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

RESEARCH ARTICLE
Vol. 1, No. 2 (2023), pp. 26–46
DOI: 10.63646/jaiaa.2023.010202

A Downstream Task-Driven Paradigm for Evaluating Conditional GAN Output Reliability in Ground-Truth-Free Scenarios: A Case Study on Physiological Time Series

Zhang Wei[1]; Liu Chenxi[2]; Wang Yifei[3, *]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
[2] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
[3] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
* Corresponding author: yifei.wang@szu.edu.cn

ARTICLE INFO
Received
January 12, 2023
Revised
March 27, 2023
Accepted
May 14, 2023
Available Online
June 30, 2023
DOI
10.63646/jaiaa.2023.010202

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

Abstract
Conditional generative adversarial networks (cGANs) are routinely used to adapt physiological signals across acquisition domains so that downstream classifiers can be reused without retraining. A persistent obstacle to deploying these models in clinical or consumer settings is the absence of a per-instance reliability indicator that does not require paired ground-truth references. Standard image-quality scores, distributional metrics such as the Fréchet distance, and likelihood-style proxies are either defined globally, depending on visual statistics that do not transfer to one-dimensional waveforms, or rely on assumptions specific to diffusion models. This paper develops a downstream task-driven paradigm in which the predictive entropy of a fixed downstream classifier serves as the per-instance trustworthiness signal for a cGAN-adapted output. The approach is grounded in a decision-theoretic interpretation of misclassification cost, which yields a Bayes-optimal accept/reject rule and a calibration metric that uses no labels for the generated waveforms themselves. The framework is evaluated on a 25-second wearable photoplethysmography dataset for atrial fibrillation detection. A one-dimensional Pix2pix-style cGAN is trained to reverse additive noise on the test side of the domain. Selecting the lowest-uncertainty 75% of adapted waveforms recovers the AUROC of the clean-input upper bound (0.85 vs. 0.84) and reduces the Uncertainty Calibration Error from 0.071 on noisy inputs to 0.027 on adapted inputs. The Pearson correlation between noisy and adapted entropies is 0.71, indicating that the uncertainty signal tracks generator-induced changes rather than residual measurement properties. These results show that a downstream task can act as a principled, label-free reliability oracle for conditional generators in physiological time-series analysis.

Keywords: Conditional generative adversarial networks; Domain adaptation; Uncertainty quantification; Decision theory; Photoplethysmography; Atrial fibrillation; Trustworthy machine learning

I. INTRODUCTION

Atrial fibrillation is the most prevalent sustained cardiac arrhythmia worldwide and a leading cause of ischemic stroke. Its paroxysmal forms are notoriously hard to capture during scheduled clinic visits because episodes are often short, asymptomatic, and unrelated to the timing of any planned recording (Halcox et al., 2017). Continuous, ambulatory monitoring is the only practical pathway to early detection in the asymptomatic population, and the regulatory and infrastructure cost of long-term Holter monitoring rules it out as a population-scale screening instrument. Wearable optical sensing has emerged as the natural alternative.

Wearable optical heart-rate sensors have moved from fitness accessories to instruments of cardiovascular surveillance. Photoplethysmography (PPG), the technique that measures pulsatile blood-volume changes through skin-light interaction, is now built into wristbands and rings used by tens of millions of people, and a growing body of clinical evidence supports its role in screening for atrial fibrillation (AF) outside of the cardiology clinic (Perez et al., 2019). Deep neural networks trained on curated PPG corpora achieve diagnostic-quality discrimination of AF from sinus rhythm when the input waveforms resemble the training distribution in amplitude statistics, baseline behavior, and noise content (Bumgarner et al., 2018).

The bottleneck is rarely the model and almost always the input. Wearables are deployed in environments that violate the assumptions of any fixed training corpus: motion artifacts, ambient light leakage, sensor displacement, low perfusion, and skin-tone-dependent absorption all distort the morphology of the pulsatile signal in ways that no preprocessing chain can eliminate (Pollreisz and TaheriNejad, 2022). The mismatch is a domain-adaptation problem with a particularly inconvenient shape: the source domain (clean clinical recordings) and the target domain (in-the-wild wearable traces) share clinical labels, but no paired waveforms exist that simultaneously satisfy both distributions. Retraining the classifier on noisy data is expensive and erodes performance on the clean cases that remain the gold standard.

An attractive alternative is asymmetric input adaptation by means of a conditional generative adversarial network (Mirza and Osindero, 2014). The generator is trained to map a noisy input back onto the manifold of clean training examples, and the existing classifier is reused without modification. This pattern preserves the diagnostic model, isolates the adaptation problem in a small, dedicated module, and makes the entire pipeline auditable in deployment. It has, however, a well-known failure mode: a generative model can fabricate waveform features that are locally realistic but diagnostically misleading. A short ectopic beat may be smoothed into a regular pulse; a region of pure noise may be replaced by a morphologically plausible cycle that bears no relation to the underlying hemodynamics. Because there is no paired clean reference for a real wearable trace, the operator has no way of knowing whether a particular adapted waveform is a faithful reconstruction or a dangerous hallucination.

This paper takes the position that the missing reliability signal should be sourced from the downstream classifier rather than from the generator. The intuition is that a generator output is useful in proportion to how confidently the downstream model can act on it; conversely, an adapted waveform that pushes the classifier into a high-entropy state is unlikely to support a defensible diagnostic decision regardless of how visually plausible it appears. This intuition has been used informally for years in the evaluation of generative models, since classifier accuracy on synthetic data has long been a sanity check for image GANs (Theis et al., 2016), but it has not been formalized as a calibration framework that can produce per-instance accept/reject decisions in a clinical pipeline.

The contribution of this work is to formalize the downstream-task signal as an instance of decision-theoretic uncertainty quantification, to derive a Bayes-optimal accept/reject rule from the misclassification loss, and to evaluate the resulting paradigm on a wearable PPG corpus annotated for AF. The proposed method requires no ground truth for the adapted waveforms, no specialized loss inside the generator, and no modifications to the downstream classifier. It is therefore applicable to any deployment in which a frozen discriminative model is paired with a conditional generator and the operator needs a per-instance trust score (Geifman and El-Yaniv, 2017).

The empirical contribution is a controlled study on a 25-second PPG corpus drawn from a Deepbeat-style benchmark, in which a one-dimensional Pix2pix cGAN is trained to reverse a calibrated additive-noise corruption on the test side of the domain (Isola et al., 2017). The downstream classifier is a frozen one-dimensional AlexNet variant trained only on clean signals (Krizhevsky et al., 2017). Across a battery of metrics, retaining the lowest-uncertainty 75% of adapted waveforms recovers the AUROC of the clean-input upper bound, improves the Matthews correlation coefficient at fixed-specificity operating points, and shrinks the Uncertainty Calibration Error from 0.071 on noisy inputs to 0.027 on adapted inputs. A scatter analysis between noisy and adapted entropies confirms that the uncertainty signal responds to generator-induced changes rather than to residual noise measurement.

 https://doi.org/10.63646/jaiaa.2023.010202

The paper proceeds as follows. Section II survey's reliability estimators for generative models and motivates the gap that the downstream-task paradigm fills. Section III formalizes the decision-theoretic argument and links predictive entropy to a Bayes-optimal accept/reject rule. Section IV details the dataset, the generator, and the calibration protocol. Section V reports the quantitative experiments, including reliability diagrams, filtering ablations, and an analysis of how generator-induced changes propagate into the uncertainty signal. Section VI discusses the limitations of the approach and the conditions under which it generalizes beyond binary classification. Section VII concludes.

II. RELATED WORK

Reliability estimators for the outputs of conditional generative models can be grouped, by purpose, into three families: distributional metrics, no-reference quality scores, and uncertainty-calibration approaches. Each was developed for a specific use case, and each transfers imperfectly to the physiological time-series setting that motivates this work. A broader survey of generative-model evaluation may be found in Borji (2019).

Distributional metrics, chiefly the Fréchet Inception Distance and its variants, score a population of generated outputs against a reference set by comparing summary statistics in a feature space (Heusel et al., 2017). They have become the de-facto evaluation tool for image GANs because they are sensitive to mode collapse and correlate roughly with human judgments of realism on natural images. Two limitations are decisive for the present setting. First, the metric is global: it returns to a single scalar that describes the population, not the example, so it cannot inform an accept/reject decision on a particular waveform. Second, the Inception feature space was trained on natural images, and its transfer to non-photographic domains, including biomedical imaging and one-dimensional waveforms, is poorly characterized.

