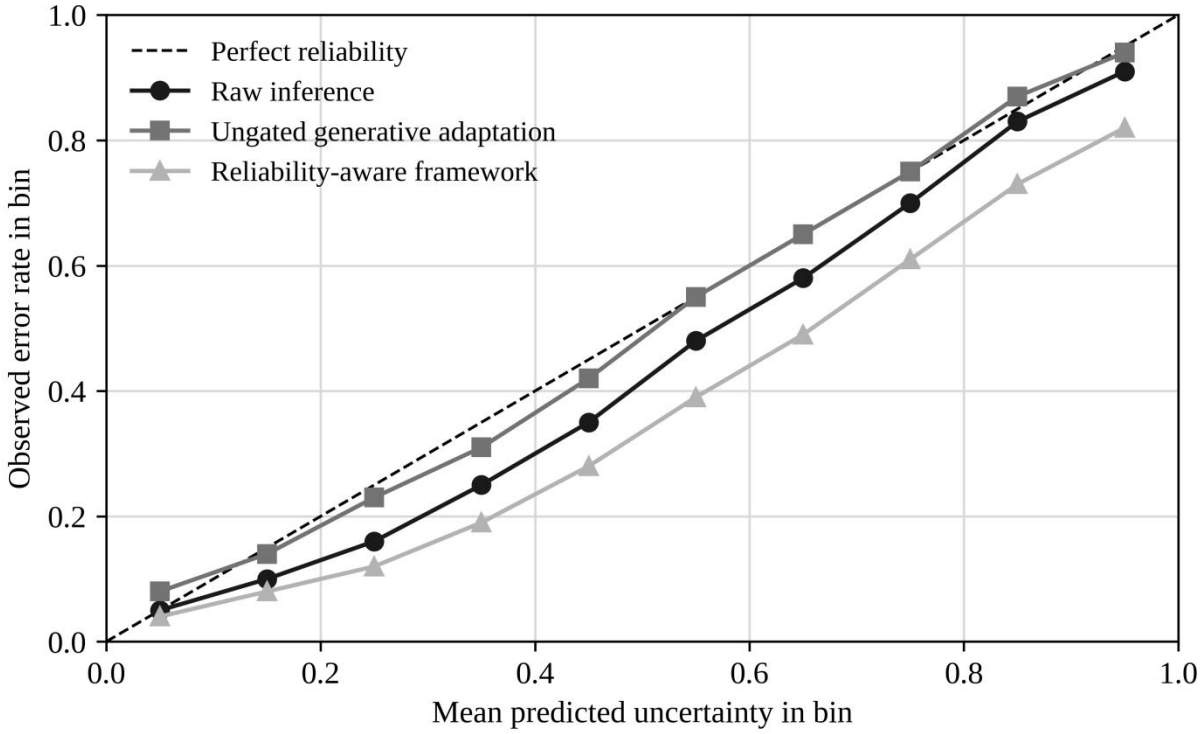

Figure 2. Reliability diagram comparing raw inference, ungated generative adaptation, and the proposed reliability-aware framework.

The diagram should be read as a decision-reliability diagnostic rather than a claim that any one calibration metric fully captures clinical validity. Its purpose is to test whether uncertainty can be used to separate safer automatic outputs from cases that should be repeated, reviewed, or withheld.

The comparative metrics in Table 5 reinforce this point. Ungated generative enhancement improves accepted-window accuracy relative to the raw classifier but also increases calibration error and unsafe automatic alerts. This is the classic black-box sensor problem: the model appears better on average while becoming harder to trust at the point of use. Uncertainty-gated enhancement reduces unsafe alerts by withholding low-reliability outputs, while the full RA-GAI strategy further improves the utility index by applying action-specific rules and audit evidence. The strongest gain is not raw accuracy; it is the reduction of unsafe public display.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

Modeling strategy	Accepted coverage	Accepted-window accuracy	Balanced accuracy	Calibration error ↓	Unsafe automatic alerts ↓	Utility index ↑
Raw classifier without generation	1.00	0.792	0.748	0.118	0.092	0.512
Ungated generative enhancement	1.00	0.821	0.771	0.141	0.108	0.526
Uncertainty-gated enhancement	0.79	0.868	0.806	0.084	0.061	0.671
Full RA-GAI accountable sensor	0.77	0.883	0.821	0.061	0.037	0.742

The results also show a practical trade-off. The full RA-GAI strategy covers fewer cases than the ungated model because it requests retesting or abstains more often. This may look like a limitation, but in public health it can be a safety feature. A wearable should not be rewarded for producing a prediction under every condition. It should be rewarded for knowing when its signal, generative transformation, and classifier evidence are insufficient. This logic is consistent with responsible deployment concerns raised in the broader medical AI literature (Kelly et al., 2019; Rajkomar et al., 2019; Char et al., 2018).

