JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

Task-Guided Confidence Scoring for Synthetic Time-Series Outputs in Health-Oriented Machine Learning Systems

Arjun S. Patel[1]; Sneha V. Kumar[2, *]

[1] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

[2] Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Karunya Nagar, Coimbatore 641114, India

* Corresponding author: sneha.kumar@karunya.edu

ARTICLE INFO
Received
July 18, 2023
Revised
September 21, 2023
Accepted
November 12, 2023
Available Online
December 30, 2023
DOI
10.63646/jaiaa.2023.010403

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Generative models are increasingly used to produce synthetic physiological time series in health-oriented machine learning, whether to denoise wearable recordings, adapt signals across acquisition domains, augment scarce training data, or impute missing segments. Yet the same flexibility that makes these models useful also lets them introduce plausible-looking but misleading artefacts, which is a serious liability when the synthetic signal feeds a clinical decision. This review argues that the trustworthiness of a synthetic output cannot be judged in isolation from the task it is meant to support, and it develops a task-guided confidence scoring perspective that grounds the quality of each generated signal in the expected cost of the downstream decision it influences. We organise the argument around four ideas: that conventional distributional and realism metrics answer the wrong question for deployment; that a useful confidence signal must be per-instance, available before ground truth, and aligned with the decision at hand; that such a signal can be derived from the behaviour of the downstream task model and externally grounded by checking whether higher scores track higher realised decision cost; and that the resulting scores enable principled gating of low-confidence outputs. Using wearable photoplethysmography and atrial-fibrillation screening as a running example, we synthesise reporting strategies across modalities, contrast their properties, and map the deployment, governance, and clinical-translation considerations that determine whether confidence scoring delivers value in practice. The perspective offers a transferable diagnostic for deciding when a synthetic time-series output is safe to use.

Keywords: Confidence scoring; uncertainty quantification; synthetic time series; generative models; domain adaptation; wearable health monitoring; decision-theoretic evaluation; trustworthy machine learning

I. INTRODUCTION

Health-oriented machine learning increasingly depends on data that a sensor never directly recorded. Models that screen for arrhythmia from a smartwatch, infer sleep stages from a wrist sensor, or estimate blood-oxygen trends from an optical signal are trained and deployed on physiological time series that have been cleaned, transformed, or reconstructed by other models before they reach the system that makes the clinical inference. The rapid maturation of deep learning for cardiac monitoring has made continuous, out-of-clinic screening realistic at population scale (Topol, 2019; Rajkomar et al., 2019; Perez et al., 2019). These advances build on deep models that approach expert-level arrhythmia detection from continuous ambulatory recordings, including methods applied to raw wrist-worn optical waveforms (Hannun et al., 2019; Aschbacher et al., 2020). At the same time, the raw signals collected by consumer wearables are noisier and more heterogeneous than the curated recordings on which clinical models are usually trained, which creates a persistent gap between the data a model expects and the data it receives in the field (Pereira et al., 2020; Charlton et al., 2023).

Generative models offer an appealing way to close that gap. Rather than retraining a fragile downstream

 https://doi.org/10.63646/jaiaa.2023.010403

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 1, No. 4 (2023), pp. 36–51**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2023.010403**

model for every new noise condition or device, one can transform an incoming signal so that it more closely resembles the distribution the downstream model was trained on, producing a synthetic time series that stands in for a measurement that was never cleanly observed. The same machinery supports several related operations: removing motion and baseline artefacts, translating a recording from one acquisition setting to another, fabricating additional training examples for rare conditions, and filling gaps left by dropout or sensor disconnection (Goodfellow et al., 2020; Isola et al., 2017; Brophy et al., 2023). In each case the output is not a passive visualisation; it becomes an input to a consequential inference about a person's health.

This is precisely where the difficulty arises. Generative models are optimised to produce outputs that look realistic, and they are very good at doing so even when the underlying content is wrong. A denoising model can invent a rhythm that was not present, smooth away a genuine abnormality, or hallucinate morphology that is statistically typical but clinically misleading. Because these artefacts are plausible by construction, they are difficult to catch by inspection, and they can degrade the accuracy of the downstream model in ways that are invisible until an error has already propagated into a decision (Kelly et al., 2019; Challen et al., 2019). The motivating question of this review is therefore not whether synthetic time series are useful, but how a system can know, for any single generated signal, whether that signal can be trusted enough to act on.

A natural response is to attach a confidence score to each synthetic output and to act only on outputs whose score is acceptable. The argument we develop is that the score must be defined with the downstream decision in view. A measure of how realistic a signal looks, or how closely a population of synthetic signals matches a population of real ones, does not answer the operative question, which is whether using this particular output will lead to a worse decision. We refer to confidence that is defined and evaluated in terms of its effect on the downstream task as task-guided confidence scoring, and we treat the need for it as the central organising idea of the review. The general case for grounding uncertainty in the medical decision it informs has been made forcefully in the clinical machine-learning literature (Begoli et al., 2019).

Figure 1 sketches the setting. A source signal that is out-of-domain or degraded passes through a generative model that produces a synthetic time series; that output feeds a downstream task model whose prediction informs a clinical decision. A confidence layer sits alongside this pipeline and scores each synthetic output according to its expected effect on the decision, enabling the system to accept the output, defer it, or route it for human review. The remainder of the review unpacks each component of this picture and the design choices it entails.

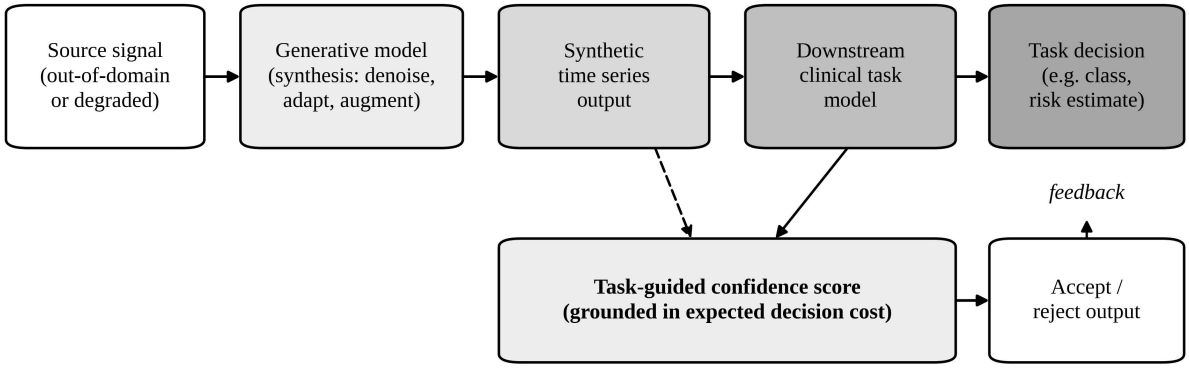

Figure 1. Task-guided confidence scoring situated within a synthetic time-series pipeline for health machine learning.

