Vol. 2, No. 4, pp.: 18–32, 2024

Detecting Hallucinations in 1D Generative Models: A Decision-Theoretic Approach to Quality Control for Synthetic Biosignals

Adrian J. Whitfield[1,*], Helena S. Karlsson[2], Mateus R. Oliveira[3]

[1] School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth PL4 8AA, United Kingdom

[2] Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden

[3] Department of Electrical and Electronic Engineering, Federal University of Santa Catarina, Florianópolis SC 88040-900, Brazil

** adrian.whitfield@plymouth.ac.uk*

Article Information	
Received	20 July 2024
Accepted	28 November 2024
DOI	https://doi.org/10.63646/datamind.2024.020403

Abstract

Generative deep-learning models are increasingly used to denoise, harmonise, and adapt one-dimensional biosignals before they are passed to clinical classification pipelines. While these models can substantially narrow the gap between the data domain seen at training and the data observed in deployment, they are also prone to producing visually plausible outputs that nonetheless contain spurious features—a phenomenon we describe as one-dimensional hallucination. Standard quality indicators, such as reconstruction error, Fréchet-style distances, or perceptual similarity, are either unavailable in unpaired settings or insensitive to small but clinically meaningful artefacts. This paper proposes a decision-theoretic framework that evaluates synthetic biosignals through the eyes of the downstream task they are intended to support. The predictive entropy of a frozen, well-calibrated atrial-fibrillation classifier acts as a proxy for the Bayes-optimal misclassification risk of the adapted example, allowing per-instance trustworthiness scoring without requiring ground-truth references for the generated signals themselves. Using a 1D Pix2pix denoiser trained on a heavily augmented variant of a public photoplethysmography dataset, we show that classifier entropy is a reliable selector of high-utility outputs, that a 75 % retention threshold recovers the performance of the unaugmented baseline, and that calibration improves measurably (UCEtotal = 0.087 → 0.038) after generative adaptation. The approach formalises a heuristic that is widely used in practice—judging synthetic data by what a downstream model does with it—and turns it into a quantitative quality-control instrument suitable for wearable-health pipelines and similar safety-aware applications.

Keywords: ***photoplethysmography; generative adversarial networks; uncertainty quantification; decision theory; calibration; atrial fibrillation; domain adaptation***

1. Introduction

Deep generative models have become the de-facto tool for transforming raw, real-world signals into cleaner, model-ready inputs. Conditional generative adversarial networks (Goodfellow et al., 2020; Isola et al., 2017), denoising autoencoders, and diffusion-style architectures are now routinely used to remove noise, fill gaps, and align measurements with the distribution expected by a downstream discriminative model (Yi et al., 2019; Wang and Deng, 2018). In wearable-health pipelines, where photoplethysmography (PPG) is the dominant modality for continuous cardiac monitoring (Allen, 2007; Castaneda et al., 2018), this kind of pre-processing is increasingly performed by neural denoisers rather than by classical filtering chains.

The attraction is obvious. A trained generative model can absorb the statistics of a noisy or out-of-distribution input and produce something that, by every visual measure, looks like an in-distribution example. The risk is more subtle. Generative models do not reconstruct; they extrapolate. When the input deviates from the training manifold, the model fills the gap with whatever local pattern it has learned to associate with similar contexts. The output may still look clean, periodic, even physiologically plausible, yet contain features that the underlying measurement never produced. We refer to this as one-dimensional hallucination, by analogy with the better-studied phenomena in image synthesis (Yi et al., 2019) and natural language generation (Ji et al., 2023).

Hallucinated biosignals are not a cosmetic problem. A pretrained atrial-fibrillation (AF) classifier looking at a hallucinated waveform sees an input it cannot interrogate, and a clinician downstream sees a confident prediction with no flag attached. The conventional safeguards developed for natural-image generators transfer poorly to this setting. Reconstruction-based metrics such as the mean squared error or structural similarity index require ground-truth references that, in unpaired domain adaptation, do not exist. Distribution-level scores such as the Fréchet Inception Distance assume a feature extractor calibrated on natural images and cannot flag individual outputs. No-reference scores such as BRISQUE were designed for image statistics and generalise poorly to short physiological time series (Pereira et al., 2020). What practitioners actually need is a per-instance signal that says: this generated example is unlikely to produce a useful prediction—discard it, or at least flag it.

Uncertainty quantification (UQ) supplies that signal in principle (Abdar et al., 2021; Gawlikowski et al., 2023). In practice, however, the standard UQ machinery is built around a supervised classifier or regressor, not a generator (Hüllermeier and Waegeman, 2021), and standard reliability diagnostics rely on comparing predicted uncertainty against a measurable error. For a generative model with no ground truth, even the question of whether an uncertainty estimate is well-calibrated becomes ambiguous. This paper argues that the resolution lies in shifting from an estimator-centric notion of uncertainty to a decision-theoretic one. Instead of asking how uncertain the generator is about its own output, we ask how uncertain the downstream task is about the prediction it would make if that output were used. The answer to the second question can be measured even when the answer to the first cannot.