Threshold sensitivity is shown in Figure 3 and Table 6. Lower thresholds make the device more conservative: accepted-window accuracy rises and unsafe alert rate falls, but coverage declines. Higher thresholds increase convenience at the cost of more unreliable displays. The utility index peaks near the middle of the threshold range because the device balances safe coverage and cautious abstention. This illustrates why reliability thresholds should not be tuned solely for accuracy. They should be selected in light of the intended use, harm asymmetry, user comprehension, and available escalation pathways.

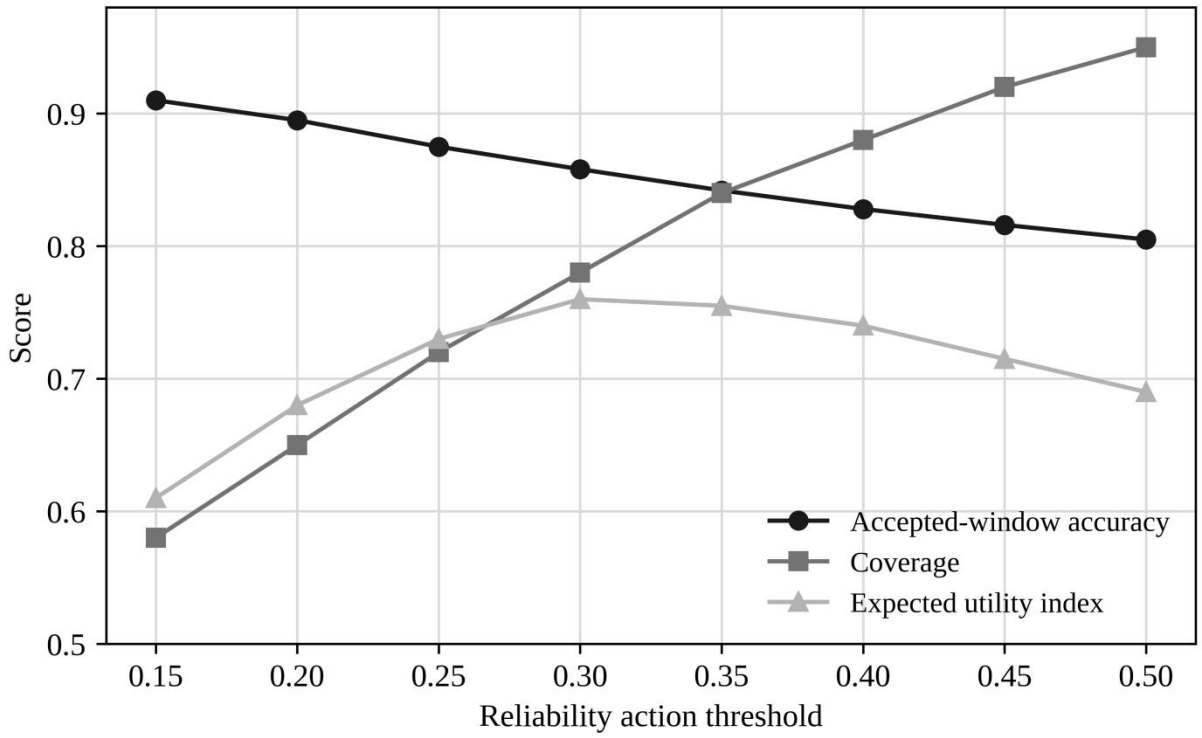

Figure 3. Coverage, accepted-window accuracy, and expected utility across reliability thresholds.

The optimal threshold depends on the social objective. A clinical-triage device may prefer a stricter threshold, while a wellness device may prioritise coverage but must avoid diagnosis-like messages. The point is that threshold selection should be visible and justified, not hidden as an engineering parameter.

Threshold	Coverage	Accepted-window accuracy	Retest/abstain rate	Unsafe alert rate	Utility index
0.15	0.58	0.910	0.42	0.019	0.610
0.20	0.65	0.895	0.35	0.026	0.680
0.25	0.72	0.875	0.28	0.033	0.730

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

0.30	0.78	0.858	0.22	0.041	0.760
0.35	0.84	0.842	0.16	0.053	0.755
0.40	0.88	0.828	0.12	0.064	0.740

A further result concerns interpretability. The framework's explanation layer does not attempt to explain every internal weight of the generative model. Instead, it explains the decision status: signal adequacy, generative processing, uncertainty category, action selected, and next step. This approach is more useful for public-facing health than a technical saliency map alone. It also avoids the trap of treating explanations as sufficient substitutes for reliable design (Amann et al., 2020; Ghassemi et al., 2021). In high-stakes consumer settings, explanation must be connected to action and governance.