The contribution of this article is conceptual and integrative rather than experimental. We bring together

three literatures that rarely speak to one another: the generative modelling of physiological time series, the quantification and calibration of predictive uncertainty, and the clinical translation of machine learning. From their intersection we draw a coherent perspective on when a synthetic output is safe to use, organised around four claims that structure the paper. First, conventional quality metrics for generated data answer a question that is misaligned with deployment. Second, a deployment-relevant confidence signal must be per-instance, available before any ground truth is known, and aligned with the specific decision the output supports. Third, such a signal can be obtained from the behaviour of the downstream task model and validated by checking that it tracks realised decision cost. Fourth, the resulting scores support principled gating that measurably improves decision quality.

The review proceeds as follows. Section II characterises the roles synthetic time series play in health systems and the trust gap each role opens. Section III contrasts conventional distributional and realism metrics with the requirements of deployment-time confidence. Section IV develops a four-step framework for constructing and grounding task-guided confidence scores. Section V synthesises scoring strategies reported across modalities and analyses their properties. Section VI examines deployment, governance, and clinical-translation considerations, including the edge and connectivity infrastructure on which continuous monitoring depends. Section VII discusses open problems and a research roadmap, and Section VIII concludes. Throughout, we use wearable photoplethysmography and atrial-fibrillation screening as a concrete running example, because it exhibits every difficulty the framework is meant to address.

II. SYNTHETIC TIME SERIES AND THE TRUST GAP IN HEALTH ML

It is worth being precise about what we mean by a synthetic time series, because the term covers a wider range of artefacts than the phrase first suggests. We use it for any physiological sequence that has been produced or substantially altered by a learned model before being consumed by a downstream task, whether the model removed noise, changed the signal's apparent acquisition conditions, generated an entirely new example, or reconstructed missing values. What unites these cases is that the consumer of the signal treats it as if it were a faithful measurement, while in fact it carries the inductive biases and failure modes of the model that produced it. The growth of generative methods for sequential physiological data has been rapid enough that surveys now struggle to keep pace with the variety of architectures and objectives in use (Brophy et al., 2023; Zhang & Lu, 2021).

The first and most established role is denoising and artefact removal. Wearable optical signals are corrupted by motion, ambient light, poor contact, and physiological confounders, and a learned model can map a corrupted segment to a cleaner estimate of the underlying waveform. Convolutional encoder-decoder architectures with skip connections, originally devised for biomedical image segmentation, have proven effective backbones for this kind of one-dimensional reconstruction, and adversarially trained variants have been applied to both electroencephalography and electrocardiography (Ronneberger et al., 2015; Brophy et al., 2022; Xu et al., 2021). The trust risk here is subtle: a denoiser that is judged only on how clean its output looks may achieve smoothness by erasing exactly the irregular morphology that signals disease.

The second role is domain adaptation. A downstream model trained on signals from one device, population, or acquisition protocol often degrades when applied to another, and adaptation seeks to align the feature distributions of the two domains so that the pretrained model continues to perform. When the target labels are derived from an independent modality, or are simply unavailable, the adaptation must operate on the input representation alone, transforming an incoming signal so that it appears to come from the training distribution (Kouw & Loog, 2021; Wang & Deng, 2018). The trust risk is that the transformation may align superficial statistics while distorting the diagnostic content, producing a signal that is in-distribution for the model but no longer faithful to the patient.

The third role is augmentation. Clinical datasets are typically imbalanced, with pathological examples far scarcer than normal ones, and synthetic examples can be generated to enrich the minority class or to expand

coverage of rare morphologies. Conditional recurrent and adversarial generators have been used to fabricate realistic physiological sequences for exactly this purpose (Esteban et al., 2017; Kiyasseh et al., 2020). The trust risk shifts in this case from a single decision to the model that is trained on the augmented data: if the synthetic minority examples are subtly unrepresentative, they can bias the downstream model in ways that only manifest later, on real patients.

The fourth role is imputation and privacy-preserving substitution. Missing segments from sensor dropout can be reconstructed, and in some settings entire synthetic datasets are released in place of sensitive records so that models can be developed without exposing patient data (Jordon et al., 2022). Here the trust risk compounds, because errors in the synthetic data are inherited by every model trained on it and are easily mistaken for properties of the real population. Across all four roles, the common thread is that the synthetic output is consumed downstream as though it were ground truth, while its reliability is neither observed nor guaranteed.

Table I organises these roles, the operation each performs, a representative health example, and the characteristic way in which trust can fail. The table makes explicit a point that recurs throughout the review: the failure modes are not generic but role-specific, and a confidence signal that is appropriate for one role may be uninformative for another. A denoiser's most dangerous error is the confident erasure of pathology; an augmenter's is systematic bias in the fabricated class; an adapter's is distributional alignment that sacrifices fidelity. This diversity is one reason that single, generic quality scores tend to disappoint, and it motivates the task-grounded view developed in later sections.

Table I. Roles of synthetic time series in health machine learning and their characteristic trust risks.

Role	Operation	Representative health example	Characteristic trust risk
Denoising	Map a corrupted segment to a cleaner estimate of the waveform	Removing motion and baseline artefacts from wrist photoplethysmography	Confident erasure of genuine pathological morphology while output looks clean
Domain adaptation	Align an input with the distribution a pretrained model expects	Translating a signal from a new device to a clinical model's training domain	Superficial statistical alignment that distorts diagnostic content
Augmentation	Generate new examples to enrich scarce or imbalanced classes	Fabricating additional rare-arrhythmia training sequences	Systematic bias in the synthetic class that misleads the trained model
Imputation / substitution	Reconstruct missing segments or stand in for sensitive records	Filling sensor-dropout gaps; releasing synthetic cohorts for model development	Errors inherited by every downstream model and mistaken for real structure

The clinical translation literature has repeatedly emphasised that machine-learning systems fail in deployment for reasons that controlled evaluation does not surface, including distribution shift, feedback effects, and the brittleness of models to inputs that differ subtly from training data (Kelly et al., 2019; Challen et al., 2019). Synthetic time series sharpen each of these concerns, because they interpose an additional learned transformation between the patient and the decision, and that transformation has its own distribution of errors. Reviews of deep learning for electrocardiographic and photoplethysmographic data note that reported performance is often optimistic relative to real-world behaviour, precisely because the conditions of evaluation are cleaner than the conditions of use (Hong et al., 2020; Pereira et al., 2020). A confidence mechanism that is honest about per-instance reliability is one way to narrow the distance between reported and realised performance.