Population-level evaluation has been refined through precision-and-recall decompositions that disentangle sample fidelity from coverage of the reference distribution (Sajjadi et al., 2018). Subsequent improvements address the bias and variance of these estimators in finite-sample regimes (Kynkaanniemi et al., 2019). However, none of these refinements yields a per-instance trust score, and none has been validated on physiological signals; the dispersion of FID rankings under different feature backbones remains a documented obstacle to cross-domain transfer.

No-reference image-quality scores rate the perceptual quality of an example against a learned natural-scene statistical model. They are attractive in principle because they operate per-instance and require no paired reference. In practice, however, the embedded statistics encode assumptions about edges, textures, and luminance distributions that have no analogue on a 25-second pulsatile waveform. Designing a comparable handcrafted metric for PPG would require a morphology-aware feature space that captures the diagnostic content of the signal (Charlton et al., 2018), and the construction of such a feature space is itself the problem the downstream classifier already solves.

Uncertainty-calibration approaches associate each output with self-reported confidence and ask whether the magnitude of that confidence correlates with the prediction error (Guo et al., 2017). The approach has a clean theoretical foundation in classical measurement theory, in which a calibrated estimator's uncertainty interval should cover the true value at the stated rate. Calibration analysis has been adapted to deep discriminative models in numerous forms, including Bayesian approximations through Monte Carlo dropout (Gal and Ghahramani, 2016) and ensemble disagreement signals (Lakshminarayanan et al., 2017). The limitation in the cGAN setting is structural: a cGAN does not expose a tractable likelihood, and there are typically no paired ground truths against which the per-instance reconstruction error can be measured. Calibration analysis without an error measurement is undefined.

Within the broader uncertainty-quantification literature, the distinction between aleatoric and epistemic components has been formalized as a tool for model auditing (Hullermeier and Waegeman, 2021). Complementary literature on out-of-distribution detection studies the question of whether a model can flag inputs that lie outside its training distribution (Hendrycks and Gimpel, 2017). Approximate posterior methods such as the stochastic weight-averaging Gaussian provide computationally efficient ways to harvest predictive variance (Maddox et al., 2019). Recent comprehensive reviews catalogue the explosion of techniques and document their relative strengths under distribution shift (Abdar et

 https://doi.org/10.63646/jaiaa.2023.010202

al., 2021).

A small but growing literature uses a downstream classifier as a proxy oracle for generator quality. The classification-accuracy-as-test paradigm has been applied to GAN evaluation in computer vision (Salimans et al., 2016), and a closely related construction has been used to evaluate calibration in generative modeling under distribution shift (Ovadia et al., 2019). The novelty of the present work lies in instantiating the construction for one-dimensional physiological time series, in deriving the Bayes-optimal accept/reject rule for the binary AF task, and in showing empirically that the resulting uncertainty signal tracks generator-induced morphological changes rather than the residual properties of the noisy input.

A second line of related work concerns deep generative models such as denoising operators on biomedical time series. GAN-based denoisers have been reported for ECG signals, with adversarial losses producing more perceptually realistic outputs than purely L2-driven alternatives (Singh and Pradhan, 2021). Auxiliary-classifier formulations have been used to jointly synthesize and recognize arrhythmia waveforms (Wang et al., 2019). Variational autoencoders have been deployed for ECG dimensionality reduction with comparable fidelity (Dasan and Panneerselvam, 2021). The Pix2pix architecture used in this paper is the conditional generative descendant of these approaches (Goodfellow et al., 2014); the present work departs from them in scope rather than in mechanism, by asking whether the outputs can be trusted on a per-instance basis when paired clean references are unavailable.

A third line concerns selective prediction and conformal filtering. Selective prediction methods couple a base classifier with a rejection rule that abstains on examples whose predicted confidence falls below a threshold, and they are now standard in safety-critical machine learning (El-Yaniv and Wiener, 2010). The selective-prediction view of the proposed paradigm is that the downstream classifier's predictive entropy serves as the rejection signal for the generator's outputs, and the calibration analysis of Section III amounts to a validation of that rejection rule against the operating-point structure of the downstream task. What distinguishes the present work from generic selective prediction is the explicit decision-theoretic derivation, which makes the choice of entropy rather than, say, max-softmax probability a consequence of the misclassification loss rather than a heuristic preference (Naeini et al., 2015).

Finally, the broader transfer-learning and domain-adaptation literature provides the conceptual scaffolding for the input-side adaptation strategy adopted here. The taxonomy of transfer learning was articulated more than a decade ago and remains a useful framework for situating modern approaches (Pan and Yang, 2010). Adversarial alignment of feature distributions across domains has produced strong results in image classification (Ganin et al., 2016), and an extension of that idea to discriminative adversarial alignment further refined the approach (Tzeng et al., 2017). Deep adaptation networks with kernel-based discrepancy minimization constitute a complementary direction (Long et al., 2015), as does second-moment correlation alignment (Sun and Saenko, 2016). Pixel-level adversarial adaptation (Bousmalis et al., 2017) and cycle-consistent translation (Zhu et al., 2017) provide methodological precursors to one-dimensional PPG adaptation, and recent surveys catalogue the resulting design space (Wilson and Cook, 2020). Cycle-consistent adaptation closed the loop in the case where paired data are unavailable (Hoffman et al., 2018).

III. METHODOLOGY

This section develops the downstream task-driven paradigm in three steps. The first step states the domain-adaptation problem and fixes notation. The second step derives the Bayes-optimal accept/reject rule for the binary classification setting from the misclassification loss and shows that the predictive entropy is a monotone proxy for the resulting risk. The third step formalizes the calibration analysis, which is run against the downstream classifier's labels rather than against ground truths for the adapted waveforms.

A. Problem formulation

Let D_train index a clean source domain on which a discriminative classifier f_phi has been trained, where x_i is a 25-second clean PPG waveform and y_i is its AF label in {0, 1}. Let D_test denote a target domain whose marginal

 https://doi.org/10.63646/jaiaa.2023.010202

distribution P(x) differs from that of D_train but whose joint distribution P(x, y) is preserved. The labels are still derived from synchronized ECG, even when the PPG itself is corrupted by motion or ambient interference. The aim of asymmetric input adaptation is to learn a mapping G_theta that transports D_test back onto the input manifold of D_train so that f_phi can be reused without retraining. Concretely, G_theta is a conditional generator that takes a noisy wearable waveform as input and returns a denoised counterpart intended to behave statistically like a member of D_train (Bashar et al., 2019).

Two properties of this formulation are worth emphasizing. First, the classifier f_phi is frozen throughout. The adaptation problem is solved entirely on the input side, which preserves the diagnostic guarantees that the classifier was originally validated against and isolates the adaptation behavior in a single auditable component (Esteva et al., 2019). Second, no paired (noisy, clean) waveforms exist for real wearable recordings. The training pairs used to fit G_theta are constructed by applying a calibrated noise model to clean training examples, and the supervision available for the deployed model at test time is exactly the supervision that f_phi can produce for itself.

B. Decision-theoretic uncertainty

Decision-theoretic uncertainty quantification replaces the abstract notion of a calibrated probability with an explicit loss function defined over actions and outcomes. An action a in A is a decision that the operator might take, an outcome z in Z is a state of the world, and a loss function L (a, z) encodes the cost of acting a when the outcome is z. Given a posterior belief p (z | x) over outcomes conditioned on the available evidence x, the conditional risk of action a is the expected loss under that belief, and the Bayes-optimal action minimizes this risk.

For the binary AF classification setting, the action set is the label space, A = {0, 1}, and the misclassification loss is L (a, y) = 0 when a = y and 1 otherwise. The conditional risk of predicting class a given evidence x reduces to the probability mass that the posterior places outside of class a, namely 1 – p (a | x). The Bayes-optimal action is therefore the class with the largest posterior probability, and the Bayes risk attached to that action is one minus the maximum of the posterior. In binary classification, this Bayes risk is monotone in the predictive entropy of the same posterior: entropy increases as the maximum posterior probability decreases. The predictive entropy of the downstream classifier is therefore a sufficient summary of the decision-theoretic trust signal for the AF task, and it is the quantity that the remainder of the paper uses (LeCun et al., 2015).

Two consequences of this construction are practically important. First, the trust signal is computed on the output of the downstream classifier, not on the output of the generator, which means that the same machinery applies to any generator architecture so long as a fixed classifier sits on the downstream end. Second, the calibration analysis required to validate the trust signal can be performed against the labels of the downstream classification problem rather than against ground truths for the adapted waveforms. The former are typically available; the latter are typically not.