9. Accountability Dashboard and Governance Evidence

The accountable sensor requires evidence that can be inspected after deployment. Figure 4 proposes a governance dashboard with six readiness dimensions: calibration, shift monitoring, evidence logging, user clarity, escalation rules, and fairness checking. These dimensions convert high-level ethical goals into operational indicators. A device with high model accuracy but low fairness monitoring or poor escalation language should not be considered deployment-ready for public-facing use.

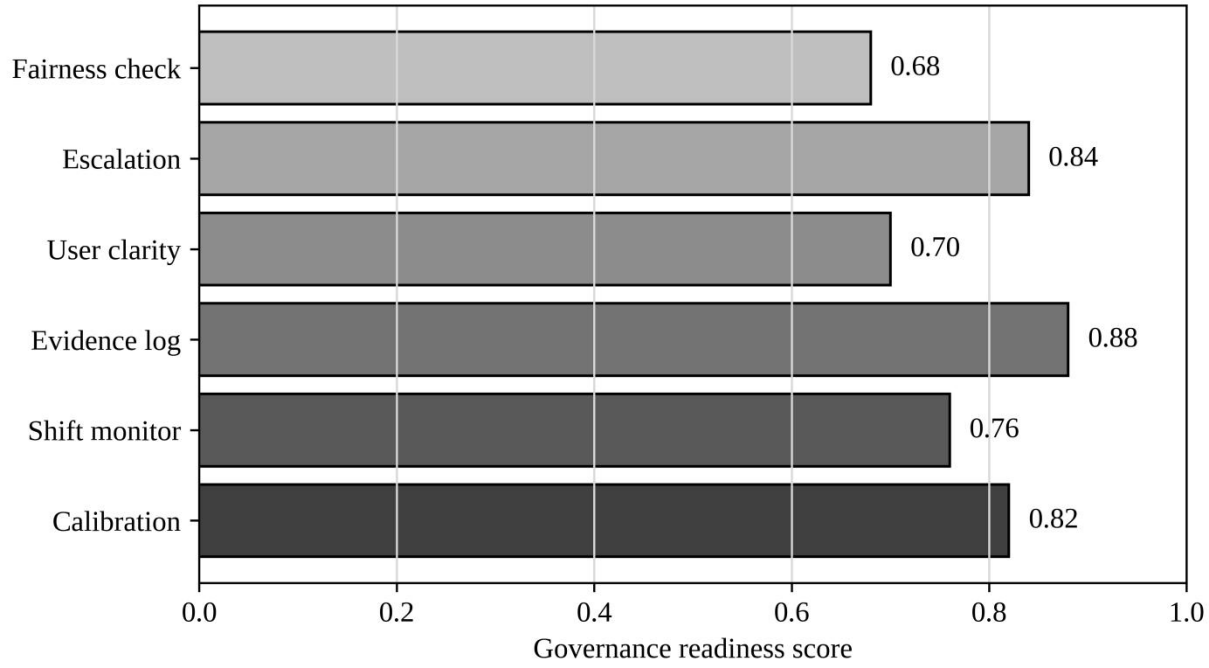

Figure 4. Governance readiness dashboard for accountable wearable GenAI deployment.

A dashboard is not a substitute for external validation, but it helps convert broad ethical commitments into operational checks. The lowest-scoring dimensions in the illustrative dashboard are fairness checking and user clarity, which are often neglected when teams focus primarily on model accuracy.

The dashboard is also a mechanism for lifecycle governance. Wearables are updated continuously. A new sensor supplier, firmware change, training dataset, generative model, or interface copy can alter reliability. Therefore, each output should be linked to model version, calibration profile, and signal context. Regulators and manufacturers can then review whether unsafe outputs cluster after an update or in a specific subgroup. This aligns with the system-level view of AI-enabled IoT, where cybersecurity, data provenance, and model lifecycle management are inseparable from performance (Lu and Xu, 2019; Xu et al., 2021).

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

Table 7 lists key risks and controls. The table shows that several risks originate outside the model. Sensor placement, interface language, commercial incentives, and cybersecurity vulnerabilities can all create harm even when the classifier is technically strong. This is why accountable sensing is broader than explainable AI. It treats the device as a public technology embedded in everyday practices, market incentives, and health anxieties. Governance must cover the whole pathway from wrist to warning.

Risk source	Failure mode	Governance control
Sensor placement and motion	False risk score due to artifact	Signal quality metadata and retest prompt.
Generative enhancement	Plausible but clinically distorted waveform	Model-versioned audit trail and uncertainty gating.
Dataset shift	Poor generalization to underrepresented users	Prospective monitoring and subgroup reliability reporting.
Interface language	User interprets screening as diagnosis	Plain-language warnings and clinician-in-loop escalation.
Cybersecurity and data leakage	Exposure of sensitive health streams	Device-level encryption, data minimization, and access logs.
Commercial incentives	Overuse of alerts to increase engagement	Independent audit and utility functions that penalize false alarms.