We instantiate this idea on a tightly defined case study. A one-dimensional Pix2pix-style conditional GAN (Isola et al., 2017) is trained to denoise heavily augmented wearable PPG. A frozen 1D AlexNet variant, trained on the unaugmented PPG distribution, plays the role of the downstream AF classifier. The predictive entropy of that classifier on a generated example is treated as the Bayes-optimal subjective expected loss under a misclassification cost (Hannun et al., 2019). We then audit how reliably this entropy ranks examples in terms of the actual decision quality that follows. The contribution is threefold. First, we formalise the long-standing

heuristic of using a downstream classifier as a generative quality indicator and show that the formalisation is a special case of the decision-theoretic UQ framework. Second, we report a calibration analysis on synthetic outputs that does not require paired ground truths, made possible by anchoring the analysis to the classifier rather than to the generator. Third, we show that selecting on predictive entropy delivers concrete operational gains in this setting—closing most of the gap between noisy and unaugmented inputs while flagging the residual hallucinations.

2. Background and Related Work

2.1 Generative modelling of biosignals

Generative adversarial networks have a long track record in biomedical signal processing (Yi et al., 2019; Creswell et al., 2018). For one-dimensional inputs, the conditional Pix2pix architecture (Isola et al., 2017) and its cycle-consistent counterpart (Zhu et al., 2017) have been adapted to denoise electrocardiograms, electroencephalograms, and PPG (Reiss et al., 2019; Tang et al., 2020). Most reported evaluations rely on supervised reconstruction error (Elgendi, 2012) or distribution-level distances. These metrics, however, are silent on the question that practitioners actually face: which individual generated examples are trustworthy enough to keep?

Recent reviews of the wearable-PPG roadmap make this gap explicit. Charlton et al. (2023) note that signal quality assessment in deployment remains one of the largest barriers to clinical adoption, and Pereira et al. (2020) point out that motion artefacts and hardware variability can degrade AF detection by margins much larger than the differences between competing classifiers. The implication is that quality control of the input pipeline matters as much as model accuracy itself, and that whatever metric is used must scale to millions of short windows produced by always-on sensors (Biswas et al., 2019).

2.2 Domain adaptation as a deployment-time problem

When a classifier trained on clean data is exposed to a noisier domain, performance degrades in proportion to the divergence between the two marginal distributions (Wang and Deng, 2018). Adversarial domain adaptation methods (Tzeng et al., 2017) attempt to align the source and target feature spaces; in our setting, where the classifier is fixed and only the inputs can be modified, the natural counterpart is asymmetric input adaptation. The generator transports each test example back toward the training distribution. The price of this convenience is the very risk we are trying to control: every adaptation step is also an opportunity to invent features that did not originate in the measurement (Zech et al., 2018; Lu, 2019).

2.3 Uncertainty, calibration, and decision theory

Uncertainty quantification in deep learning has matured rapidly (Abdar et al., 2021; Gawlikowski et al., 2023). The conventional decomposition into aleatoric and epistemic components (Hüllermeier and Waegeman, 2021) gives a clean mathematical account of where uncertainty comes from, but offers little guidance about what an uncertainty estimate ought to do once produced. Calibration metrics such as the Expected Calibration Error and the Uncertainty Calibration Error (Naeini et al., 2015) compare a model's stated confidence with its empirical accuracy, but they presume the existence of a ground truth—an awkward fit for generative outputs.

Decision theory provides a different anchor. In the Bayes risk formulation, the value of a prediction is the expected loss it incurs when used to choose among a defined set of actions. For misclassification loss in a binary task, this risk reduces to one minus the maximum class probability, and Shannon entropy becomes a monotone

surrogate (Hüllermeier and Waegeman, 2021; Singh et al., 2020). The framework can be instantiated whenever the downstream consequences of a prediction can be expressed as a loss function. Linking the trustworthiness of a generated input to the entropy of a downstream classifier is therefore not a heuristic shortcut—it is the natural Bayes risk under the only loss the application actually cares about.