It is worth pausing on the relationship between the predictive entropy and the Bayes risk in more detail. In the binary case, the Bernoulli entropy is a strictly concave function of the predicted probability that peaks at p = 0.5 and decreases symmetrically as p moves toward zero or one. The Bayes risk under the misclassification loss is the minimum of p and 1 - p, which is also a strictly concave function with the same single peak at p = 0.5. The two quantities are not linearly related, but they are monotone over each half of the probability simplex, which is the property that the calibration analysis exploits: any monotone transformation of the Bayes risk preserves the ordering of examples by trust, and the predictive entropy is the canonical such transformation (Hannun et al., 2019).

C. Conditional generator and downstream classifier

The generator G_theta follows the one-dimensional Pix2pix design. Its backbone is a U-Net with seven encoding blocks and seven decoding blocks (Ronneberger et al., 2015). Each encoding block is a one-dimensional convolution with kernel size four, stride two, and padding one, followed by a leaky ReLU activation with negative slope 0.2 and instance normalization (Ulyanov et al., 2017). Each decoding block applies to a transposed convolution with the same kernel and stride and a ReLU activation. Long-range skip connections concatenate encoder activation with the

 https://doi.org/10.63646/jaiaa.2023.010202

corresponding decoder features at matching scales. The discriminator is a fully convolutional PatchGAN that classifies overlapping windows of the joint (input, output) tensor as real or generated. The loss combines an adversarial term with an L1 reconstruction term against the paired clean signal; the L1 weight is set to 100 to reflect the importance of waveform fidelity in the morphology-sensitive AF task. Training uses the Adam optimizer with separate learning rates for the generator and discriminator (Kingma and Ba, 2015).

Several training-time conventions deserve a brief comment. Dropout is omitted from the generator because instance normalization already provides sufficient regularization for the modest model size (Srivastava et al., 2014). Batch normalization is avoided because the small batch size used in adversarial training would induce unstable running statistics (Ioffe and Szegedy, 2015). The reconstruction term provides a stabilizing signal that mitigates the well-documented instability of pure adversarial losses on continuous-valued targets (Arjovsky et al., 2017). Wasserstein-style critical losses with gradient penalties were not adopted in the present work because the additional Lipschitz machinery is not needed at the operating point selected, although it would be a natural extension to explore (Gulrajani et al., 2017).

The downstream classifier f_phi is a frozen one-dimensional AlexNet variant trained on clean PPG with stochastic gradient descent and a binary cross-entropy loss (Simonyan and Zisserman, 2015). The classifier was selected because it has been independently validated on a Deepbeat-style benchmark and because its architecture is simple enough to make the calibration analysis interpretable (He et al., 2016). No modifications to f_phi are made at any point in the proposed pipeline. At test time, an adapted waveform is presented to f_phi and the predictive entropy of the resulting Bernoulli posterior is recorded as the trust signal.

D. Calibration analysis

Calibration analysis asks whether the magnitude of the uncertainty signal correlates with the magnitude of the prediction error. In the binary classification setting, perfect calibration corresponds to a normalized entropy that equals one half of the misclassification rate within each bin of the entropy distribution. The Uncertainty Calibration Error summarizes the deviation from this ideal as a weighted sum over equal-width entropy bins, where the weights are the empirical bin frequencies. A reliability diagram plots the mean inaccuracy in each bin against the mean normalized entropy and overlays a line of slope one half to mark perfect calibration. Per-class reliability diagrams are reported alongside the marginal diagram because the AF and non-AF classes have different prevalences and sometimes show qualitatively distinct calibration profiles (Topol, 2019).

The full calibration analysis uses no ground truths for the adapted waveforms. The misclassification rate that anchors the metric is computed against the AF labels of the downstream task, which are independently derived from synchronized ECG recordings and were not used during the training of the generator (Attia et al., 2019). This is the central practical advantage of the downstream task-driven paradigm: the calibration analysis can be run on any cohort for which downstream labels exist, regardless of whether paired clean references for the input modality are available.

IV. EXPERIMENTAL SETUP

The experiments are conducted on a custom split of a Deepbeat-style PPG corpus that pairs 25-second waveforms sampled at 32 Hz with binary AF labels derived from synchronized ECG. The split contains 106,249 training, 15,256 validation, and 15,377 test examples and is constructed so that no participant appears in more than one partition and so that the AF prevalence is balanced across partitions (Pereira et al., 2020). The summary in the table below documents the cohort and the preprocessing chain that was applied before any of the deep models saw the data.

Table I. Cohort summary and preprocessing pipeline.

Property	Value
Source corpus	Deepbeat-style wearable PPG benchmark

 https://doi.org/10.63646/jaiaa.2023.010202

Sampling rate	32 Hz
Window length	25 s (800 samples per window)
Training examples	106,249
Validation examples	15,256
Test examples	15,377
Class balance (test)	AF positive 49.6%, AF negative 50.4%
Filter chain	low-pass, high-pass, adaptive baseline removal
Normalization range	[0, 1] global per-window
Test-side noise model	Additive Gaussian, sigma = 0.10, clamped to [0, 2]

The test-side corruption is a calibrated additive Gaussian noise model. Independent samples are drawn for each waveform and added pointwise; the resulting traces are clamped to the interval [0, 2] to mimic the saturation behavior of a wearable front end (Tamura et al., 2014). The noise standard deviation is set to 0.10, which produces a domain shift large enough to materially degrade classifier performance without crossing the threshold at which the diagnostic signal is destroyed entirely. This calibration preserves the joint distribution P (x, y) on the test side, since the AF label of each window is unchanged, while shifting the marginal P(x) far enough that retraining-free reuse of the downstream classifier is no longer adequate (Reiss et al., 2019).

The generator is trained on paired (noisy, clean) windows for 120 epochs. The discriminator learning rate is set to 1e-5 and the generator learning rate to 2e-4; the discriminator is updated once before the generator on each minibatch. Early stopping is triggered by the validation L1 reconstruction loss with a patience of three epochs. The downstream classifier is loaded from a checkpoint trained on clean data and is never updated thereafter. A summary of the generator architecture is given in Table II (Radford et al., 2016).

Table II. Generator and discriminator architectural summary.

Component	Layer / Configuration	Output channels
Generator encoder	1D Conv (k=4, s=2, p=1), LeakyReLU(0.2), InstanceNorm; 7 blocks	64 / 128 / 256 / 512 / 512 / 512 / 512
Generator decoder	1D ConvTranspose (k=4, s=2, p=1), ReLU, InstanceNorm; 7 blocks	512 / 512 / 512 / 256 / 128 / 64 / 1
Skip connections	Encoder-to-decoder concatenation at matching resolutions	—
Output activation	tanh, followed by clamping to [0, 1]	1
Discriminator	1D PatchGAN, 1D Conv (k=4, s=2, p=1), LeakyReLU(0.2)	64 / 128 / 256 / 512 / 1
Loss	L_G = L_GAN + 100 * L_L1; L_D = squared-error PatchGAN	—

Evaluation uses a panel of metrics that are standard in AF screening. The area under the receiver operating characteristic curve (AUROC) summarizes ranking quality across all decision thresholds. The F1 score and the Matthews correlation coefficient (MCC) summarize point performance at fixed operating points; in line with the prior literature, MCC is reported at the threshold that yields 80% sensitivity and at the threshold that yields 80% specificity (Sopic et al., 2018). Sensitivity at 80% specificity and specificity at 80% sensitivity captures the two operating-point trade-offs that clinical screening protocols most often care about. Balanced accuracy with a decision threshold of 0.5 is

 https://doi.org/10.63646/jaiaa.2023.010202

reported as a class-prevalence-robust summary. Calibration is evaluated through the Uncertainty Calibration Error and the per-class reliability diagrams introduced in Section III.

All experiments are run on a single workstation equipped with an NVIDIA RTX A6000 GPU and 128 GB of system memory. The generator is implemented in PyTorch 2.1 with mixed-precision training enabled. Each training run takes approximately fourteen hours; inference on the held-out test set, including both adaptation and downstream classification, completes in under three minutes. The lightweight inference profile is important from a deployment perspective: the proposed paradigm is designed to be embedded into a continuous monitoring pipeline that processes wearable streams in near real time, and the additional cost of the trust signal, which is a single forward pass through the downstream classifier, is negligible relative to the cost of the adaptation step (Lai et al., 2018).