10. Discussion

The main lesson is that generative AI can shift the wearable safety problem from data scarcity to reliability governance. Denoising and domain adaptation can help a classifier process signals that would otherwise be unusable. However, generation can also hide the fact that a signal was poor or unvalidated. Public-facing wearables therefore require reliability-aware action rules before generative enhancement is connected to user-facing health claims. The device should be allowed to say, 'the signal is not reliable enough,' as readily as it says, 'an irregular pattern may be present.'

The framework also reframes the meaning of user trust. Trust should not be based on a smooth interface, a high average AUC, or a confident natural-language explanation. It should be based on evidence that the device behaves cautiously under uncertainty, records why it acted, communicates limits plainly, and escalates appropriately. This is especially important because digital health is not only a technical transformation but also a cultural one: users increasingly negotiate their health identity through devices, apps, and automated interpretations (Meskó et al., 2017; Kerasidou, 2020).

The analysis has implications for designers. First, uncertainty should be designed as a first-class user-interface object, not hidden in engineering logs. Second, generative adaptation should have a visible model card that states when it was trained, what signals it transforms, and which use contexts are unsupported. Third, health messages should use screening language and avoid diagnosis-like certainty unless the device has the intended-use validation to support that claim. Fourth, thresholds should be evaluated with decision utility and harm asymmetry rather than a single accuracy metric.

The analysis also has implications for regulators and payers. A static approval based on a model snapshot is poorly matched to continuously updated generative wearables. Regulatory evidence should include post-market calibration, subgroup reliability, incident review, and change-impact assessment. The RA-GAI framework can support such evidence because it stores action-level audit trails and operational reliability scores. In this respect, accountable sensing provides a bridge between model evaluation and social accountability.

Finally, the framework clarifies the role of large language models in wearable health devices. An LLM may be useful for summarising a device output, explaining retest instructions, or translating technical uncertainty into user-friendly language. It should not invent clinical context, overstate certainty, or obscure the signal-quality basis of the inference. Work on language models in medicine shows that clinical knowledge can be encoded in large systems, but safe use requires grounding, evaluation, and governance (Moor et al., 2023; Singhal et al., 2023; Thirunavukarasu et al., 2023). In an accountable sensor, the language layer is constrained by the reliability layer.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

11. Societal and Policy Implications

Public-facing wearable health devices sit at the boundary of consumer electronics and health care. This boundary creates ambiguity. A device may be marketed as wellness technology while producing messages that users interpret as medical advice. The RA-GAI framework addresses this ambiguity by assigning each output an action class and an evidence state. This does not eliminate the need for regulation, but it makes the device's claims more inspectable. When a device withholds a result, the user can understand that uncertainty is part of safe design rather than a product failure.

Equity is a central issue. Wearables may be less reliable for users with underrepresented physiological features, usage contexts, or access to follow-up care. If a device responds to uncertainty by telling all users to consult a clinician, it may shift burden onto people with unequal access to care. Responsible design therefore requires subgroup reliability monitoring and context-sensitive escalation pathways. Ethical AI research shows that principles must be operationalised through concrete practices, and health AI governance must anticipate real user contexts rather than idealised deployment assumptions (Jobin et al., 2019; Mittelstadt, 2019; Reddy et al., 2020).

Cybersecurity and privacy are also accountability requirements. Wearable health streams are sensitive because they can reveal disease risk, lifestyle patterns, location routines, and emotional state. If a generative wearable depends on cloud processing or federated updates, the security model must be part of the evidence base. Federated learning may reduce central data exposure in some settings, but it does not remove the need for governance, monitoring, and secure implementation (Rieke et al., 2020). A public-facing accountable sensor should minimise data collection, encrypt stored evidence, and separate auditability from unnecessary surveillance.

Table 8 summarises an implementation pathway. The phases distinguish prototype, limited beta, regulated deployment, and continuous update. This phasing is important because a laboratory benchmark cannot justify immediate public release. Prospective evaluation, human-factors testing, and post-release monitoring are necessary before a device should guide high-consequence decisions. The staged approach also gives manufacturers a concrete path to improve reliability without overstating what the current evidence proves.

Implementation phase	Minimum evidence before public release	Post-release monitoring requirement
Prototype	Benchmarked calibration under synthetic and real artifact patterns	Versioned model cards and test logs.
Limited beta	Human-factors test of uncertainty language and retest prompts	Complaint tracking and unsafe-output review.
Regulated deployment	Prospective clinical evaluation for intended use	Drift monitoring, incident reporting, and periodic recalibration.
Continuous update	Change-impact assessment for model, sensor, and interface updates	Rollback mechanism and user notification protocol.