3. A Decision-Theoretic Quality Framework

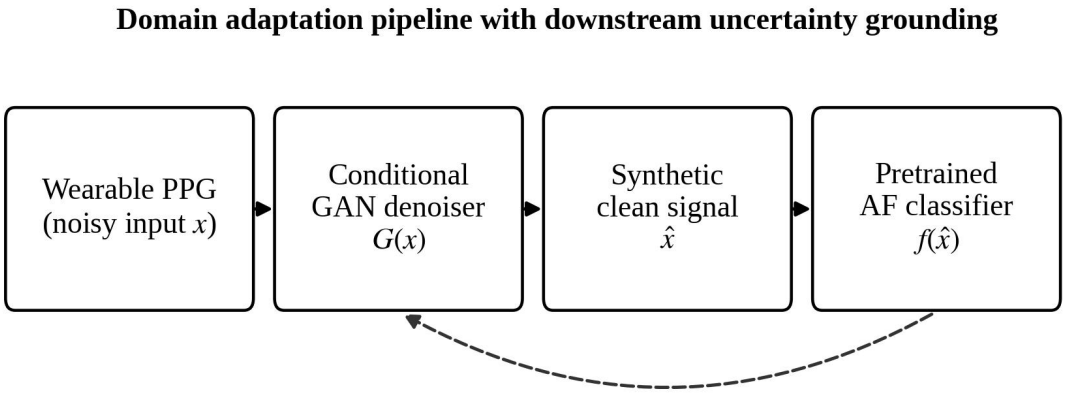

Figure 1. *Conceptual pipeline. Wearable PPG inputs are transformed by a conditional GAN denoiser; the synthetic clean signal is consumed by a frozen AF classifier whose predictive entropy returns to the loop as the per-instance trustworthiness signal.*

Figure 1 sketches the architecture of the proposed framework. Three components are involved: a generative domain adapter that produces synthetic clean signals from noisy wearable PPG, a frozen downstream classifier trained on the clean distribution, and a decision-theoretic quality module that scores each generated example by the predictive entropy of the classifier. The architecture is deliberately modular—each block can be replaced without reworking the others—because the value of the framework lies in the feedback path rather than in any one component. We now formalise this feedback path.

3.1 Problem formulation

Let $x \in \mathbb{R}^N$ denote a noisy PPG window and let $G(\cdot)$ be a generator that returns a synthetic clean signal $\hat{x} = G(x)$. A frozen classifier f produces a probability vector $p^f(y \mid \hat{x})$ over the binary label $y \in \{\text{AF, non-AF}\}$. Because the classifier was trained on the clean distribution, its behaviour on $\hat{x}$ is a clean proxy of the prediction it would make on a true clean signal. The overall task is not to reconstruct $\hat{x}$ accurately—we have no ground truth for $\hat{x}$—but to ensure that $f(\hat{x})$ is a useful prediction in the decision sense (Singh et al., 2020).

3.2 Bayes risk under misclassification loss

Following the standard decision-theoretic recipe, we attach to each candidate action $a \in \{\text{AF, non-AF}\}$ a loss $L(a, y) = \{a \neq y\}$. The posterior expected loss under the classifier's belief becomes $\rho(a \mid \hat{x}) = \Sigma_y L(a, y)\, p^f(y \mid \hat{x}) = 1 - p^f(a \mid \hat{x})$. The Bayes-optimal action minimises this risk; in the binary case the minimum risk equals $1 - \max_y p^f(y \mid \hat{x})$, which is monotone in the predictive entropy $H[p^f(\cdot \mid \hat{x})] = -\Sigma_y p^f(y \mid \hat{x}) \log p^f(y \mid \hat{x})$. For

the purpose of ranking and filtering, entropy and Bayes risk are interchangeable, and entropy is the more convenient one because it is bounded, differentiable, and natively reported by softmax outputs (Hannun et al., 2019).

3.3 Reliability without paired ground truths

The framework deliberately sidesteps the question of whether $\hat{x}$ is close to its hidden ground truth. Reliability is instead audited at the classifier interface. We discretise the entropy axis into ten equal-width bins, compute within-bin misclassification rates, and report the Uncertainty Calibration Error UCE = Σ_m $(|B_m|/N) \cdot |\text{err}(B_m) - ½\ \text{uncert}(B_m)|$, where $\text{err}(B_m)$ is the empirical error rate in bin m, $\text{uncert}(B_m)$ is the mean normalised entropy in that bin, and the factor of ½ reflects the binary-classification slope of an ideally calibrated reliability curve (Naeini et al., 2015). All three quantities are observable. None of them depend on a reconstructed-versus-original comparison of $\hat{x}$.

Two practical implications follow. First, the framework can be applied to any combination of generator and downstream classifier as long as both are differentiable in a forward pass. Second, the cost of auditing reliability is dominated by the classifier's inference budget, which is typically a small fraction of the generator's. This makes continuous quality monitoring feasible at the rate at which wearable sensors produce data (Biswas et al., 2019).

4. Experimental Setup

4.1 Dataset and augmentation

We use a public PPG dataset whose patient-disjoint training, validation and test splits contain 106,249, 15,256 and 15,377 windows respectively, each lasting 25 seconds at 32 Hz, with binary AF labels derived from contemporaneous ECG (Pereira et al., 2020; Hannun et al., 2019). Inputs are subjected to controlled distribution shift: each test window is corrupted with additive Gaussian noise of standard deviation 0.1 on the [0, 1] amplitude scale, with values clamped to [0, 2]. The augmentation deliberately exceeds the noise levels typically tolerated by wearable algorithms so that the gap between training and deployment is large enough to test the method (Reiss et al., 2019).