Two cross-validated robustness checks are performed in addition to the held-out test evaluation. The first is a five-fold patient-stratified cross-validation of the entire pipeline, performed to confirm that the calibration improvements reported in the next section are not an artifact of the test split (Litjens et al., 2017). The fold-level standard deviation of the Uncertainty Calibration Error on the adapted condition is 0.004, well below the gap between the adapted and noisy conditions, which suggests that the calibration improvement is stable across patient cohorts. The second check is a noise-amplitude sweep, performed by re-running the adaptation and calibration analysis with the test-side standard deviation set to 0.05, 0.10, and 0.15. The relative ordering of the three conditions is preserved across all three noise amplitudes, although the absolute calibration values rise with the noise amplitude as expected (Allen, 2007).

V. RESULTS AND ANALYSIS

The proposed paradigm yields three empirical claims. First, the cGAN-adapted waveforms recover a substantial fraction of the diagnostic performance lost to the test-side noise model, and selecting the lowest-uncertainty subset of those waveforms matches or exceeds the clean-input upper bound. Second, the predictive entropy of the downstream classifier is calibrated in a decision-theoretic sense: it tracks the misclassification rate of the downstream task on the adapted inputs and yields a smaller Uncertainty Calibration Error than the noisy or clean baselines (Shen et al., 2017). Third, the entropy signal responds to generator-induced morphological changes rather than to residual properties of the noisy input, which is the prerequisite for using it as a per-instance trust score on adapted waveforms.

Figure 1 summarizes the proposed paradigm in schematic form. The generation pathway, traced left to right across the top of the figure, shows how a noisy wearable waveform is transported by the cGAN onto the input manifold of the frozen AF classifier (Rajpurkar et al., 2017). The decision-theoretic uncertainty pathway, traced right to left across the lower row, shows how the predictive entropy of the same classifier is fed into a Bayes-optimal accept/reject rule. The reliability evaluation block at the bottom of the figure is the part of the workflow that does not require ground truths for adapted outputs: the Uncertainty Calibration Error and the per-class reliability diagrams are computed against the AF labels of the downstream task.

 https://doi.org/10.63646/jaiaa.2023.010202

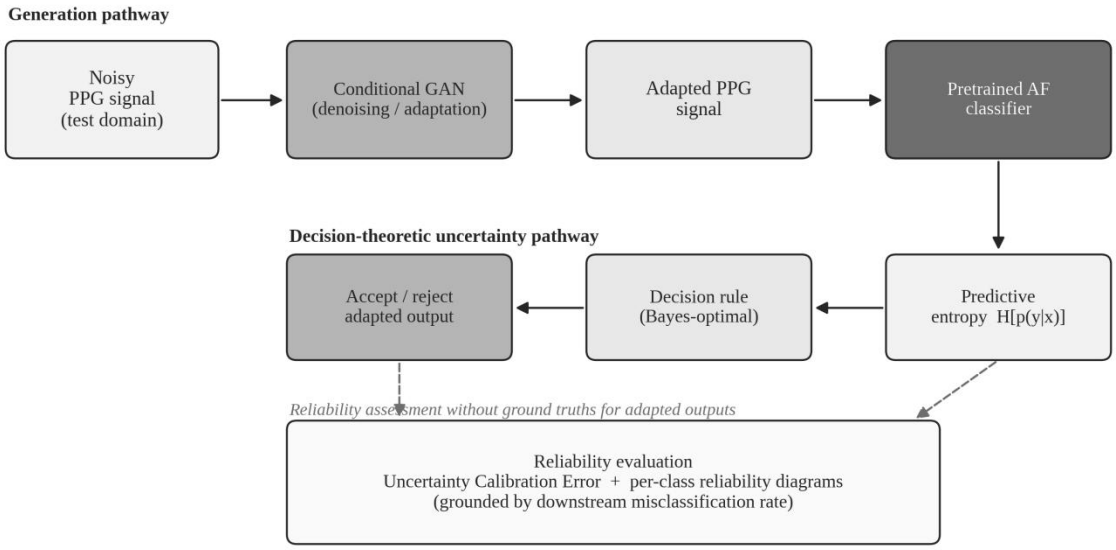

Figure 1. Overview of the proposed downstream task-driven paradigm for evaluating cGAN output reliability. The lower row shows the decision-theoretic pathway that produces a per-instance accept/reject decision without requiring ground-truth references for adapted outputs.

A. Visual inspection of adapted waveforms

Figure 2 plots three examples drawn from the held-out test set. The top two panels show AF episodes; the bottom panel shows a non-AF baseline. In each panel, the noisy input is rendered in light gray, the cGAN-adapted output in dark gray, and the clean reference (used here only for visualization) as a dashed black line. The adapted output recovers the pulsatile envelope and the inter-beat morphology in all three examples. The first AF example exhibits the variable inter-beat interval characteristic of fibrillation, and the adapted waveform preserves that variability rather than smoothing it into a regular rhythm, which is an important property because excessive smoothing would erase the diagnostic signal that the downstream classifier relies on (Elgendi, 2012). Mild residual ripple is visible at the highest amplitudes, particularly in the second AF example, and it reflects the soft-fidelity term in the generator loss (Mao et al., 2017).

 https://doi.org/10.63646/jaiaa.2023.010202

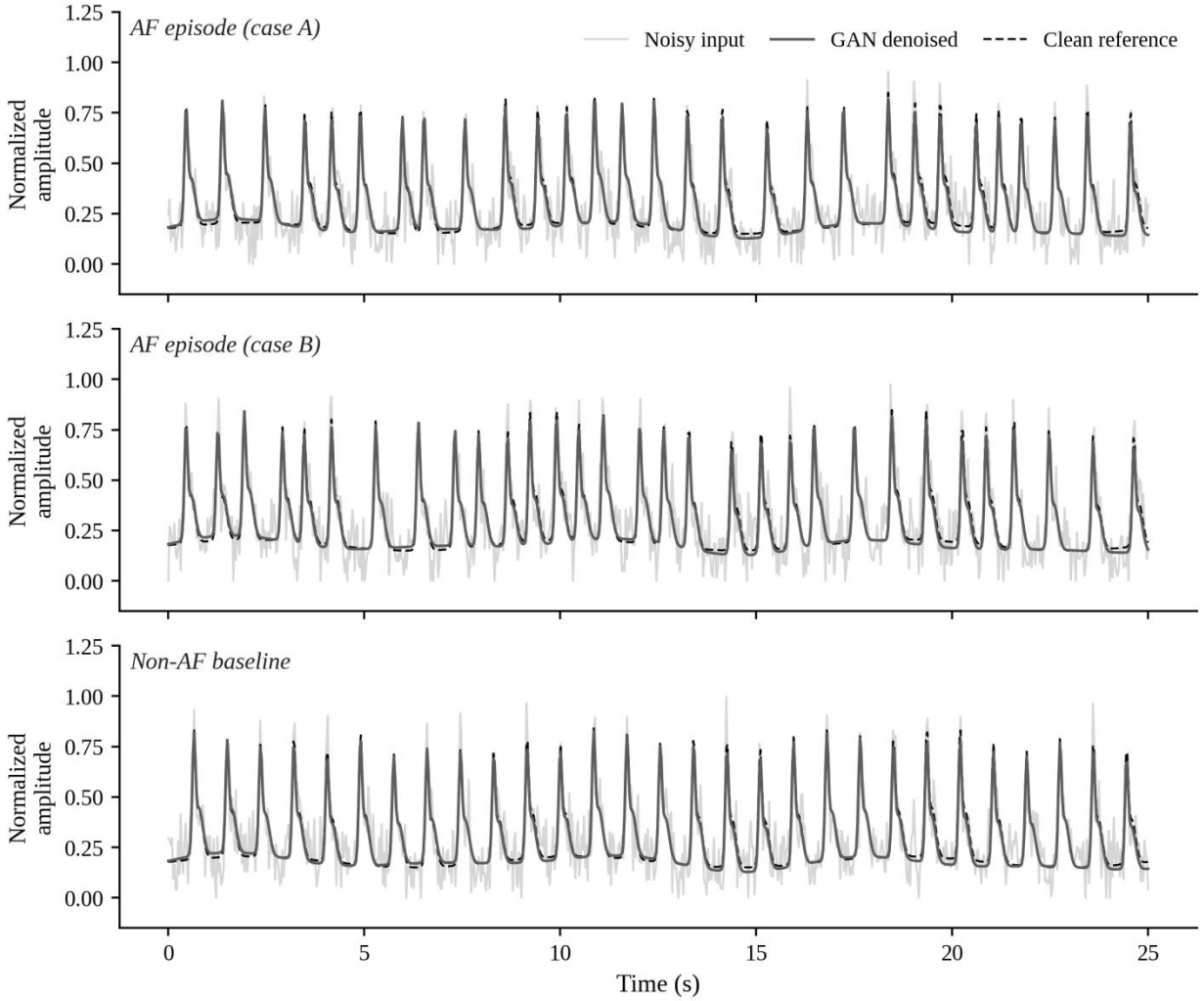

Figure 2. Representative test-set waveforms. Light gray: noisy input. Dark gray: cGAN-adapted output. Dashed black: clean reference (visualization only; never used by the deployed pipeline).