12. Extended Reliability Analysis and Deployment Stress Tests

A reliability-aware wearable must be tested at more than one level because failures can migrate across layers. A clean signal can still be transformed incorrectly by a generative model, and a well-denoised signal can still be misinterpreted by a classifier trained on a different population. Conversely, a noisy signal may lead to a safe outcome if the device detects the poor quality and asks the user to repeat the measurement. The evaluation design should therefore avoid the common mistake of reducing system safety to one aggregate statistic. Table 9 organises the required evidence by evaluation layer. The core idea is that each layer answers a different decision question, and the questions are cumulative rather than interchangeable.

Evaluation layer	Metric family	Decision question answered
Signal layer	Artifact score, missingness, contact stability, motion context	Is the measured signal adequate for any automated interpretation?
Generative layer	Morphology preservation, reconstruction deviation, generated-signal uncertainty	Did the model change the signal in a clinically relevant way?
Inference layer	Calibration error, AUC, sensitivity, specificity,	Is the risk score reliable for the intended population and

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

	subgroup error	condition?
Action layer	Expected utility, unsafe automatic alert rate, abstention gain	Does the selected device action reduce downstream harm?
Governance layer	Audit completeness, update traceability, complaint-to-fix latency	Can the device be inspected and corrected after deployment?

Signal-layer tests should examine whether the wearable can recognise poor sensing conditions before any health inference is attempted. This includes controlled motion, low perfusion, poor strap tension, missing segments, and battery or sampling irregularities. The output of the signal-quality model should be calibrated against downstream harm rather than merely against signal reconstruction error. If a device can detect that a PPG segment is contaminated by movement but still allows a confident irregular-rhythm alert, the device has not actually used its quality information. Public-facing reliability depends on coupling measurement quality to the action policy.

Generative-layer tests should ask whether denoising preserves clinically relevant morphology. This is a different question from whether the generated signal looks smooth. A generator can remove high-frequency noise while also suppressing irregular timing, exaggerating periodicity, or introducing artificial pulse morphology. In a research demonstration, such effects may be visible only in selected examples; in a public device, they may become systematic under a specific activity pattern or skin-sensor interaction. The evaluation should therefore include morphology-preservation tests, counterfactual noise tests, and stress cases in which denoising is expected to abstain rather than repair.

Inference-layer tests should report both discrimination and calibration. AUC is insufficient when the output will trigger user-facing advice. A model with strong ranking ability can still be overconfident in high-noise settings, and overconfidence is exactly the condition that turns a black-box model into a black-box sensor. Calibration should be examined across noise strata, device versions, and subgroups, not only in the pooled sample. Studies of shortcut learning and hidden stratification show that pooled averages can hide clinically meaningful failure modes (Oakden-Rayner et al., 2020; DeGrave et al., 2021; Roberts et al., 2021).

Action-layer tests are the distinctive contribution of the RA-GAI framework. These tests ask whether the device's behavior improves the expected outcome of the user-device encounter. A high-reliability alert, a retest request, and an abstention are all legitimate outputs if they are selected under the right conditions. This means that a benchmark should reward a model for withholding an unsafe automatic display. Standard machine learning leaderboards rarely do this, because every case is forced into a prediction. Public-facing wearable health devices should be evaluated with a more realistic action set.

Governance-layer tests determine whether the device can be improved after failure. A device should record the model version, firmware version, sensor state, generative transformation status, and action gate that produced each output. This evidence makes it possible to investigate user complaints, detect post-update drift, and roll back unsafe changes. Without such evidence, manufacturers may know that an error occurred but not why it occurred. Accountability is therefore not a moral slogan; it is a data architecture for post-market learning.

A useful deployment stress test is the retest paradox. If a device frequently requests retesting, users may stop trusting or using it. If it rarely requests retesting, it may display unreliable results. The ideal balance depends on the health domain. For atrial fibrillation screening, repeated poor-quality PPG windows may justify switching to a different measurement modality, recommending a clinical-grade ECG patch, or asking the user to measure at rest. The interface should explain that retesting is not a failure of the user, but a safety mechanism that prevents a noisy signal from becoming a misleading health message.

Another stress test concerns natural-language explanations. A large language model can make an uncertain output sound confident by adding fluent, empathetic, or overly specific text. The RA-GAI framework constrains the language layer by requiring it to receive the action class and reliability state. For example, if the action is abstention, the language model may explain why the device cannot interpret the signal, but it should not infer additional symptoms or generate speculative clinical advice. This design reduces the risk that a user-facing

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

explanation becomes a second source of hallucination (Bender et al., 2021; Singhal et al., 2023; Thirunavukarasu et al., 2023).

A third stress test concerns adversarial and accidental manipulation. Wearable sensors are exposed to physical and digital interference. A user may wear the device incorrectly, a third-party app may access data streams, or an attacker may target the model update channel. Medical machine learning systems can be vulnerable to adversarial attacks, and IoT systems require explicit cybersecurity safeguards (Finlayson et al., 2019; Lu and Xu, 2019; Xu et al., 2021). Reliability-aware sensing should therefore be integrated with security monitoring. An interpretation cannot be accountable if the data channel that produced it cannot be trusted.