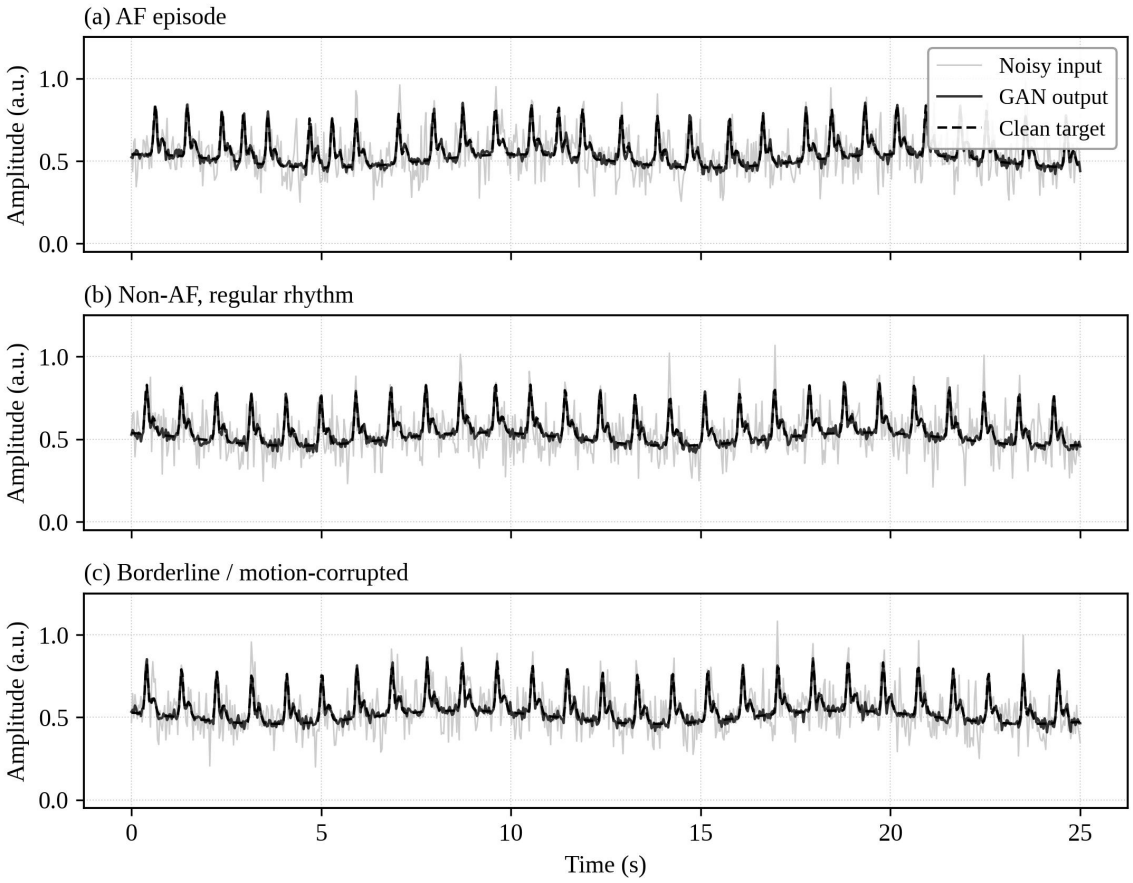

Figure 2. *Three illustrative PPG windows. Light grey traces show the noisy input, solid black traces show the GAN-denoised output, and the dashed black trace shows the unaugmented clean target retained for visualisation only. Panels (a) and (b) are typical AF and non-AF cases; panel (c) is a borderline case dominated by motion-like noise.*

Figure 2 illustrates the qualitative effect of the augmentation and of the subsequent generative denoising on three representative windows. The recovery is clean enough to be visually convincing, which is precisely the regime in which hallucinations are most insidious (Tang et al., 2020). The visual agreement with the dashed reference does not, by itself, provide any guarantee that the AF morphology has been preserved: that question is what the rest of the paper sets out to answer.

4.2 Generator and discriminator

The generator is a one-dimensional U-Net (Ronneberger et al., 2015) with seven encoder and seven decoder blocks. Each encoder block consists of a Conv1D layer with kernel size 4, stride 2 and padding 1, followed by leaky ReLU (slope 0.2) and instance normalisation. Decoder blocks mirror the encoder using transposed convolutions and ReLU activations, with skip connections at matching scales. The discriminator follows the PatchGAN convention (Isola et al., 2017), producing a 1-channel logit map along time. The combined loss is $L_G = L_{GAN} + \lambda \cdot L_1$ with $\lambda = 100$; the Adam optimiser uses learning rates 2×10^{-4} and 1×10^{-5} for generator and discriminator respectively. Early stopping is based on validation L_1 loss with a patience of three epochs.

4.3 Classifier and metrics

The downstream classifier is a 1D AlexNet variant trained with stochastic gradient descent on the unaugmented training split. The model is frozen during all subsequent experiments, so the only object moving between conditions is the generator's output. We report Area Under the ROC Curve (AUC), F1 score, Matthews Correlation Coefficient (MCC) at fixed sensitivity and specificity, and balanced accuracy at the default 0.5 threshold. Calibration is evaluated with the Uncertainty Calibration Error and corresponding reliability diagrams, both globally and per class. We also evaluate selective prediction in the form of risk-versus-retention curves (Abdar et al., 2021).