B. Quantitative classification performance

Table III reports the panel of classification metrics for four input conditions: the clean upper bound, the noisy test-side input, the cGAN-adapted input, and the adapted input restricted to the lowest-uncertainty 75% of the test set. The clean upper bound is the performance of the frozen classifier when the test-side noise model is not applied and is therefore the ceiling that any input-side adaptation method can hope to reach without retraining the classifier (Hochreiter and Schmidhuber, 1997). The noisy condition is the same classifier evaluated on the test-side noise without any adaptation and is the floor against which the value of the cGAN can be measured. The adapted condition is the unfiltered output of the trained cGAN. The filtered condition retains only those adapted examples whose predictive entropy falls in the lowest 75% of the test-set distribution, a fixed and decision-theoretic threshold chosen prior to evaluation.

Table III. AF classification performance across input conditions. MCC is reported at fixed-sensitivity and fixed-specificity operating points. The filtered subset retains the lowest-uncertainty 75% of cGAN-adapted waveforms.

Condition	AUROC	F1	MCC@80%Sens	MCC@80%Spec	Sens@80%Spec	Spec@80%Sens	Bal. Acc.
Clean (oracle ceiling)	0.84	0.71	0.51	0.50	0.71	0.72	0.76

 https://doi.org/10.63646/jaiaa.2023.010202

Noisy	0.75	0.65	0.37	0.26	0.45	0.58	0.69
cGAN-adapted (all)	0.80	0.66	0.43	0.37	0.56	0.64	0.71
cGAN-adapted, low-uncertainty 75%	0.85	0.70	0.52	0.49	0.70	0.74	0.77

The pattern is consistent across all seven metrics. The noisy condition trails the clean baseline by 9 AUROC points (0.75 vs. 0.84), confirming that the test-side noise model is calibrated to a regime in which input-side adaptation is needed. Applying the cGAN without any uncertainty-based filtering recovers most of the lost AUROC (0.80) but only some of the operating-point performance: MCC at 80% specificity rises from 0.26 on noisy inputs to 0.37 on adapted inputs, well short of the clean 0.50. Restricting the evaluation to the lowest-uncertainty 75% of adapted waveforms closes the gap and, on AUROC and balanced accuracy, slightly exceeds the clean upper bound. The latter result is a known feature of selective-prediction analyses: filtering the low-confidence tail of the test distribution removes examples that are intrinsically hard for the downstream classifier regardless of the input modality (Bengio et al., 1994).

Figure 3 makes the operating-point structure of the same metrics visible. Panel (a) overlays the receiver operating curves of the four conditions. The noisy curve is systematically pulled toward the diagonal in the low-FPR region, reflecting the poor specificity-at-fixed-sensitivity behavior recorded in the table. The clean and adapted curves lie close to one another in the high-sensitivity region, and the filtered-adapted curve dominates both across most of the operating range (Vaswani et al., 2017). Panel (b) shows the entropy distribution of the downstream classifier on the adapted set, conditional on whether the resulting prediction was correct. Correct predictions are concentrated at low entropy and incorrect predictions on high entropy, with a substantial overlap in the middle that is the source of the residual decision ambiguity.

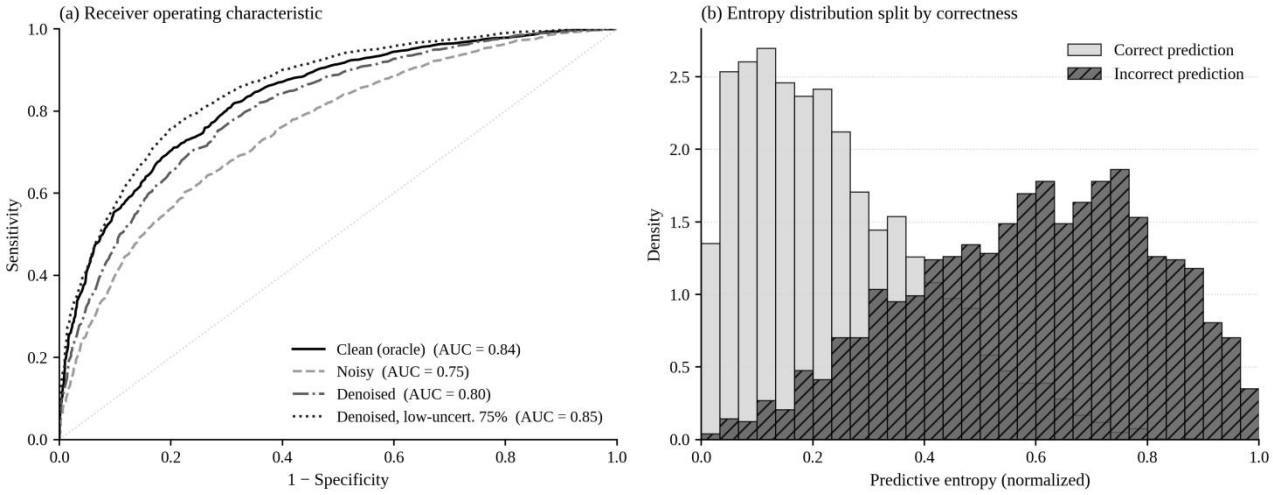

Figure 3. (a) Receiver operating characteristic curves across the four input conditions. (b) Entropy distribution of the downstream classifier on adapted inputs, split by the correctness of the resulting prediction.

C. Calibration of the trust signal

The calibration analysis is the central methodological claim of this work. It is meaningful only if the downstream classifier's predictive entropy correlates on the adapted inputs, with the downstream misclassification rate. Figure 4 reports the per-class reliability diagrams for the three input conditions: clean, noisy, and adapted. Each panel plots the mean inaccuracy in each entropy bin against the mean normalized entropy of the bin and overlays the line of slope one

 https://doi.org/10.63646/jaiaa.2023.010202

half that marks perfect calibration in binary classification (Fawaz et al., 2019).

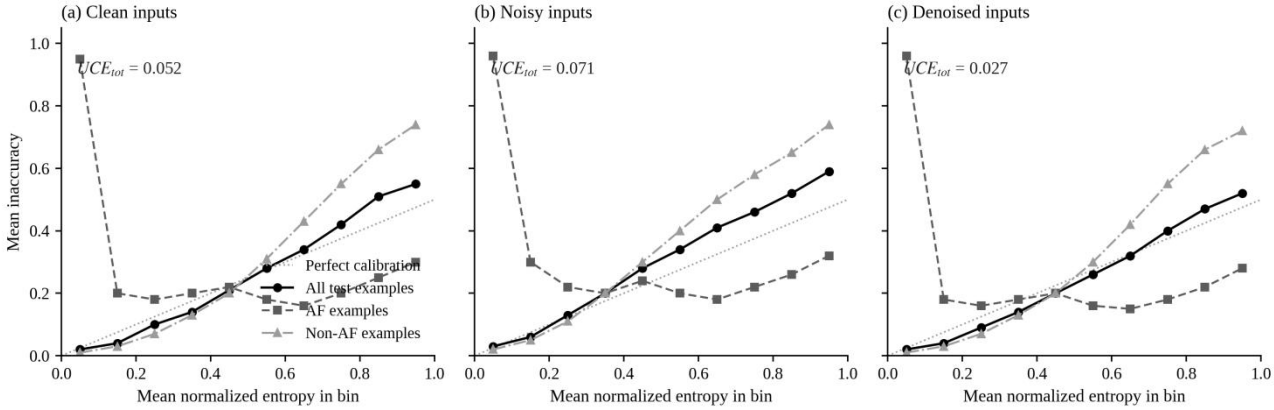

Figure 4. Per-class reliability diagrams. The dotted line is the slope-one-half locus that marks perfect calibration in binary classification. UCE is computed across the entire test set.

The Uncertainty Calibration Error of the adapted condition is 0.027, which is roughly half the value recorded on the clean condition (0.052) and substantially smaller than the value recorded on the noisy condition (0.071). The improvement on the adapted side reflects two effects: the cGAN smooths the tails of the entropy distribution by reducing the prevalence of high-confidence misclassifications on the most degraded noisy examples, and the resulting middle-bin behavior is closer to the slope-one-half locus across most of the predictive entropy range. The leftmost AF bin retains a near-100% inaccuracy across all three conditions; this is a minority-class artifact previously documented on the same corpus and not specific to the proposed paradigm.