Two framing assumptions bound the scope of the review. First, we are concerned with settings in which the synthetic output feeds a defined downstream decision, because it is the decision that gives the confidence score its meaning; open-ended generation without a downstream task is outside our scope. Second, we focus on established health systems operating under clinical and regulatory expectations, where the cost of acting on a bad output is high and the value of deferring or flagging it is correspondingly large. Within these bounds, the

wearable-photoplethysmography example is representative: the signals are degraded, the downstream task is a consequential binary screen, ground truth is expensive, and the broader survey literature on artificial intelligence has long anticipated this move toward continuous, model-mediated monitoring (Lu, 2019).

III. FROM GENERIC QUALITY METRICS TO TASK-GUIDED CONFIDENCE

The most widely used measures of generated-data quality were developed for a different purpose than deployment-time decision support, and understanding the mismatch is the key to seeing why a task-guided alternative is needed. Broadly, the established measures fall into three families. The first is global distributional comparison, which asks how closely a population of synthetic outputs resembles a population of real ones in some feature space. The Frechet-style distances popularised for image synthesis exemplify this family, and analogous constructions have been proposed for sequential data (Heusel et al., 2017; Shmelkov et al., 2018). These measures are valuable for comparing models during development, but they are corpus-level by construction and say nothing about any individual output.

The second family consists of per-instance realism or plausibility scores, which rate a single synthetic signal by its proximity to a learned manifold of real signals or by the judgement of a discriminator. These do provide a signal for each output, and recent work has formalised sample-level fidelity and authenticity metrics for generative models (Alaa et al., 2022). The third family consists of reference-based error metrics, which compare a generated signal against a known clean version and report a reconstruction error. The difficulty in the generative setting is that the clean reference is usually unavailable at deployment, which is the very reason the generator was needed; and even when a reference exists, an error metric measures faithfulness to that reference rather than the consequence of using the output downstream (Stenger et al., 2024).

A fourth and increasingly important line of work treats confidence as the predictive uncertainty of a model, and asks whether that uncertainty is calibrated against observed error. The uncertainty-quantification literature distinguishes reducible epistemic uncertainty from irreducible aleatoric uncertainty, develops Bayesian and ensemble estimators, and emphasises that raw neural-network confidences are frequently miscalibrated and require correction (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Guo et al., 2017; Abdar et al., 2021; Gawlikowski et al., 2023). Related techniques flag inputs that lie outside the training distribution and construct distribution-free prediction sets with coverage guarantees (Hendrycks & Gimpel, 2017; Angelopoulos & Bates, 2023). This family is much closer to what deployment needs, because it is per-instance and can be available before ground truth, but on its own it still evaluates the estimate of a quantity rather than the quality of the decision the estimate informs.

Stating the deployment requirements plainly clarifies the gap. A confidence signal that is useful for gating a synthetic output must satisfy three properties at once. It must be per-instance, because the system acts on one output at a time and a corpus-level verdict cannot tell it which outputs to trust. It must be available ex ante, before the true outcome is known, because the purpose of the score is to decide whether to act, not to explain an error after the fact. And it must be task-aligned, expressed in terms of the decision the output supports, because the same generated signal may be perfectly adequate for one decision and dangerously inadequate for another. None of the first three metric families satisfies all three properties simultaneously; predictive uncertainty satisfies the first two but only partially the third.

Figure 2 contrasts the two ends of this spectrum directly. On the left, a global distributional evaluation consumes a population of real and synthetic signals, computes an aggregate distance, and returns a single corpus-level verdict; it operates after the fact, depends on access to a reference distribution, and is agnostic to any particular use. On the right, a per-instance task-guided evaluation consumes a single synthetic signal, derives a confidence score from how the downstream task model responds to it, and returns an accept-or-reject decision for that one output; it operates before ground truth, requires no clean reference, and is aligned with the task. The two are complementary, but only the latter supports the moment-to-moment gating that deployment requires.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

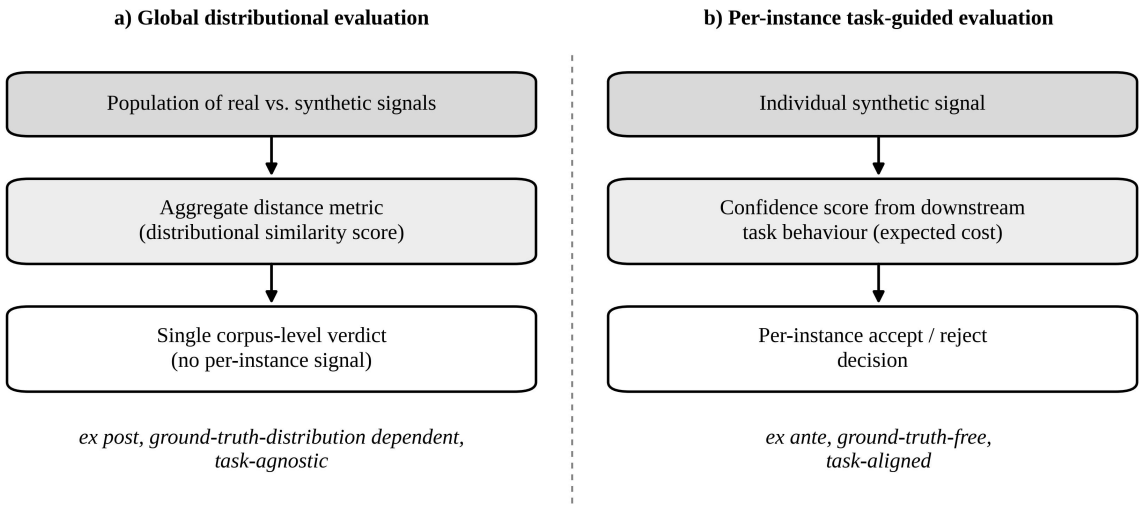

Figure 2. Two evaluation paradigms for synthetic outputs: global distributional assessment versus per-instance task-guided confidence.