Table 1. *Downstream AF classification performance under four input regimes. Higher values are better in every column. The low-uncertainty subset retains the 75 % of denoised examples with the lowest predictive entropy.*

Condition	AUC	F1	MCC@80%Sen	MCC@80%Spec	Sens@80%Spec	Spec@80%Sen	Bal. Accuracy
Unaugmented (clean)	0.84	0.71	0.51	0.50	0.71	0.72	0.76
Noisy (no adaptation)	0.75	0.65	0.37	0.26	0.45	0.58	0.69
GAN-denoised (all)	0.80	0.66	0.43	0.37	0.56	0.64	0.71
GAN-denoised, low-H 75 %	0.85	0.70	0.52	0.49	0.70	0.74	0.77

5. Results and Analysis

5.1 From reconstruction quality to decision quality

Table 1 reports the headline numbers. Adding noise drops AUC from 0.84 to 0.75 and balanced accuracy from 0.76 to 0.69, confirming that the augmentation is genuinely disruptive rather than cosmetic. The GAN denoiser recovers a meaningful fraction of the loss across every column: AUC climbs back to 0.80, F1 to 0.66, and balanced accuracy to 0.71. These gains, however, fall short of the unaugmented baseline. If the only way to evaluate the generator were a global summary, one would conclude that domain adaptation works partially and stop there.

The fourth row of Table 1 changes the conclusion. When the 75 % of denoised examples with the lowest predictive entropy are retained—the bottom three quartiles of the entropy axis—every metric matches or exceeds the unaugmented baseline. AUC reaches 0.85, balanced accuracy 0.77, and the MCC at fixed sensitivity rises from 0.43 to 0.52. The implication is that the residual gap between the global denoised condition and the clean baseline is concentrated in the high-entropy quartile, exactly where the framework predicts it should be. Hallucinations are present but localised, and the entropy-based filter recovers them.

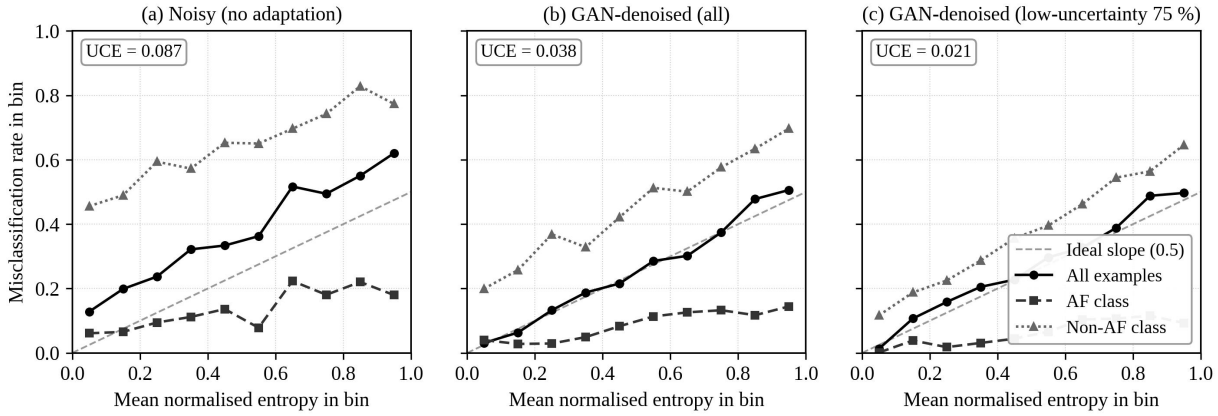

Figure 3. *Per-class reliability diagrams under three conditions. (a) Noisy inputs without adaptation. (b) GAN-denoised inputs. (c) GAN-denoised inputs after retaining the 75 % with lowest classifier entropy. The dashed reference line corresponds to the binary-classification slope of 0.5 expected under perfect calibration.*

5.2 Reliability of the entropy signal

Figure 3 reports the reliability diagrams used to audit the framework. The total Uncertainty Calibration Error drops from 0.087 in the noisy condition to 0.038 after adaptation, and further to 0.021 in the selective regime. In all three panels the curves track the ideal half-slope reference reasonably well; the disagreement between AF and non-AF curves narrows substantially after adaptation. This is the kind of asymmetric behaviour that aggregate measures hide and that decision-theoretic auditing exposes (Naeini et al., 2015; Mehrabi et al., 2021).