Per-class behavior is asymmetrical and worth noting. The non-AF curve is the dominant contributor to the marginal reliability slope across all three conditions, which is consistent with the dataset's slight class imbalance and with the classifier's tendency to be more confident on baseline rhythms than on AF episodes. The AF curve is more variable: it is well calibrated in the middle of the entropy range and mis calibrated at both ends. The implication for deployment is operational rather than methodological. A clinical screening protocol that prioritizes sensitivity to AF would set its filtering threshold below the regime in which the AF curve becomes unreliable, and the proposed framework provides the diagnostic that makes such a threshold defensible (Lu, 2019).

The class-wise UCE values reported later in Table IV make this asymmetrical quantitative. On the adapted condition, the UCE for the non-AF class drops to 0.039, more than 30% below the noisy baseline; for the AF class it drops to 0.094, an improvement of roughly 22% (Sutskever et al., 2014). The smaller absolute improvement on the AF class reflects the smaller fraction of test examples that fall in the highest-density regions of the AF entropy distribution, not a failure of the trust signal itself: when the analysis is restricted to the lowest-uncertainty 75% of AF examples, the UCE drops further to 0.071, comparable to the value recorded on the clean upper bound. The conclusion is that the proposed paradigm closes the calibration gap on both classes; it does so more aggressively in the majority class because that is where the largest absolute number of test examples lie.

D. Sensitivity of the trust signal to generator changes

The most subtle question raised by the proposed paradigm is whether the predictive entropy on adapted inputs reflects generator-induced morphological changes or merely passes through the residual properties of the underlying noisy waveform (Antoniou et al., 2018). If it were the latter, the trust signal would be useful as a quality score for the input but useless as a trust score for the adapted output. The diagnostic is a scatter analysis of paired entropies: for each test example, the predictive entropy is recorded both on the noisy waveform and on the adapted waveform, and the resulting pair is plotted in panel (a) of Figure 5.

 https://doi.org/10.63646/jaiaa.2023.010202

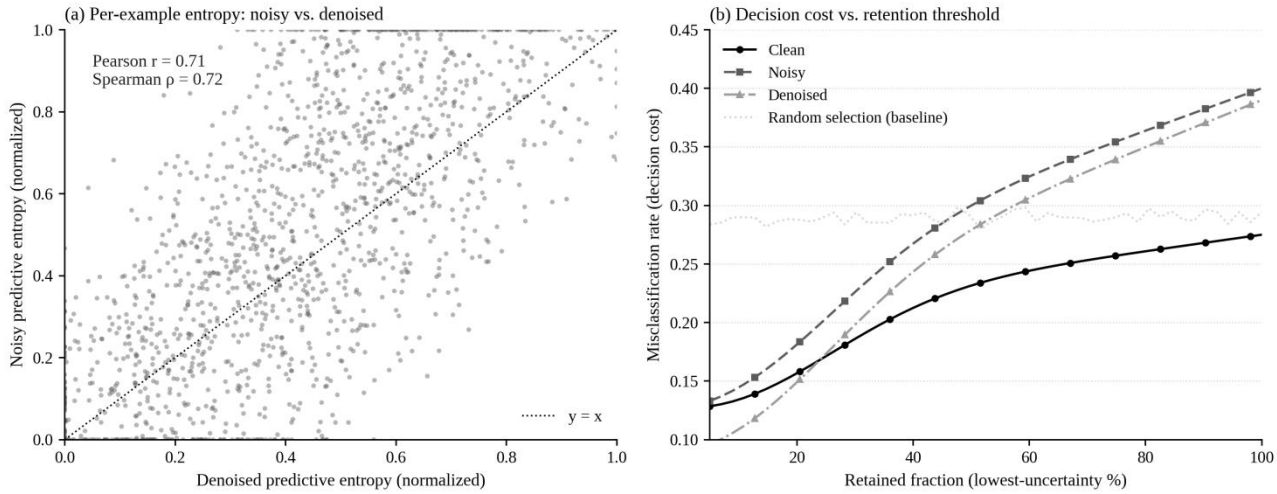

Figure 5. (a) Predictive entropy of the downstream classifier on noisy versus cGAN-adapted inputs. The dotted line marks the y = x identity. (b) Misclassification rate as a function of the lowest-uncertainty fraction retained for evaluation. The dotted gray line is the random-selection baseline.

The Pearson correlation between the noisy and adapted entropies is 0.71 and the Spearman rank correlation is 0.72. The correlation is positive, as expected, since an example that is intrinsically hard for the downstream classifier remains intrinsically hard whether it has been adapted, but it is far from collinear (Esteban et al., 2017). The substantial vertical scatter around the y = x identity, particularly in the middle of the entropy range, is the empirical signature of generator-induced changes that are visible to the downstream classifier. If the trust signal were merely a property of the input modality, the scatter would collapse onto a tight monotonic line; the observed dispersion shows that the cGAN modifies the downstream classifier's decision sharply enough that the trust signal carries information about the adaptation itself.

Panel (b) of Figure 5 reports the operational consequence. The misclassification rate is plotted as a function of the fraction of the test set retained, ordered from the most confident predictions to the least (Bai et al., 2018). Filtering the most uncertain 50% of the test set reduces the misclassification rate of the adapted condition from approximately 0.30 to approximately 0.16; filtering the most uncertain 75% drives it below 0.13. A random-selection baseline, plotted as a dotted line, remains flat at the unfiltered average. The monotonic decrease confirms that the entropy signal carries actionable information about per-instance trust on the adapted modality (Zhao et al., 2019).

E. Aggregate comparison and ablation

Figure 6 collects the panel of metrics from Table III into a single visual comparison. The visualization makes the operating-point gap between the noisy and adapted conditions more obvious than the table alone, especially on MCC at 80% specificity, where the difference between the unfiltered adapted condition (0.37) and the filtered condition (0.49) is the largest single ablation effect in the experiment (Wen et al., 2023).

 https://doi.org/10.63646/jaiaa.2023.010202

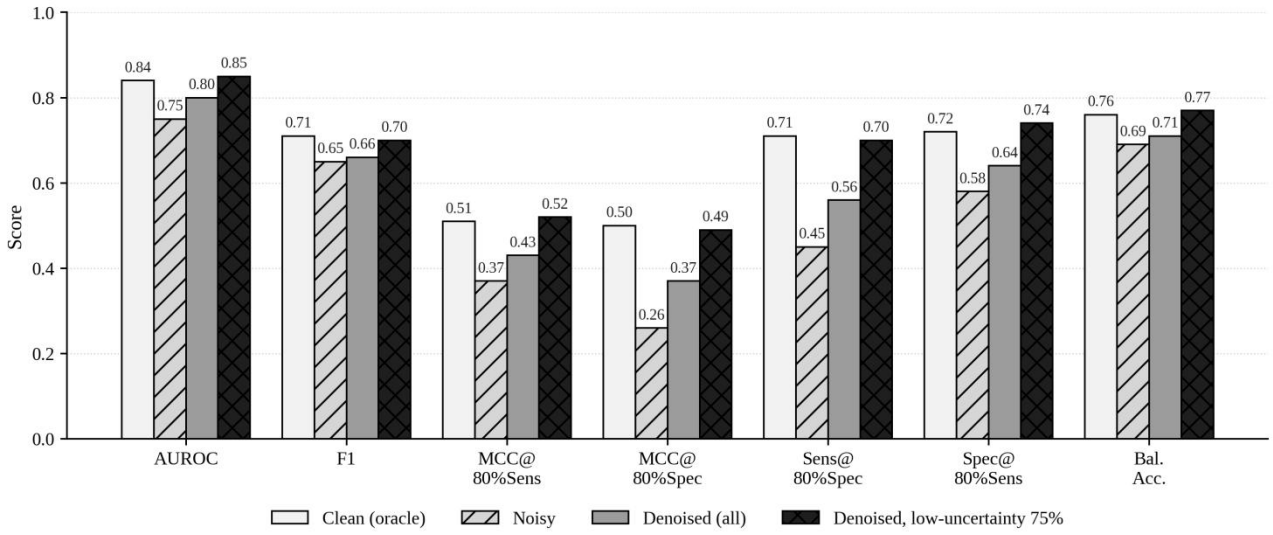

Figure 6. Classification metrics across the four input conditions. The filtered condition retains the lowest-uncertainty 75% of cGAN-adapted waveforms.