The reframing has a practical payoff beyond conceptual tidiness. Because a task-guided score is defined through the downstream model, it lets a system reuse the rich toolkit of uncertainty quantification that was developed for supervised models, applying it to the task model rather than to the generator. This is attractive because generative models are comparatively awkward to equip with calibrated uncertainty, whereas calibration methods for discriminative classifiers are mature (Guo et al., 2017; Abdar et al., 2021). It also means the confidence signal inherits whatever ground truth the downstream task already provides, sidestepping the chronic absence of ground truth for generated outputs. The next section turns this reframing into a concrete construction.

IV. A FRAMEWORK FOR TASK-GUIDED CONFIDENCE SCORING

The framework rests on a simple shift of perspective: instead of asking how good a synthetic signal is in the abstract, ask what decision it will inform and what it would cost to decide wrongly. This is the logic of statistical decision theory, in which the value of an action depends on a loss defined over outcomes, and the rational choice minimises the expected loss under the model's beliefs. Recent work has applied this decision-grounded view to rethink what uncertainty should mean for modern machine-learning systems and to express subjective confidence in terms of the decisions a model supports (Smith et al., 2024; Wang & Holmes, 2024). Applied to synthetic time series, it yields a confidence score whose units are the expected cost of the downstream decision, rather than an abstract quality index.

Figure 3 presents the framework as four design decisions arranged in a loop. The starting point is the intended use case: the system designer must name the decision that the synthetic output will inform, because every later choice depends on it. The four steps then proceed from that decision and feed back to it, so that evidence gathered downstream continually refines the definition of confidence.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

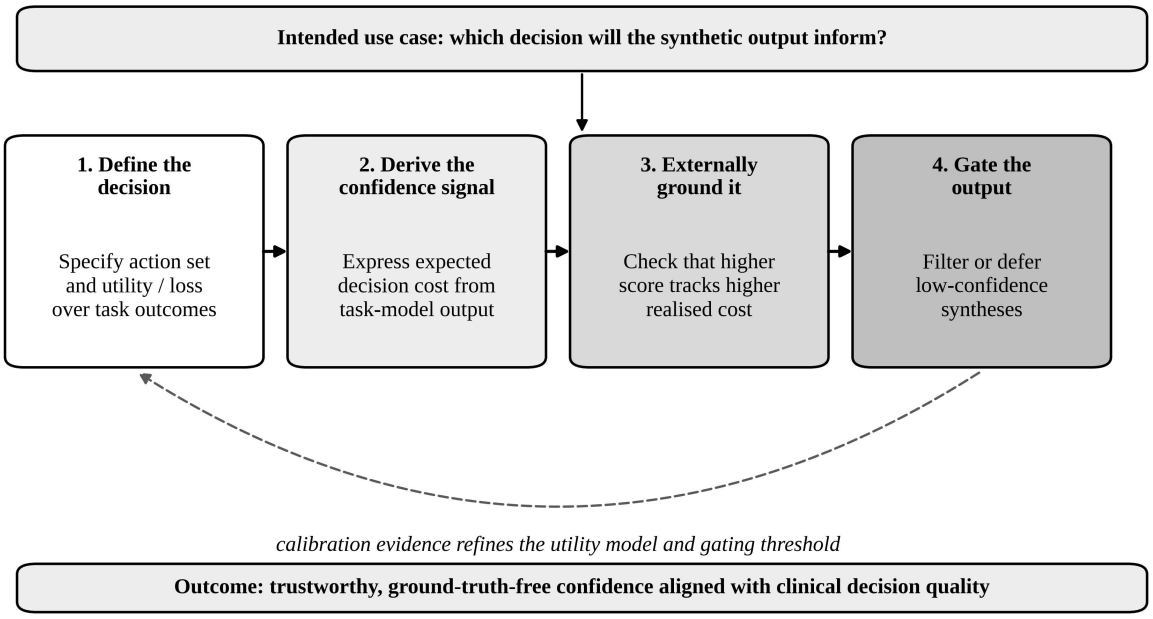

Figure 3. A four-step framework for constructing and grounding task-guided confidence scores.

The first step is to define the decision. This means specifying the set of actions available to the system and a loss or utility over the possible task outcomes. In a binary screening setting such as atrial-fibrillation detection, the actions are the candidate classifications and a natural loss is the misclassification cost, possibly weighted to reflect the asymmetry between missing disease and raising a false alarm. The discipline of writing down the loss is valuable in itself, because it forces an explicit statement of the clinical trade-off that the confidence score is meant to protect, a trade-off that the broader literature on machine-assisted medical decisions argues should never be left implicit (Begoli et al., 2019).

The second step is to derive the confidence signal from the downstream task model. Given the loss, the relevant quantity is the expected decision cost of acting on the synthetic output, computed under the task model's predictive distribution. For the misclassification loss, the expected cost reduces to a simple function of the model's predicted class probabilities, and in the binary case the predictive entropy of those probabilities is a monotone proxy for it. The score is thus read off the behaviour of the downstream model when it is fed the synthetic signal, which is exactly the quantity that matters: not how the signal looks, but how it moves the decision. This makes the predictive entropy or an equivalent expected-cost expression the natural confidence signal, and connects the construction to the established treatment of calibrated uncertainty in classifiers (Guo et al., 2017; Kendall & Gal, 2017).

The third step is to externally ground the signal. A confidence score is only useful if higher scores genuinely correspond to worse decisions, and this correspondence cannot be assumed; it must be demonstrated against realised outcomes. Grounding therefore means checking, on representative held-out data, that outputs the score deems low-confidence do in fact incur higher decision cost than outputs it deems high-confidence. This is the step that turns a plausible heuristic into an evidence-backed instrument, and it is also where calibration and reliability analysis enter, because the shape of the relationship between score and realised cost is itself a calibration question (Lakshminarayanan et al., 2017; Abdar et al., 2021). Critically, the grounding uses the downstream task's ground truth, not any ground truth for the synthetic signal, which is what makes the approach feasible when clean references are unavailable.

The fourth step is to gate the output. Once the score is grounded, the system can act on it: retain outputs

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

whose expected decision cost is acceptable, and defer, flag, or route for human review those whose cost is too high. The gating threshold is not a statistical afterthought but a clinical parameter, because it sets the operating point on the trade-off between acting on uncertain outputs and discarding useful ones. The loop in Figure 3 closes here: the evidence gathered when gated decisions are eventually adjudicated feeds back to refine the utility model and the threshold, so that the confidence instrument improves with use. The use of a downstream model's predictive behaviour to vet generated outputs is a long-standing heuristic; the framework's contribution is to give it a decision-theoretic justification and a grounding procedure (Shmelkov et al., 2018; Smith et al., 2024).