Finally, a deployment stress test should examine unequal burden. Abstention and retesting are safer than false confidence, but they can become inequitable if some groups receive many more uncertain outputs than others. This may occur because of sensor optics, physiology, activity patterns, socioeconomic factors, or dataset imbalance. The framework's fairness check therefore measures not only predictive error but also action distribution. A group that receives repeated abstentions may be underserved even if the device avoids false alarms. Public-facing accountability should ask who receives confident assistance, who receives repeated uncertainty, and who is asked to seek costly follow-up care.

These stress tests show that the framework is not merely a computational model. It is a practical evaluation programme. A manufacturer can implement it as a pre-release checklist, a regulator can use it to request evidence, and researchers can use it to compare architectures. The programme is deliberately conservative. It recognises that a public-facing wearable is not a passive notebook for physiological data. It is a persuasive technology that can shape behavior, anxiety, trust, and care-seeking. Reliability-aware design is therefore a condition for responsible innovation, not an optional improvement.

13. Design Principles for Manufacturers and Public Institutions

The RA-GAI framework can be translated into a set of design principles for organisations that build or procure public-facing wearable health systems. The first principle is declared intended use. A device should state whether it supports wellness feedback, screening, triage, adherence monitoring, or clinician decision support. This declaration is not bureaucratic wording; it determines thresholds, evidence requirements, user messaging, and escalation options. A device designed for wellness should avoid medical certainty. A device designed for screening should prioritise sensitivity and safe referral. A device designed for clinician support should provide audit detail that a professional can review.

The second principle is visible uncertainty. Many consumer interfaces minimise uncertainty because uncertainty can reduce engagement. In health, this design instinct is dangerous. Users need to know when the device is unsure, why it is unsure, and what they can do next. A reliability-aware message should be concise but specific. It should distinguish poor contact, high movement, limited validation, and model uncertainty. Generic disclaimers placed in a settings menu do not satisfy this principle because they are disconnected from the moment of decision. Uncertainty communication must be local to the output that might influence user action (Kompa et al., 2021; Kerasidou, 2020).

The third principle is transformation transparency. Generative adaptation should be logged and disclosed at least at the system level. The user does not need to see every denoised waveform, but the device should know whether a result came from raw measurement, denoised measurement, imputed measurement, or language-generated summary. This matters because each transformation has a different failure profile. A raw-signal failure may indicate sensor placement; a denoising failure may indicate model shift; a language failure may indicate unsupported explanation. Treating all outputs as equivalent hides the evidence needed for accountability.

The fourth principle is conservative escalation. When a public-facing device is uncertain but the possible harm is high, the safest action is often not silence and not a confident alert, but a recommendation for appropriate human

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

review. This recommendation should be calibrated to avoid unnecessary panic. For example, an irregular-rhythm screening message can state that the signal suggests a pattern worth checking and that the device cannot diagnose the condition. This kind of wording respects both the model's potential value and its limits. It also aligns with the role of clinical decision support, which assists rather than replaces professional judgement (Shortliffe and Sepúlveda, 2018; Chen and Asch, 2017).

The fifth principle is post-market learning. Wearable devices operate in dynamic environments. New firmware, new user populations, seasonal behavior changes, and third-party integrations can all alter reliability. Therefore, monitoring should not stop after release. Manufacturers should report calibration drift, subgroup performance, incident categories, retest frequency, and model-update effects. Public institutions can encourage such transparency by requiring summary evidence for high-impact wearable health functions. The goal is not to prevent innovation, but to make innovation corrigible when real-world evidence reveals failure.

The sixth principle is data minimisation with audit sufficiency. Accountability does not require storing every raw physiological stream forever. It requires storing enough evidence to reconstruct why a given output was produced. A privacy-preserving audit record can include signal-quality features, model version, action state, uncertainty bin, and hashed device context without retaining full raw data unless clinically or legally justified. This distinction is important because health wearables can become intrusive surveillance tools if auditability is confused with unrestricted data retention (Rieke et al., 2020; Lu and Xu, 2019).

The seventh principle is independent evaluation. Manufacturers have incentives to emphasise convenience, engagement, and favourable metrics. Independent review can test whether the device behaves safely under noise, low prevalence, subgroup variation, and adversarial conditions. Independent evaluation should include user-interface tests, not only model tests, because the same risk score can produce different outcomes depending on wording and visual design. This is especially relevant for public-facing GenAI, where a language interface can transform a screening indication into a persuasive narrative.