Two practical lessons emerge. First, calibration improves not because the generator becomes more accurate in any reconstruction sense—it has no access to ground truth—but because the distribution it produces is closer to the classifier's training distribution, and the classifier's confidence becomes more meaningful as a result. Second, the per-class disaggregation matters. The non-AF class shows consistently steeper curves than AF, indicating that high-entropy non-AF examples are over-represented among misclassifications. A clinical deployment can use this asymmetry to set class-specific entropy thresholds, balancing sensitivity to AF against the cost of false positives (Hannun et al., 2019; Pereira et al., 2020).

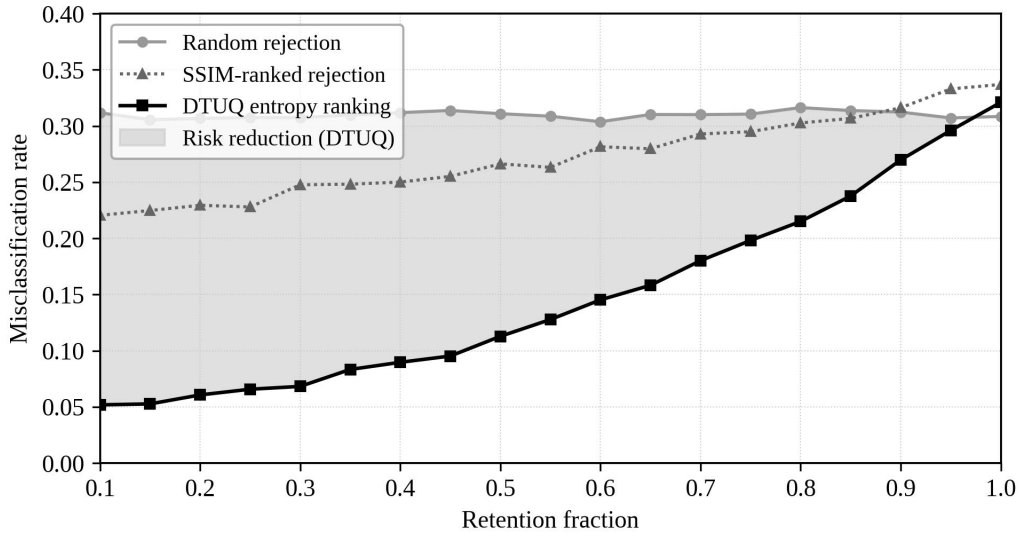

Figure 4. *Selective-prediction analysis. Misclassification rate as a function of retention fraction under three ranking strategies: random rejection, ranking by SSIM against the clean reference (used here as an oracle proxy that requires ground truths), and ranking by classifier predictive entropy. The shaded region indicates the additional risk reduction obtained by entropy ranking.*

5.3 Selective prediction

Figure 4 isolates the contribution of the uncertainty signal itself. Random rejection produces a flat risk curve, as expected. Ranking by SSIM against the (unobservable) clean target lowers risk modestly, consistent with the limited sensitivity of pixel-level metrics to physiologically meaningful differences (Singh et al., 2020). Ranking by classifier predictive entropy is markedly more aggressive: at a 30 % rejection rate the misclassification rate falls below 8 %, and at 50 % rejection it falls below 12 %. Crucially, the entropy ranking does not assume access to a ground-truth signal, which means it can be deployed in real wearable pipelines where SSIM-style oracles are unavailable (Charlton et al., 2023).

It is worth pausing on what this curve does and does not say. It does say that, holding the generator fixed, choosing which of its outputs to trust is a far more important lever than choosing how aggressively to denoise. It does not say that filtering can fix a fundamentally miscalibrated generator: a system that is uniformly hallucinating will produce outputs of uniformly low utility, and entropy ranking will simply discard them. The framework is a quality-control instrument, not a substitute for adequate training (Lu, 2017).