A caveat about the choice of proxy deserves emphasis, because it determines the validity of the whole construction. In binary screening the predictive entropy is a monotone function of the maximum class probability and therefore preserves the ordering of expected decision cost, which is what licenses its use as a confidence signal. In multi-class settings this monotonicity no longer holds in general, and entropy may reorder outputs relative to their true expected cost, so a different expression that respects the loss should be used instead. The lesson is that the confidence signal must be derived from the loss rather than chosen for convenience, and that an order-preserving relationship between the score and the decision cost is the property to verify (Smith et al., 2024; Gawlikowski et al., 2023).

V. ANALYTICAL SYNTHESIS OF SCORING STRATEGIES

Having set out the framework, we now situate it among the strategies actually used to judge synthetic and uncertain outputs, so that its advantages and limitations are visible by comparison rather than by assertion. Table II arranges five representative strategies against the three deployment requirements identified in Section III, together with a fourth column for whether the strategy yields a per-instance signal. The pattern in the table is the analytical core of the review: as one moves down the rows, the strategies acquire the properties that deployment demands, culminating in task-guided confidence, which is the only row that satisfies all four.

Table II. Comparison of scoring strategies for synthetic and uncertain outputs against deployment requirements.

Scoring strategy	Per-instance	Available ex ante	Ground-truth-free	Task-aligned
Global distributional distance (e.g. distributional similarity of corpora)	No	No	No	No
Per-instance realism or plausibility score	Yes	Yes	Partly	No
Reference-based error metric (versus a known clean signal)	Yes	No	No	No
Predictive uncertainty / calibration of the task model	Yes	Yes	Yes	Partly
Task-guided confidence (expected decision cost)	Yes	Yes	Yes	Yes

Read row by row, the table tells a coherent story. Global distributional distances are powerful for model selection but fail every deployment criterion, because they are computed over corpora, after the fact, against a reference distribution, and without reference to any task. Per-instance realism scores recover the per-instance and ex-ante properties but remain task-agnostic and only partially independent of a reference manifold (Alaa et al., 2022). Reference-based error metrics are per-instance but require a clean signal that deployment rarely provides, and they measure faithfulness rather than decision quality (Stenger et al., 2024). Predictive uncertainty and calibration of the task model satisfy the first three requirements and come close on the fourth, which is precisely why the framework builds on them; task-guided confidence completes the final step by tying the score to the decision's expected cost (Guo et al., 2017; Abdar et al., 2021).

The decisive empirical question for any confidence instrument is whether the score is externally grounded,

meaning that higher scores reliably correspond to higher realised decision cost. Figure 4 illustrates the distinction that grounding is meant to detect. The curves are schematic rather than measured, drawn to make the concept legible: a well-grounded score rises monotonically with the realised misclassification rate, so that thresholding it cleanly separates outputs that should be retained from those that should be deferred, whereas a poorly-grounded score is roughly flat or non-monotone and offers no useful operating point. Verifying which case obtains is an empirical exercise on representative data, and it is the same kind of reliability analysis that the uncertainty-quantification literature applies to classifier confidences (Guo et al., 2017; Lakshminarayanan et al., 2017).

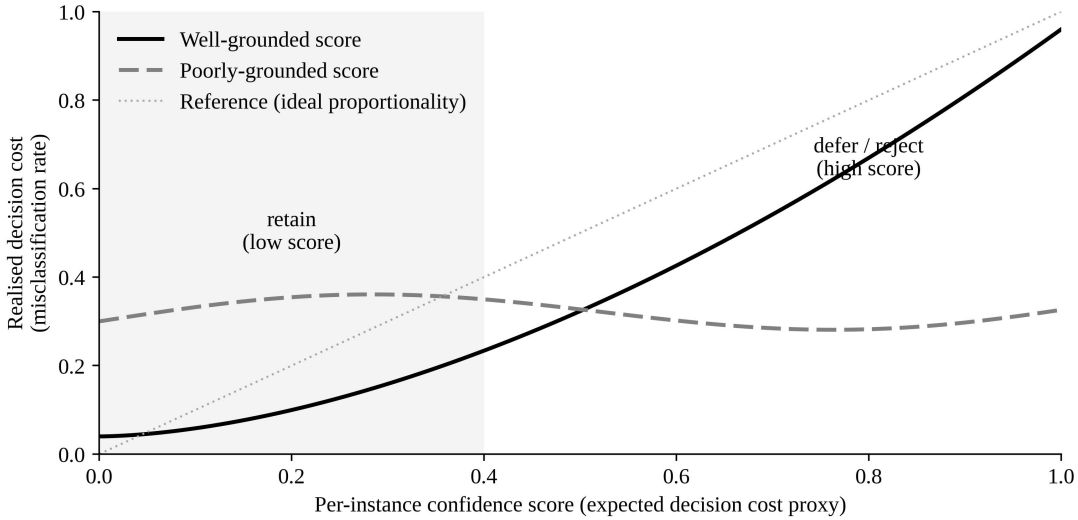

Figure 4. Schematic relationship between a per-instance confidence score and realised decision cost for a well-grounded and a poorly-grounded score.

The grounding analysis also exposes where confidence scoring is hardest, which is exactly the regime in which it is most needed. Out-of-distribution inputs, atypical morphologies, and underrepresented patient subgroups are the cases where a generator is most likely to hallucinate and where a task model's uncertainty is least likely to be calibrated, so the score can be least trustworthy precisely when the stakes are highest (Ovadia et al., 2019; Hendrycks & Gimpel, 2017). This is not an argument against confidence scoring but an argument for evaluating it under shift rather than only on average, and for treating aggregate reliability metrics as a starting point rather than a conclusion. Reliability that holds on a convenient test set may not survive contact with a new device or a new clinic.

A further subtlety is that a confidence score should reflect the changes the generator introduces, not merely the difficulty of the original measurement. If a score assigned to a denoised signal is driven entirely by properties of the noisy input, it tells us about the measurement rather than about the synthesis, and it cannot distinguish a good generation from a bad one. A useful diagnostic is therefore to compare the score of the synthetic output with the score of its source signal: a moderate rather than near-perfect association indicates that the score is sensitive to what the generator did, which is the property that lets it filter poor generations rather than merely poor measurements. This kind of decomposition is rarely reported but is essential to interpreting what a confidence score actually measures (Alaa et al., 2022; Stenger et al., 2024).