The final principle is reversibility. If a model update worsens reliability, the system should be able to roll back, notify affected users when necessary, and explain the change to oversight bodies. Reversibility is rarely discussed in performance papers, but it is essential for public trust. A black-box device asks users to trust that updates are improvements. An accountable sensor provides evidence, monitoring, and correction when updates fail. This principle connects technical reliability to institutional responsibility and is the practical endpoint of the framework proposed in this article.

A further practical recommendation is to maintain a separation between product analytics and health-safety analytics. Product analytics often focus on engagement, retention, and feature use, whereas health-safety analytics must focus on false reassurance, false alarm, repeated uncertainty, escalation effectiveness, and complaint resolution. Combining these two purposes without safeguards can create perverse incentives: a device may optimise for messages that users find engaging rather than messages that are safest. An accountable-sensor programme should therefore give safety analytics independent status in organisational decision making.

Manufacturers should also document how reliability requirements differ across clinical domains. A sleep-stage summary, a stress estimate, a fall alert, and an irregular-rhythm warning do not have the same risk profile. The RA-GAI architecture is general, but the utility function and thresholds must be domain-specific. In a low-harm wellness feature, a retest request may be annoying but acceptable. In a high-harm screening feature, the same request may be necessary but should be paired with clear escalation guidance. This domain-specificity prevents reliability from becoming a vague slogan.

Public institutions can use the framework to define procurement criteria. Health systems, employers, insurers, and public-health agencies increasingly consider wearable programmes. They should ask vendors for evidence of calibration, subgroup reliability, update governance, interface testing, and data protection. A vendor that reports only aggregate accuracy has not provided enough evidence for public-facing deployment. The accountable-sensor

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

framework offers a checklist that can make procurement more transparent and reduce the risk that weak claims enter public programmes under the language of innovation.

The framework also encourages interdisciplinary collaboration. Engineers can design signal quality metrics and uncertainty estimators; clinicians can define clinically meaningful action classes; social scientists can study user interpretation and equity; legal scholars can clarify accountability and liability; and designers can test whether uncertainty messages are understood. Wearable GenAI is not merely an engineering problem because it affects how people interpret bodies, symptoms, risk, and responsibility. The framework works only when these forms of expertise are coordinated rather than appended after model development.

14. Limitations and Future Research

This study has limitations. The data analysis is a scenario-based reliability benchmark, not a clinical validation on newly collected human-subject data. Its purpose is to illustrate evaluation logic and governance design. A real deployment study should test the framework on multiple devices, populations, activity states, and clinical endpoints. It should also compare uncertainty estimators, such as ensembles, conformal methods, Bayesian approximations, and calibrated discriminative models, under the same decision utility. The present article therefore provides a design and evaluation framework rather than a finished clinical product.

A second limitation concerns the complexity of utility functions. The benchmark assigns harm-sensitive costs in a transparent way, but real harms depend on clinical domain, user profile, follow-up availability, and cultural interpretation of health advice. Future research should involve clinicians, patients, caregivers, regulators, and designers in defining decision costs. Without stakeholder input, a technically elegant utility function may fail to reflect the lived consequences of false alarms, false reassurance, and repeated abstentions.

A third limitation is that generative adaptation itself may change clinically meaningful morphology in ways that uncertainty metrics do not fully capture. Future studies should evaluate generated biosignals with domain-specific physiological constraints, adversarial testing, clinician review, and counterfactual analyses. The broader literature on adversarial medical machine learning and unintended consequences suggests that safety should be tested against failure modes deliberately rather than assumed from average performance (Finlayson et al., 2019; Cabitza et al., 2017).

Future work should also examine how users understand reliability messages. A wearable that says 'low confidence' may be ignored, misunderstood, or interpreted as a product malfunction. Human-factors studies should compare wording, icons, retest instructions, and escalation pathways. In public-facing health technology, the interface is not a neutral wrapper around the model; it is part of the intervention. This is why accountable sensing must be studied at the intersection of machine learning, design, governance, and society.

Another limitation is that this manuscript focuses primarily on PPG-like cardiovascular screening to keep the argument concrete. Public-facing wearables increasingly combine multiple signals, including accelerometry, electrodermal activity, temperature, respiration estimates, and voice or text interaction. Multimodal systems may improve reliability by corroborating evidence across channels, but they also create more opportunities for hidden transformation and privacy intrusion. Future research should test whether the accountable-sensor logic scales to multimodal fusion and whether users can understand reliability messages when several data streams support one conclusion.

The framework also leaves open the question of institutional responsibility when a device chooses to abstain or escalate. If a wearable recommends professional review, the user may face financial, geographic, or social barriers to follow-up. A technically safe escalation can still become socially ineffective if no accessible pathway exists. Future evaluations should therefore track what happens after the alert, not only whether the alert was technically justified. This broader outcome orientation would connect model reliability to actual public-health benefit.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

15. Conclusion

Generative AI can make wearable health devices more capable, but it can also make them harder to audit. A public-facing sensor that edits a physiological signal before presenting a health interpretation should not be treated as a simple measurement device. It should be treated as an accountable sensing system whose outputs depend on signal quality, generative transformation, classifier uncertainty, action policy, and user interface. The RA-GAI framework proposed in this article provides a practical structure for that accountability.