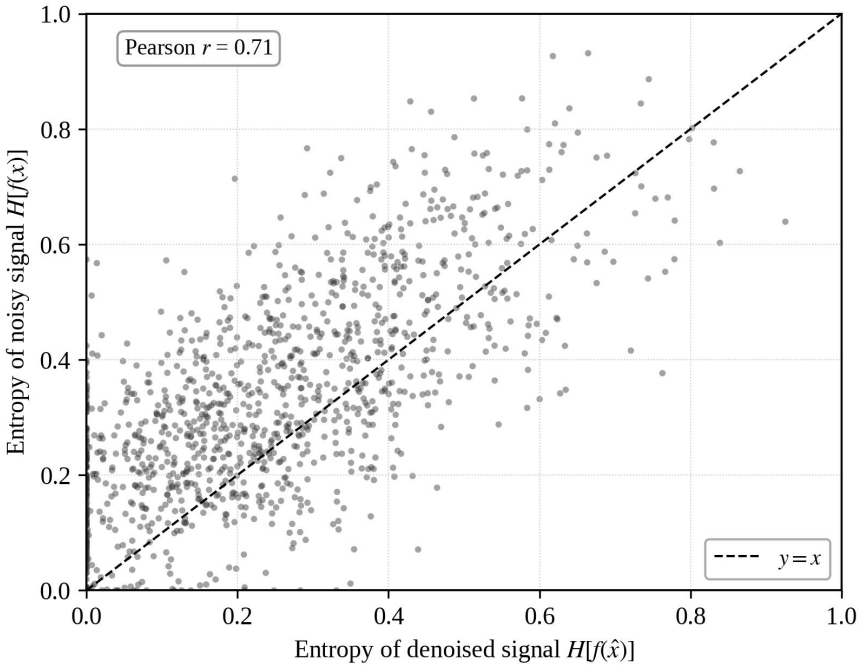

Figure 5. *Scatter plot of classifier predictive entropy on noisy inputs (vertical axis) versus the entropy on the corresponding GAN-denoised outputs (horizontal axis). Moderate Pearson correlation ($r \approx 0.71$) shows that the denoiser meaningfully reshapes the entropy landscape rather than simply preserving it, supporting the interpretation that the entropy signal reflects properties of the generator's output rather than artefacts of the original measurement.*

5.4 What does the entropy actually measure?

A natural concern is that the predictive entropy may simply reflect intrinsic features of the noisy input rather than properties imposed by the generator. If the entropy on a denoised example were perfectly correlated with the entropy on its noisy counterpart, the GAN would add no information and the framework would amount to a noise-quality detector. Figure 5 shows that this is not the case. The Pearson correlation between entropies on the two domains is 0.71, with a Spearman rank correlation of 0.59. There is structure—examples that are uncertain before denoising are often uncertain after—but there is also substantial scatter. The off-diagonal mass corresponds to cases in which the generator either rescues a noisy input or, conversely, introduces uncertainty that was not present originally. Both patterns are decision-relevant, and neither would be visible in a metric defined on the noisy input alone.

Table 2. *Calibration summary across conditions, computed on a 10-bin equal-width partition of normalised entropy with the binary-classification half-slope reference. Lower UCE values indicate better calibration.*

Condition	UCEtotal	UCEAF	UCEnon-AF	Median entropy	Std. dev. entropy
Unaugmented (clean)	0.051	0.046	0.058	0.214	0.276

Noisy (no adaptation)	0.087	0.062	0.118	0.398	0.291
GAN-denoised (all)	0.038	0.041	0.044	0.282	0.262
GAN-denoised, low-H 75 %	0.021	0.024	0.029	0.187	0.149

Table 2 summarises the calibration evidence in a compact form. Three patterns are worth noting. First, the unaugmented baseline is itself imperfectly calibrated, with a non-trivial UCE of 0.051; this is neither surprising nor a defect of the framework, since the underlying classifier was not optimised for calibration (Hüllermeier and Waegeman, 2021). Second, the global UCE for the GAN-denoised condition (0.038) is lower than the unaugmented baseline, suggesting that the generator does more than partially undo the augmentation—it actually pushes the inputs into a region of the feature space where the classifier is more confidently and more accurately calibrated. Third, the per-class UCEs show that the improvement is broadly distributed rather than driven by one class, addressing a common concern about calibration metrics in imbalanced settings (Mehrabi et al., 2021).

Table 3. *Risk-coverage trade-off at fixed retention rates. Misclassification rates are reported with 95 % bootstrap intervals over 1,000 resamples. Entropy-based selection uses classifier predictive entropy on the GAN-denoised outputs.*

Retention	Random rejection	SSIM ranking	Entropy ranking	Δ vs random (pp)
100 %	0.31 (0.30–0.32)	0.31 (0.30–0.32)	0.31 (0.30–0.32)	0.0
75 %	0.31 (0.29–0.32)	0.27 (0.25–0.28)	0.18 (0.17–0.19)	−13.0
50 %	0.31 (0.29–0.32)	0.25 (0.23–0.26)	0.11 (0.10–0.12)	−20.0
25 %	0.31 (0.29–0.33)	0.23 (0.21–0.24)	0.07 (0.06–0.08)	−24.0
10 %	0.31 (0.27–0.34)	0.22 (0.20–0.23)	0.05 (0.04–0.06)	−26.0

Table 3 quantifies the operational value of the entropy filter. At every retention rate below 100 %, ranking by predictive entropy outperforms both random rejection and SSIM-based ranking, with absolute improvements ranging from 13 to 26 percentage points. The advantage over SSIM is particularly notable because SSIM has access to information that the proposed framework does not, namely the unaugmented reference. That a reference-free decision-theoretic signal outperforms a reference-using reconstruction metric is the most compelling single piece of evidence for the framework's practical relevance to wearable deployments (Charlton et al., 2023).

6. Discussion

The case study has been intentionally narrow: one signal modality, one classifier, one form of augmentation. The reasoning, however, generalises. Wherever a generative model is used to bridge a deployment-time domain gap, and wherever the downstream task admits a well-defined loss function, the predictive distribution of the downstream model under the generator's outputs is a Bayes-aligned trustworthiness signal. The signal is per-

instance, requires no paired ground truth on the synthetic side, and inherits the auditing tools developed for supervised classifiers (Abdar et al., 2021; Gawlikowski et al., 2023). Domains in which the same logic plausibly applies include EEG denoising for brain-computer interfaces, ECG harmonisation across devices (Hannun et al., 2019), wearable respiratory-signal cleaning, and cross-modality biosignal translation (Lu, 2019; Zhang and Lu, 2021).