It is worth making the cost structure of gating explicit, because it is what converts a confidence score into a decision rule. Suppose that, on representative data, a system can estimate two quantities as a function of the gating threshold: the fraction of outputs retained, and the realised decision cost among the retained outputs. Lowering the threshold retains fewer outputs but improves their average reliability, while raising it retains more outputs at the price of admitting riskier ones. The operating point is therefore a trade-off between coverage and

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

reliability, and the right point depends on what the system does with a deferred output: if deferral routes the case to a clinician or to a confirmatory test, aggressive gating is cheap; if deferral means the patient is simply not screened, it is expensive. Reporting the whole coverage-reliability curve, rather than a single accuracy number, is the honest way to communicate what a confidence instrument buys, and it mirrors the selective-prediction framing that the uncertainty literature has developed for classifiers (Abdar et al., 2021; Gawlikowski et al., 2023).

A concrete reading of the running example makes the analysis tangible. In wearable atrial-fibrillation screening, adding noise to a clean signal degrades the downstream classifier, and denoising recovers much but not all of the lost performance; the residual gap is exactly the population of generations that the confidence layer should catch. When the score is well grounded, restricting action to the lower-cost majority of denoised outputs recovers performance close to that achieved on uncorrupted signals, which is the empirical signature of a useful instrument: the gated subset behaves as though the corruption had never occurred. The pattern also exposes a class-dependent hazard, namely that the rarer, pathological class is typically where calibration is weakest and where confident errors concentrate, so aggregate reliability can flatter a system that is unreliable precisely on the cases that matter most (Pereira et al., 2020; Hong et al., 2020).

Finally, the analysis clarifies the relationship between confidence scoring and the more familiar idea of downstream utility, in which synthetic data are judged by the performance of a model trained or evaluated on them. Downstream utility is a corpus-level, retrospective summary; task-guided confidence is its per-instance, prospective counterpart. The two are consistent, and the framework can be seen as formalising the long-standing practice of using a downstream model to vet generated data, but it sharpens that practice into a per-output instrument with a decision-theoretic meaning (Shmelkov et al., 2018; Esteban et al., 2017). Seen this way, task-guided confidence is less a departure from existing practice than a disciplined version of what practitioners already do informally.

VI. DEPLOYMENT, GOVERNANCE, AND CLINICAL TRANSLATION

A confidence instrument that is sound in principle still has to operate inside a real monitoring system, and the conditions of deployment reshape what is feasible. Continuous health monitoring is increasingly carried out on wearable and edge devices that sit at the periphery of a larger information-processing fabric, an arrangement anticipated by the cyber-physical and Industry 4.0 visions of sensing systems that couple physical measurement tightly to computation (Lu, 2017; Lu, 2025). In that setting the synthetic-output pipeline and its confidence layer must run under tight power, memory, and latency budgets, because the device is small, battery-limited, and expected to operate for long periods without intervention. Table III summarises the principal deployment considerations and a practical response to each.

Table III. Deployment considerations for task-guided confidence scoring in continuous health monitoring.

Deployment consideration	Why it matters for confidence scoring	Practical response
Distribution shift at the edge	Uncertainty estimates degrade on unseen patients and acquisition settings, the regime where gating is most needed	Monitor score behaviour in subgroups; recalibrate against realised cost over time
Latency and on-device compute	Continuous monitoring runs on constrained wearable and edge hardware with tight power budgets	Keep the scoring computation lightweight; reserve heavy analysis for cloud or hybrid tiers
Threshold and decision cost	The accept/defer boundary encodes a clinical trade-off between missed disease and false alarms	Set the gating threshold from an explicit utility model agreed with clinical stakeholders
Governance and accountability	A deferred or overridden output must be auditable and explainable to clinicians and regulators	Log scores and gating decisions; expose the rationale

		at the point of care
Prospective validation	Retrospective grounding may miss failure modes that emerge after deployment	Pair retrospective calibration with prospective monitoring of decision outcomes

The first consideration is distribution shift at the edge. The whole value of a confidence score is realised when it encounters inputs unlike those seen in development, yet that is also where uncertainty estimates are least reliable, so the score most needed for gating is the one most likely to be miscalibrated (Ovadia et al., 2019). The practical response is to treat grounding as an ongoing activity rather than a one-time validation: monitor how the score behaves within clinically meaningful subgroups, and recalibrate against realised decision cost as new outcomes accrue. The general survey literature on artificial intelligence has long stressed that models must be maintained against drift rather than deployed and forgotten (Lu, 2019; Zhang & Lu, 2021).

The second consideration is latency and on-device computation. Because the scoring signal in the framework is read from the downstream task model rather than from the generator, it can be made lightweight, which is fortunate given the constraints of wearable hardware. Where heavier analysis is required, a tiered architecture can keep time-critical gating on the device while delegating periodic recalibration and audit to a hybrid or cloud tier. The connectivity and resource-allocation questions raised by such tiered deployments connect to a substantial body of work on next-generation networks and on the economics of cloud provisioning for distributed intelligent systems (Lu & Zheng, 2020; Lu et al., 2020). The architecture also rests on secure device-to-cloud communication, an area where the security of the underlying internet-of-things substrate is a recognised precondition rather than an afterthought (Lu & Xu, 2019; Xu et al., 2021).

The third consideration is the gating threshold and its relationship to decision cost. The accept-or-defer boundary is where the abstract loss function meets operational reality, and it encodes a clinical judgement about the relative harm of acting on a bad output versus discarding a usable one. Setting it well is a decision-analytic exercise that should involve the people who own the clinical trade-off, and it benefits from the same management-analytic framing that treats threshold choice as a deliberate optimisation under stated preferences rather than a default (Lu et al., 2024a; Lu et al., 2024b). A threshold chosen without an explicit utility model will silently embed whatever trade-off happens to be convenient, which is precisely the opacity that clinical governance is meant to prevent.

The fourth consideration is governance and accountability. When a system defers or overrides an output, that action must be explainable to clinicians and auditable by regulators, and the confidence score and gating decision must therefore be logged and surfaced rather than hidden inside the pipeline. The clinical-AI literature has been sharply critical of post-hoc explanations that rationalise a decision without faithfully reflecting how it was made, and a confidence instrument should not add to that problem (Ghassemi et al., 2021; Challen et al., 2019). A virtue of task-guided confidence in this respect is that its meaning is transparent: the score is the expected cost of the decision, and the threshold is the agreed tolerance for that cost, so the rationale for gating can be stated in terms a clinician can interrogate.