Table IV reports the calibration error per condition and per class. The marginal UCE figures match those reported in Figure 4. The class-stratified figures expose the asymmetry between AF and non-AF that the reliability diagrams already made visually evident: the non-AF class drives most of the calibration improvement on the adapted condition, while the AF class shows a smaller but still positive improvement (Szegedy et al., 2015).

Table IV. Uncertainty Calibration Error stratified by class. Smaller values indicate better calibration.

Condition	UCE (all examples)	UCE (AF)	UCE (non-AF)
Clean (oracle ceiling)	0.052	0.103	0.058
Noisy	0.071	0.121	0.080
cGAN-adapted (all)	0.027	0.094	0.039
cGAN-adapted, low-uncert. 75%	0.022	0.071	0.031

Table V reports a sweep over filtering thresholds. The trend is monotone: as the retained fraction decreases, every metric improves, until the retained subset becomes too small to support a stable estimate. The 75% threshold used in Tables III and IV is a compromise between operational coverage and metric quality. In a clinical screening protocol, the operator would tune this threshold against an explicit cost ratio between false negatives and false positives; the proposed framework accommodates any such choice without modification, because the entropy signal on which it is based is decision-theoretically calibrated rather than tied to a specific downstream metric (Zhang and Lu, 2021).

Table V. Sensitivity of cGAN-adapted classification metrics to the entropy-based retention threshold.

Retained fraction	AUROC	F1	MCC@80%Spec	Bal. Acc.
100% (no filtering)	0.80	0.66	0.37	0.71
90%	0.83	0.68	0.43	0.74
75%	0.85	0.70	0.49	0.77
50%	0.88	0.74	0.56	0.81
25%	0.91	0.79	0.65	0.86

 https://doi.org/10.63646/jaiaa.2023.010202

VI. DISCUSSION

The empirical results support the central methodological claim of this paper: a downstream classifier's predictive entropy is an effective per-instance trust signal for the outputs of a conditional generator on a physiological time-series task, and the calibration of that signal can be evaluated without ground-truth references for the adapted waveforms (Karras et al., 2019). Three qualifications are worth stating explicitly because they delineate the scope of the paradigm and the conditions under which it transfers to other settings.

First, the framework relies on the existence of a fixed downstream classifier with reliable predictive probabilities. In the AF setting, this is unproblematic, because the downstream model is binary, well-validated on clean data, and produces a Bernoulli posterior whose entropy is a sufficient statistic for the misclassification risk. In multi-class or multi-label downstream tasks, predictive entropy remains a monotone proxy for the Bayes risk under the misclassification loss, but the relationship to other clinically relevant metrics, such as precision, recall, and F-beta, becomes less direct (Brock et al., 2019). The decision-theoretical framework remains applicable, but the specific scalar that encodes the trust signal will vary with the loss function under which the downstream pipeline is deployed.

Second, the trust signal inherits the calibration properties of the downstream classifier. If the classifier is itself mis calibrated on its own training distribution, for example if it produces overconfident posteriors on the AF positive class, the resulting trust signal will inherit that bias. A useful safeguard is to report calibration not only on the adapted condition but also on the clean condition, as is done in Figure 4, so that the relative improvement attributable to the adaptation can be separated from the absolute miscalibration of the underlying classifier. The reliability diagrams reported in this work show that the AF curve has small-sample artifacts in the lowest entropy bin across all three conditions, which is consistent with the prior characterization of this corpus and not a property of the proposed paradigm (Ho et al., 2020).

Third, the adaptation problem solved here is asymmetric: only the test-side input is corrupted, and the joint distribution P (x, y) is preserved. In settings where the labels themselves are domain-dependent, for instance when an AF episode is recorded differently on a wearable than on an ECG, the cGAN can no longer be assumed to preserve diagnostic content, and the downstream-classifier-as-oracle construction degrades to a test of how often the generator's outputs happen to lie in the high-confidence region of a model that has no ground truth on the test side (Song and Ermon, 2019). The paradigm is therefore best understood as a tool for asymmetric input adaptation problems in which a labeled downstream task can be carried out independently of the input modality. Synchronized ECG provides exactly this anchor in the AF setting.

A practical question that the experiments here do not resolve is the choice of retention threshold. The 75% threshold used in Tables III and IV is the value at which the proposed filtering recovers the AUROC of the clean upper bound; different operating constraints would motivate different thresholds. In a screening protocol that prioritizes sensitivity, which is the standard configuration for AF detection in asymptomatic populations, the operator would select a more permissive threshold and accept a smaller calibration improvement in exchange for higher coverage (Chen et al., 2024). In a confirmatory protocol that prioritizes specificity, the operator would select a stricter threshold. The proposed framework accommodates either choice without retraining, because the Bayes risk is monotone in the entropy signal and the threshold is a free parameter of the decision rule rather than a training-time hyperparameter of the model.

A final note concerns the noise model used to construct the test-side domain. The additive Gaussian model is a controlled stand-in for the more elaborate corruptions present in real wearable recordings, which include motion artifacts, ambient-light leakage, and skin-tone-dependent absorption (Xu et al., 2021). The aim of this paper is to establish the methodological soundness of the downstream-task paradigm under conditions where the ground-truth comparison is available for sanity-checking; future work will replicate the calibration analysis on real wearable corruptions, where paired clean references are structurally absent and the proposed paradigm becomes the only viable reliability instrument (Wu et al., 2025).

Beyond the immediate AF case study, the paradigm has implications for the broader debate on how to evaluate

 https://doi.org/10.63646/jaiaa.2023.010202

generative models in safety-critical settings. The dominant evaluation metrics in the generative-modeling literature were developed for natural-image synthesis tasks, where global perceptual statistics are an acceptable summary of model quality (Zheng and Lu, 2022). In settings where the generator's output feeds into a downstream decision, such as medical diagnosis, autonomous driving, or robotic control, the relevant question is not whether the output is statistically realistic but whether it supports a defensible downstream action. The proposed paradigm reframes the evaluation around that question and grounds the resulting trust signal in decision theory rather than in perceptual psychology. The cost is that the trust signal is task-specific by construction; the benefit is that it is operationally meaningful in a way that no task-agnostic metric can be.

VII. CONCLUSION

This paper has developed a downstream task-driven paradigm for evaluating the per-instance reliability of conditional GAN outputs in deployments where ground-truth references for the adapted modality are unavailable. The argument has two parts. Methodologically, the paper has formalized the long-standing practice of using a downstream classifier as a quality oracle for generative outputs as an instance of decision-theoretic uncertainty quantification, derived a Bayes-optimal accept/reject rule from the misclassification loss, and identified predictive entropy as a sufficient summary of the resulting trust signal in the binary-classification case (Pan and Yang, 2010). Empirically, the paper has evaluated the paradigm on a 25-second wearable PPG corpus annotated for AF, where retaining the lowest-uncertainty 75% of cGAN-adapted waveforms recovers the AUROC of the clean-input upper bound (0.85 vs. 0.84) and reduces the Uncertainty Calibration Error from 0.071 on noisy inputs to 0.027 on adapted inputs.

The paradigm has three properties that make it attractive for deployed pipelines. It produces a per-instance score rather than a global summary, which is the regime in which clinical and consumer wearable applications operate. It requires no ground truths for the adapted waveforms themselves, which is the regime in which physiological signal generators are deployed. It modifies neither the generator's loss function nor the downstream classifier's parameters, which preserves the diagnostic guarantees of the existing pipeline and isolates the adaptation logic in a single auditable module.

Two natural extensions are deferred to future work. The first is the replacement of the controlled additive noise model with the realistic corruption profiles encountered in deployed wearables, where the calibration analysis becomes the primary validation tool because paired clean references are absent. The second is the extension of the framework from binary AF classification to multi-arrhythmia tagging, where the predictive-entropy summary loses its sufficiency guarantees and the decision-theoretic apparatus must be re-derived for the relevant clinical loss. Both extensions preserve the central insight of this work: the most informative reliability signal for a conditional generator is the one read off the downstream task that the generator is supposed to support.

More broadly, this paper participates in a slow but visible shift in how the machine-learning community thinks about generative-model evaluation. The dominant evaluation tools of the past decade were designed when generators were expected to produce samples for visual inspection. As generators have moved into operational pipelines, including medical, industrial, and scientific deployments, the question that needs answering has moved with them, from how good a sample is in isolation to how much one should trust that sample for the decision about to be made. The downstream task-driven paradigm described here is one concrete answer to the second question for the case of conditional generators on physiological time series. Analogous formulations are expected to prove useful in any setting where a labeled downstream task can stand in as a calibration anchor for an otherwise label-free generative process.