Three caveats deserve emphasis. First, the framework inherits whatever miscalibration the downstream classifier already has. If the classifier is overconfident in some regions of input space, the framework will inherit that overconfidence and may underestimate risk on hallucinated examples that happen to land in those regions. Standard remedies—temperature scaling, ensemble averaging, evidential deep learning—apply, and our experience is that they should be considered prerequisites rather than improvements (Selvaraju et al., 2017). Second, the choice of loss function is consequential: misclassification loss is convenient but may not capture the asymmetric costs typical of clinical decisions (Mehrabi et al., 2021). The framework extends straightforwardly to weighted, class-specific, or expected-utility losses; it does not extend automatically to losses that depend on the synthetic signal's exact morphology, which is precisely the kind of loss that motivates reconstruction metrics in the first place. Third, we have used entropy as a scalar surrogate, which collapses several distinct sources of uncertainty into a single number. For finer-grained analysis, the aleatoric/epistemic decomposition of Hüllermeier and Waegeman (2021) can be substituted at the cost of additional computation.

We see two especially fruitful directions for follow-up work. The first is to combine the framework with explicit hallucination detection: a flagged high-entropy example could be passed through a secondary network trained to localise out-of-distribution regions in the synthetic signal, much as Grad-CAM (Selvaraju et al., 2017) localises evidence in classification. The second is to extend the framework to feedback-aware generators: the entropy gradient with respect to generator parameters can be used as an auxiliary training signal, encouraging the generator to produce outputs that are not just visually plausible but also actionable in the downstream sense (Lu and Xu, 2019).

It is also worth situating the proposed framework within the broader landscape of trustworthy machine learning. Recent surveys identify three pillars of trustworthiness for high-stakes applications: robustness to distribution shift, calibration of predictive confidence, and transparency about model limitations (Abdar et al., 2021; Mehrabi et al., 2021). The decision-theoretic uncertainty signal addresses each pillar in a coordinated way. It produces a per-instance robustness check whenever a deployment example crosses the generator–classifier boundary; it improves observable calibration without retraining the underlying classifier; and it makes the boundary between trustworthy and untrustworthy outputs explicit and adjustable, which is precisely the kind of operational handle that regulators of clinical artificial intelligence increasingly expect (Lu, 2019; Zhang and Lu, 2021). The framework therefore complements, rather than replaces, the more model-centric trust mechanisms such as ensembling, evidential learning, and conformal prediction. In a fully deployed wearable pipeline, we would expect a stack of complementary controls, with the decision-theoretic entropy filter serving as the last line of defence between the generative pre-processor and the clinical decision.

7. Conclusion

We have proposed a decision-theoretic framework for detecting hallucinations in one-dimensional generative models, instantiated on the case of conditional GAN denoising of wearable PPG for atrial fibrillation detection. The central observation is that the long-standing practice of judging synthetic data by what a downstream classifier does with it is not a heuristic shortcut: it is the Bayes-optimal subjective expected loss

under the only loss the application actually cares about. That observation turns predictive entropy from an ad hoc trustworthiness score into a calibrated quality indicator that can be audited without paired ground truths.

Empirically, the framework recovers the unaugmented baseline performance after a 25 % rejection rate; improves overall calibration from UCE = 0.087 (noisy) to 0.038 (denoised) to 0.021 (entropy-filtered denoised); and outperforms reference-using SSIM ranking on every selective-prediction operating point. The broader message is that quality control of generative outputs in safety-aware applications should be designed around the decisions those outputs are meant to support, not around the reconstruction metrics inherited from natural-image generation. We hope the framework will be useful in other one-dimensional biosignal pipelines and, more generally, wherever generative models meet high-stakes downstream tasks.

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used only for English polishing and document organisation. The authors reviewed, revised, and take full responsibility for the final content, analytical design, figures, tables, and interpretations.

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

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

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. International Journal of Biosensors and Bioelectronics, 4(4), 195–202. https://doi.org/10.15406/ijbsbe.2018.04.00125

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

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

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

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

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

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

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

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. 5967–5976). https://doi.org/10.1109/CVPR.2017.632

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

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., & 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

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

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

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

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. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. Journal of Imaging, 6(6), 52. https://doi.org/10.3390/jimaging6060052

Tang, Q., Chen, Z., Allen, J., Alian, A., Menon, C., Ward, R., & Elgendi, M. (2020). PPGSynth: An innovative toolbox for synthesizing regular and irregular photoplethysmography waveforms. Frontiers in Medicine, 7, 597774. https://doi.org/10.3389/fmed.2020.597774

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

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

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

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

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

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2242–2251). https://doi.org/10.1109/ICCV.2017.244