The fifth consideration is prospective validation. A retrospective demonstration that the score tracks decision cost on historical data is necessary but not sufficient, because new failure modes appear after deployment when the system meets patients, devices, and settings that the development data never contained (Futoma et al., 2020; Kelly et al., 2019). The recommended posture is to pair the retrospective grounding of Section IV with prospective monitoring of decision outcomes, so that the confidence instrument is audited continuously against the very quantity it is meant to predict. This is the deployment analogue of the grounding step, and it is what keeps a confidence score honest over the life of a system.

Two broader translational points cut across the table. First, the methods that surround confidence scoring are themselves evolving quickly, and emerging computational paradigms, including quantum approaches to machine learning, may change the cost and capability of the models that generate and consume synthetic signals;

a confidence framework defined at the level of decisions rather than implementations is robust to such change (Lu et al., 2024c). Second, the infrastructural maturity that makes continuous monitoring possible, from sensing to connectivity to analytics, is the same maturity that makes confidence scoring necessary, because it places model-mediated signals in front of consequential decisions at a scale that manual oversight cannot match (Lu, 2025; Charlton et al., 2023).

VII. DISCUSSION AND RESEARCH ROADMAP

The perspective developed here reframes a familiar engineering instinct, that one should check whether a downstream model still works when fed a generated signal, as a principled instrument with a decision-theoretic meaning and an explicit grounding procedure. The reframing matters because it changes what counts as evidence of quality. Under the task-guided view, the question is never whether a synthetic time series is realistic in the abstract, but whether using it will degrade the specific decision it informs, and that question can be answered per output, before ground truth, and in the units of the decision itself. This is a more demanding standard than distributional similarity, and a more useful one (Smith et al., 2024; Begoli et al., 2019).

Several open problems follow directly from the framework. The first concerns losses that resist per-instance expression. Misclassification cost yields a clean per-output confidence signal, but many clinically important objectives, such as the balance of sensitivity and specificity captured by an F-style score, are defined over populations and do not decompose neatly into a single instance's contribution. Constructing per-instance confidence signals for such objectives, or principled approximations to them, is an open methodological challenge that the broader uncertainty-quantification literature has only begun to address (Gawlikowski et al., 2023; Angelopoulos & Bates, 2023).

The second problem concerns confidence under realistic, structured corruption. Much of the evidence base relies on simple additive noise to create the domain gap that adaptation then closes, whereas deployment confronts motion, contact loss, and physiological confounders whose statistics are far richer (Charlton et al., 2023; Pereira et al., 2020). Whether confidence scores grounded under simple noise remain grounded under realistic corruption is an empirical question that deserves systematic study, and it connects to the wider concern that synthetic-data evaluation must be stress-tested rather than reported on convenient benchmarks (Stenger et al., 2024).

The third problem concerns the joint behaviour of generator and task model under shift. Because the confidence signal is read from the task model, its reliability depends on the task model's calibration, which itself degrades out of distribution; the two failure modes can compound, so that a hallucinated synthetic signal is scored by a miscalibrated classifier (Ovadia et al., 2019; Kouw & Loog, 2021). Disentangling these contributions, and designing confidence instruments that remain informative when both the generator and the task model are stressed, is a substantial research agenda. Methods that detect out-of-distribution inputs upstream of the task model may have a role here (Hendrycks & Gimpel, 2017).

The fourth problem is methodological and concerns evaluation protocol. The field would benefit from reporting conventions that treat per-instance grounding, subgroup reliability, and the score's sensitivity to generator-induced change as first-class results rather than afterthoughts, in the same way that calibration and robustness have become expected reports for predictive models (Guo et al., 2017; Hong et al., 2020). A shared protocol would make confidence instruments comparable across studies and would discourage the optimistic reporting that the clinical-translation literature has repeatedly criticised (Futoma et al., 2020; Kelly et al., 2019).

A fifth direction concerns the human factors of confidence-aware systems, which are easy to neglect and decisive in practice. A score is only useful if the people who receive it understand what it means and act on it appropriately, and there is a well-documented risk that decision-makers either defer reflexively to an automated signal or ignore it entirely. The clinical-AI literature warns that confidence and explanation can both be misused, lending unwarranted authority to a number or rationalising a decision after the fact, and a task-guided score is

RESEARCH ARTICLE
Vol. 1, No. 4 (2023), pp. 36–51
DOI: 10.63646/jaiaa.2023.010403

not immune to these failure modes simply because its definition is principled (Ghassemi et al., 2021; Challen et al., 2019). Designing the presentation of confidence so that it calibrates human trust rather than displacing human judgement is a research problem in its own right, and one that the broader literature on artificial intelligence and on management analytics for decision support is well placed to inform (Lu, 2019; Lu et al., 2024b).

A sixth problem is quantitative and concerns how the gating threshold should be set when the two error types carry very different costs, as they almost always do in screening. A missed atrial-fibrillation episode and a false alarm are not interchangeable, and the expected-cost formulation makes this asymmetry explicit by allowing the loss attached to each outcome to be weighted before the threshold is chosen. The practical consequence is that the operating point cannot be read off a single accuracy number; it follows from a cost-weighted sweep across candidate thresholds, in which each choice fixes both the fraction of synthetic outputs retained and the cost-weighted error among them. Reporting that full curve, rather than one summary statistic, is the honest analogue of the coverage-risk reporting that the selective-prediction literature has made standard, and it lets a clinical team locate the operating point that matches its own tolerance for the two kinds of mistake (Charlton et al., 2023; Kelly et al., 2019).

A seventh consideration is that many clinically consequential decisions are not binary, and the convenient mathematics of the binary case does not survive the move to richer action spaces. When the action set has more than two members, the entropy of the task model's prediction is no longer a monotone function of the expected decision cost, so it can rank two outputs in the opposite order to the quantity that actually matters. The remedy is to compute the expected loss directly from the predicted distribution and the cost structure, rather than to lean on entropy as a proxy, even though doing so forgoes the analytical convenience that makes the binary case attractive. This is less a limitation of the framework than a reminder that the confidence signal must be derived from the loss that governs the decision, and that any shortcut is valid only while the property that justifies it continues to hold (Gawlikowski et al., 2023; Kendall & Gal, 2017).