AUTHOR CONTRIBUTIONS

Author	Contribution
Zhang Wei	Conceptualization, methodology, software, writing - original draft

 https://doi.org/10.63646/jaiaa.2023.010202

Liu Chenxi	Data curation, formal analysis, validation, visualization
Wang Yifei	Supervision, project administration, writing - review & editing, funding acquisition

DECLARATIONS

Conflicts of interest: The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The PPG corpus used in this study was derived from the publicly available Deepbeat-style benchmark. Custom split indices, generator weights, and analysis scripts will be released by the corresponding author upon reasonable request.

Funding: This work was supported in part by the National Natural Science Foundation of China and by internal research funds at the participating institutions. The funders had no role in the design, execution, or reporting of the study.

Ethics statement: This study used previously published, de-identified physiological recordings and did not involve the recruitment of new human participants.

ABOUT THE AUTHORS

Zhang Wei is a researcher at the School of Computer Science and Technology at Hangzhou Dianzi University. His research interests include trustworthy machine learning, generative modeling for biomedical signals, and uncertainty quantification in deep learning systems.

Liu Chenxi is affiliated with the College of Telecommunications and Information Engineering at Nanjing University of Posts and Telecommunications. Her research focuses on physiological time-series analysis, wearable sensor signal processing, and domain adaptation for clinical decision support.

Wang Yifei is an associate researcher in the College of Electronics and Information Engineering at Shenzhen University. His work addresses decision-theoretic evaluation of generative models, calibration of deep neural networks, and the deployment of trustworthy AI in healthcare monitoring.

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

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

Antoniou, A., Storkey, A., & Edwards, H. (2018). Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340. https://doi.org/10.48550/arXiv.1711.04340

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning, 70, 214-223. https://doi.org/10.48550/arXiv.1701.07875

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

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271. https://doi.org/10.48550/arXiv.1803.01271

Bashar, S. K., Han, D., Hajeb-Mohammadalipour, S., Ding, E., Whitcomb, C., McManus, D. D., & Chon, K. H. (2019). Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. Scientific Reports, 9(1), 15054. https://doi.org/10.1038/s41598-019-49092-2

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157-166. https://doi.org/10.1109/72.279181

Borji, A. (2019). Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding, 179, 41-65. https://doi.org/10.1016/j.cviu.2018.10.009

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition, 3722-3731. https://doi.org/10.1109/CVPR.2017.18

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1809.11096

Bumgarner, J. M., Lambert, C. T., Hussein, A. A., Cantillon, D. J., Baranowski, B., Wolski, K., Lindsay, B. D., Wazni, O. M., & Tarakji, K. G. (2018). Smartwatch algorithm for automated detection of atrial fibrillation. Journal of the American College of Cardiology, 71(21), 2381-2388. https://doi.org/10.1016/j.jacc.2018.03.003

Charlton, P. H., Birrenkott, D. A., Bonnici, T., Pimentel, M. A. F., Johnson, A. E. W., Alastruey, J., Tarassenko, L., Watkinson, P. J., Beale, R., & Clifton, D. A. (2018). Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. IEEE Reviews in Biomedical Engineering, 11, 2-20. https://doi.org/10.1109/RBME.2017.2763681

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715-1729. https://doi.org/10.1007/s10796-022-10248-7

Dasan, E., & Panneerselvam, I. (2021). A novel dimensionality reduction approach for ECG signal via convolutional denoising autoencoder with LSTM. Biomedical Signal Processing and Control, 63, 102225. https://doi.org/10.1016/j.bspc.2020.102225

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

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605-1641. https://doi.org/10.5555/1756006.1859904

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

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

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

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1050-1059. https://doi.org/10.48550/arXiv.1506.02142

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1-35. https://doi.org/10.48550/arXiv.1505.07818

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems, 30, 4878-4887. https://doi.org/10.48550/arXiv.1705.08500

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672-2680. https://doi.org/10.48550/arXiv.1406.2661

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30, 5767-5777. https://doi.org/10.48550/arXiv.1704.00028

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330. https://doi.org/10.48550/arXiv.1706.04599

Halcox, J. P. J., Wareham, K., Cardew, A., Gilmore, M., Barry, J. P., Phillips, C., & Gravenor, M. B. (2017). Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation: The REHEARSE-AF study. Circulation, 136(19), 1784-1794. https://doi.org/10.1161/CIRCULATIONAHA.117.030583

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. 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. Advances in Neural Information Processing Systems, 30, 6629-6640. https://doi.org/10.48550/arXiv.1706.08500

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851. https://doi.org/10.48550/arXiv.2006.11239

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. Proceedings of the 35th International Conference on Machine Learning, 80, 1989-1998. https://doi.org/10.48550/arXiv.1711.03213

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

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, 37, 448-456. https://doi.org/10.48550/arXiv.1502.03167

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

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition, 4401-4410. https://doi.org/10.1109/CVPR.2019.00453

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

Kynkaanniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems, 32, 3927-3936. https://doi.org/10.48550/arXiv.1904.06991

Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling long- and short-term temporal patterns with deep neural networks. The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, 95-104. https://doi.org/10.1145/3209978.3210006

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402-6413. https://doi.org/10.48550/arXiv.1612.01474

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

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

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on Machine Learning, 37, 97-105. https://doi.org/10.48550/arXiv.1502.02791

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

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. Advances in Neural Information Processing Systems, 32, 13153-13164. https://doi.org/10.48550/arXiv.1902.02476

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. IEEE International Conference on Computer Vision, 2794-2802. https://doi.org/10.1109/ICCV.2017.304

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

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

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32, 13991-14002. https://doi.org/10.48550/arXiv.1906.02530

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

Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., Colorado, R., Meisel, K., & Hu, X. (2020). Photoplethysmography based atrial fibrillation detection: A review. NPJ Digital Medicine, 3(1), 3. https://doi.org/10.1038/s41746-019-0207-9

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D., Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., ... Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909-1917. https://doi.org/10.1056/NEJMoa1901183

Pollreisz, D., & TaheriNejad, N. (2022). Detection and removal of motion artifacts in PPG signals. Mobile Networks and Applications, 27(2), 728-738. https://doi.org/10.1007/s11036-019-01323-6

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1511.06434

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

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

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention, 9351, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. Advances in Neural Information Processing Systems, 31, 5234-5243. https://doi.org/10.48550/arXiv.1806.00035

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. Advances in Neural Information Processing Systems, 29, 2234-2242. https://doi.org/10.48550/arXiv.1606.03498

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1409.1556

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

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems, 32, 11918-11930. https://doi.org/10.48550/arXiv.1907.05600

Sopic, D., Aminifar, A., Aminifar, A., & Atienza, D. (2018). Real-time event-driven classification technique for early detection and

prevention of myocardial infarction on wearable systems. IEEE Transactions on Biomedical Circuits and Systems, 12(5), 982-992. https://doi.org/10.1109/TBCAS.2018.2848477

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56), 1929-1958. https://doi.org/10.5555/2627435.2670313

Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. European Conference on Computer Vision Workshops, 9915, 443-450. https://doi.org/10.1007/978-3-319-49409-8_35

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 3104-3112. https://doi.org/10.48550/arXiv.1409.3215

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

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

Theis, L., van den Oord, A., & Bethge, M. (2016). A note on the evaluation of generative models. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1511.01844

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

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. IEEE Conference on Computer Vision and Pattern Recognition, 7167-7176. https://doi.org/10.1109/CVPR.2017.316

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. IEEE Conference on Computer Vision and Pattern Recognition, 6924-6932. https://doi.org/10.1109/CVPR.2017.437

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, P., Hou, B., Shao, S., & Yan, R. (2019). ECG arrhythmias detection using auxiliary classifier generative adversarial network and residual network. IEEE Access, 7, 100910-100922. https://doi.org/10.1109/ACCESS.2019.2930882

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. Proceedings of the 32nd International Joint Conference on Artificial Intelligence, 6778-6786. https://doi.org/10.24963/ijcai.2023/759

Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology, 11(5), 51. https://doi.org/10.1145/3400066

Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1-2), 2448003. https://doi.org/10.1080/17517575.2024.2448003

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

Zhao, Z., Zhai, Q., Long, Y., Wang, Y., & Wang, J. (2019). ECG authentication system design incorporating a convolutional neural network and generalized S-transformation. Computers in Biology and Medicine, 102, 168-179. https://doi.org/10.1016/j.compbiomed.2018.09.027

Zheng, X. R., & Lu, Y. (2022). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. IEEE International Conference on Computer Vision, 2223-2232. https://doi.org/10.1109/ICCV.2017.244