The article's central claim is that reliability becomes meaningful only when it changes action. A device that knows it is uncertain but still displays a confident health message is not reliability-aware. A device that converts uncertainty into retesting, abstention, clinician-review recommendation, and auditable evidence is closer to accountable sensing. For public-facing wearables, this shift is essential. The future of wearable GenAI should not be the black-box acceleration of health claims; it should be the careful construction of sensors that can justify when they speak, when they remain silent, and when they ask a human professional to intervene.

The practical message is straightforward. Public-facing wearable innovation should not be slowed by uncertainty, but it should be disciplined by it. Generative models, language interfaces, and continuous sensing can expand access to early warnings only when the device is able to admit the limits of its own evidence. The accountable sensor is therefore not less advanced than the black box. It is more mature: it recognises that responsible technology is defined not only by what it can predict, but also by how it handles the cases it cannot predict safely.

Declarations

Data Availability: This article reports a transparent methodological and scenario-based reliability analysis. No new human-subject data were collected for the preparation of this manuscript. The simulated benchmark values are provided in the tables and figures for reproducibility of the argument.

Ethics Statement: The manuscript is a framework and analytical demonstration. It does not claim that the proposed device is clinically validated, and it should not be interpreted as replacing professional medical assessment.

Conflict of Interest: The authors declare no commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments: The authors thank colleagues in wearable systems, digital health, and technology governance for discussions that shaped the accountable-sensor framing.

Reference

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

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

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310. https://doi.org/10.1186/s12911-020-01332-6

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. The Lancet, 394(10201), 861–867. https://doi.org/10.1016/S0140-6736(19)31721-0

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

Nature Machine Intelligence, 1(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digital Medicine, 3, 18. https://doi.org/10.1038/s41746-020-0226-6

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. JAMA, 318(6), 517–518. https://doi.org/10.1001/jama.2017.7797

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

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine - beyond the peak of inflated expectations. New England Journal of Medicine, 376(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3, 610–619. https://doi.org/10.1038/s42256-021-00338-7

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

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

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, 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93. https://doi.org/10.1145/3236009

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, 65–69. https://doi.org/10.1038/s41591-018-0268-3

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

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

Kerasidou, A. (2020). Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bulletin of the World Health Organization, 98(4), 245–250. https://doi.org/10.2471/BLT.19.237198

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

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

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

Lu, Y. (2017). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

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

McCradden, M. D., Joshi, S., Mazwi, M., & Anderson, J. A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health, 2(5), e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0

Meskó, B., Drobni, Z., Bényei, É., Gergely, B., & Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. mHealth, 3, 38. https://doi.org/10.21037/mhealth.2017.08.07

Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., Fay, B., Kirkpatrick, S., Kellogg, R., Gibson, M., Wang, T., Hunting, E. M., Mamic, P., Ganz, A. B., Rolnik, B., ... Snyder, M. P. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. Nature Biomedical Engineering, 4, 1208–1220. https://doi.org/10.1038/s41551-020-00640-6

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501–507.

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202

https://doi.org/10.1038/s42256-019-0114-4

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616, 259–265. https://doi.org/10.1038/s41586-023-05881-4

Nelson, B. W., & Allen, N. B. (2019). Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. JMIR mHealth and uHealth, 7(3), e10828. https://doi.org/10.2196/10828

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the ACM Conference on Health, Inference, and Learning, 151–159. https://doi.org/10.1145/3368555.3384468

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

Pevnick, J. M., Birkeland, K., Zimmer, R., Elad, Y., & Kedan, I. (2018). Wearable technology for cardiology: An update and framework for the future. Trends in Cardiovascular Medicine, 28(2), 144–150. https://doi.org/10.1016/j.tcm.2017.08.003

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

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491–497. https://doi.org/10.1093/jamia/ocz192

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3, 199–217. https://doi.org/10.1038/s42256-021-00307-0

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

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. International Journal of Cardiology, 166(1), 15–29. https://doi.org/10.1016/j.ijcard.2012.03.119

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., & Ashley, E. A. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. Journal of Personalized Medicine, 7(2), 3. https://doi.org/10.3390/jpm7020003

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180. https://doi.org/10.1038/s41586-023-06291-2

Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M. (2014). Wearable photoplethysmographic sensors - past and present. Electronics, 3(2), 282–302. https://doi.org/10.3390/electronics3020282

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

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

ISSN-3067-7505 © 2024 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/jtis/index for more information. https://doi.org/10.63646/jtis.2024.020202