An eighth question concerns transportability: a confidence instrument validated at one site, on one device, and in one patient population may not retain its reliability when any of those conditions changes. Because the score inherits whatever calibration the downstream task model possesses, and because that calibration is known to degrade under distribution shift, an instrument that looks dependable in a retrospective study can quietly lose its grounding when a new sensor or an unfamiliar cohort enters the stream. The corpus surveyed here suggests that the safest posture is to treat grounding as a property that must be re-established whenever the deployment context moves materially, and to pair the initial retrospective check with prospective monitoring of realised decision cost, so that drift in the relationship between score and outcome is caught before it propagates into care (Futoma et al., 2020; Pereira et al., 2020; Hong et al., 2020; Lu, 2025).

Looking further ahead, the convergence of generative modelling with continuous, sensor-rich monitoring suggests that synthetic signals will become more, not less, central to health machine learning, and that the decisions they inform will grow in consequence (Topol, 2019; Rajkomar et al., 2019; Perez et al., 2019). As composite systems combine components that mature at different rates, a confidence framework anchored to decisions rather than to particular models or modalities offers a durable point of reference, because it asks the same question, what will it cost to be wrong, regardless of how the underlying technology evolves (Lu et al., 2024c; Lu, 2025). The research roadmap above is, in effect, a programme for making that question answerable in ever more realistic settings.

VIII. CONCLUSION

Synthetic time series have become a quiet but load-bearing element of health-oriented machine learning, standing in for measurements that were never cleanly observed and feeding decisions that matter for patients. This review has argued that the trustworthiness of such an output cannot be judged apart from the decision it informs, and it has developed a task-guided confidence scoring perspective in response. The central claims are

that conventional distributional and realism metrics answer the wrong question for deployment; that a deployment-relevant confidence signal must be per-instance, available before ground truth, and aligned with the decision at hand; that such a signal can be derived from the downstream task model and grounded by checking that higher scores track higher realised decision cost; and that the resulting scores enable principled gating of low-confidence outputs.

Framed this way, confidence scoring is not an add-on to a generative pipeline but the mechanism that makes the pipeline safe to use, and its value is greatest exactly where the system is most exposed, on unfamiliar patients, devices, and settings. The wearable photoplethysmography example used throughout shows both the promise and the difficulty: the framework gives a clear account of when a denoised signal is safe to act on, while the open problems of non-decomposable losses, realistic corruption, compounding shift, and evaluation protocol mark the work that remains. We hope the perspective offers a transferable diagnostic, applicable well beyond the running example, for deciding when a synthetic time-series output deserves to be trusted.

ACKNOWLEDGEMENT

The authors thank colleagues who provided feedback on the framing of this review. This work did not involve the collection of new human-participant data.

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

Alaa, A., Van Breugel, B., Saveliev, E. S., & van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Proceedings of the 39th International Conference on Machine Learning (pp. 290–306). https://doi.org/10.48550/arXiv.2102.08921

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

Aschbacher, K., Yilmaz, D., Kerem, Y., Crawford, S., Benaron, D., Liu, J., Eaton, M., Tison, G. H., Olgin, J. E., & Li, Y. (2020). Atrial fibrillation detection from raw photoplethysmography waveforms: A deep learning application. Heart Rhythm O2, 1(1), 3–9. https://doi.org/10.1016/j.hroo.2020.02.002

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

Brophy, E., Redmond, P., Fleury, A., De Vos, M., Boylan, G., & Ward, T. (2022). Denoising EEG signals for real-world BCI applications using GANs. Frontiers in Neuroergonomics, 2, 805573. https://doi.org/10.3389/fnrgo.2021.805573

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

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

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

Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv. https://doi.org/10.48550/arXiv.1706.02633

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health, 2(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

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

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

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

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330). https://doi.org/10.48550/arXiv.1706.04599

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

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1610.02136

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Advances in Neural Information Processing Systems (Vol. 30). https://doi.org/10.48550/arXiv.1706.08500

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

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1125–1134). https://doi.org/10.1109/CVPR.2017.632

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). Synthetic data — what, why and how? arXiv. https://doi.org/10.48550/arXiv.2205.03257

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

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30). https://doi.org/10.48550/arXiv.1703.04977

Kiyasseh, D., Tadesse, G. A., Thwaites, L., Zhu, T., & Clifton, D. (2020). PlethAugment: GAN-based PPG augmentation for medical diagnosis in low-resource settings. IEEE Journal of Biomedical and Health Informatics, 24(11), 3226–3235. https://doi.org/10.1109/JBHI.2020.2979608

Kouw, W. M., & Loog, M. (2021). A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3), 766–785. https://doi.org/10.1109/TPAMI.2019.2945942

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30). https://doi.org/10.48550/arXiv.1612.01474

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

Lu, Y. (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. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

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. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

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

Lu, Y., & Zheng, X. (2020). 6G: A survey on technologies, scenarios, challenges, and the related issues. Journal of Industrial Information Integration, 19, 100158. https://doi.org/10.1016/j.jii.2020.100158

Lu, Y., Ivanov, L. A., Wang, F., Pisarenko, Z. V., & Ye, C. (2024a). Management analytics: A bibliometric analysis. Nanotechnologies in Construction, 16(3), 257–266. https://doi.org/10.15828/2075-8545-2024-16-3-257-266

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024b). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431–440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334–351. https://doi.org/10.1080/17517575.2019.1669827

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems (Vol. 32). https://doi.org/10.48550/arXiv.1906.02530

Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., Colorado, R., Meisel, K., & Hu, X. (2020).

Photoplethysmography based atrial fibrillation detection: A review. npj Digital Medicine, 3, 3. https://doi.org/10.1038/s41746-019-0207-9

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

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

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Shmelkov, K., Schmid, C., & Alahari, K. (2018). How good is my GAN? In Proceedings of the European Conference on Computer Vision (pp. 213–229). https://doi.org/10.1007/978-3-030-01216-8_14

Smith, F. B., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., & Rainforth, T. (2024). Rethinking aleatoric and epistemic uncertainty. arXiv. https://doi.org/10.48550/arXiv.2412.20892

Stenger, M., Leppich, R., Foster, I., Kounev, S., & Bauer, A. (2024). Evaluation is key: A survey on evaluation measures for synthetic time series. Journal of Big Data, 11, 66. https://doi.org/10.1186/s40537-024-00924-7

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

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, Z., & Holmes, C. (2024). On subjective uncertainty quantification and calibration in natural language generation. arXiv. https://doi.org/10.48550/arXiv.2406.05213

Xu, B., Liu, R., Shu, M., Shang, X., & Wang, Y. (2021). An ECG denoising method based on the generative adversarial residual network. Computational and Mathematical Methods in Medicine, 2021, 5527904. https://doi.org/10.1155/2021/5527904

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

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