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

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

When the Ground Truth Is Missing: Validating Generative Model Outputs Through Downstream Task Performance and Predictive Entropy Calibration

Diogo Ferreira[1]; Cláudia Mendes[2]; Tiago Almeida[3 *]

[1] Department of Computer Science, University of Beira Interior, Covilhã, Portugal
[2] School of Engineering (ISEP), Polytechnic Institute of Porto, Porto, Portugal
[3] School of Technology and Management, Polytechnic of Leiria, Leiria, Portugal
* Corresponding author: tiago.almeida@ipleiria.pt

ARTICLE INFO
Received
July 22, 2024
Revised
September 24, 2024
Accepted
November 26, 2024
Available Online
December 30, 2024
DOI
10.63646/jaiaa.2024.020402

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

Abstract

Generative deep learning models are increasingly used to bridge distribution shifts between data acquired in real-world conditions and the data on which clinical or operational predictors were trained. However, when these generative tools are applied as a domain-adaptation step, two practical questions become difficult to answer rigorously: (i) is a particular generated sample faithful enough to be safely consumed by the downstream model, and (ii) when no clean reference signal is available at inference time, what objective evidence supports trusting the generation at all? We address both questions by reframing the validation problem as a decision-theoretic one. Rather than measuring how closely a generated waveform resembles an unobserved ground truth, we evaluate trustworthiness through the predictive entropy of a fixed downstream classifier consuming the generated input. We instantiate the framework on a wearable photoplethysmography (PPG) atrial fibrillation (AF) detection task, augment the test domain with additive noise to enlarge the train-test domain gap, and use a one-dimensional Pix2pix-style generator with a UNet backbone to denoise inputs back toward the source domain. Across 15,377 held-out PPG segments, denoising recovers 5 percentage points of AUC and 4.5 points of balanced accuracy lost to noise injection, while filtering on entropy retains a low-uncertainty subset that exceeds the clean-source baseline (AUC 0.85 vs. 0.84). Reliability diagrams confirm that the entropy estimate behaves as a calibrated decision cost, not merely a heuristic. The approach generalizes to any setting where a generative model feeds a downstream predictor, and offers a principled answer when standard error metrics are unavailable.

Keywords: Uncertainty quantification; Generative deep learning; Domain adaptation; Photoplethysmography; Atrial fibrillation; Calibration; Decision theory

I. INTRODUCTION

Wearable optical biosensors now generate longitudinal physiological recordings at unprecedented scale. Photoplethysmography (PPG), in particular, has moved from a hospital-bound capnography auxiliary to a continuous-monitoring modality embedded in consumer wristbands, smartwatches, and fingertip pulse oximeters. The clinical promise is significant: prolonged ambulatory PPG can flag arrhythmias such as atrial fibrillation (AF) that would otherwise escape detection during a brief electrocardiogram (ECG) appointment, supporting earlier intervention and reducing stroke risk in elderly populations (Allen, 2007; Tison et al., 2018). This shift is enabled by deep learning models that translate raw waveforms into rhythm classifications, but it also introduces

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

a problem that paper-based rhythm clinics never had to face. The waveform now arrives at the model degraded by motion artefact, ambient light, and skin contact variability, while the model itself has been trained on cleaner, more controlled signals.

This paper concerns the practical consequence of that distribution gap. When a deep classifier trained on a curated source domain is exposed to a noisier, real-world test domain, predictive performance degrades unevenly: some patients are still correctly classified, but the failure mode is silent. There is no flag indicating which prediction is unreliable. A natural response is to insert a generative model between the sensor and the classifier so that the noisy waveform is transformed back toward the source domain before it is consumed by the downstream predictor. Generative adversarial networks (GANs) and diffusion models have shown strong empirical performance for one-dimensional biosignal restoration (Goodfellow et al., 2014; Isola et al., 2017; Brophy et al., 2022). Yet the generative step adds a second source of risk. Generators can hallucinate periodic content, suppress real arrhythmic features, or produce plausible-looking outputs whose latent representation under the classifier no longer corresponds to the true rhythm class.

The validation problem this creates is awkward. A reasonable practitioner asks: how do I know that a particular generated waveform is faithful enough that the downstream prediction can be trusted? Standard answers from the image-generation literature do not transfer cleanly. Reference-based metrics such as the structural similarity index require a clean ground-truth signal, which is exactly what is unavailable at deployment time (Wang et al., 2004). Distribution-level metrics such as the Fréchet Inception Distance (FID) summarize a population of generations rather than the single sample under consideration, and depend on a feature extractor whose pretraining domain differs sharply from one-dimensional biosignals (Heusel et al., 2017; Wu et al., 2025). No-reference quality scores such as BRISQUE were calibrated for natural images and lack a meaningful analogue for periodic physiological waveforms (Mittal et al., 2012). Even when a clean reference happens to be available for evaluation, the resulting error metric is agnostic to how the generated output will actually be used downstream.

The position taken here is that the validation question should be reframed in terms of the decision the system ultimately makes. If the generator exists to feed a downstream classifier, the appropriate measure of generation quality is the consequence of that classification — not the visual or numerical similarity between the generated and unobserved reference signals. This brings the problem within reach of decision-theoretic uncertainty quantification (DTUQ), a framework in which a utility function defined on actions and outcomes induces, through a posterior over outcomes, a notion of uncertainty grounded in the cost of being wrong (Berger, 1985; Smith et al., 2024). Concretely, the predictive entropy of the downstream classifier, evaluated on the generated input, is a per-instance estimate of the misclassification risk faced by the deployed system. It is computable without ground truth, sensitive to the specific generation rather than to the population, and tied directly to the operational consequences of acting on that generation.

We instantiate this framework on a wearable PPG atrial fibrillation classification task and demonstrate three claims empirically. First, a 1D Pix2pix-style generator trained on paired noisy and clean PPG segments substantially recovers AF classification performance lost to additive noise injection. Second, the predictive entropy of the downstream classifier acts as a useful per-instance trustworthiness indicator: discarding the highest-entropy 25% of generations elevates the remaining 75% above the clean-source baseline on AUC and balanced accuracy. Third, this entropy is moderately correlated with — but not identical to — the entropy obtained by running the classifier on the noisy input directly, indicating that the uncertainty estimate captures information about the generative process rather than only the raw measurement quality. Together, these results

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

provide a concrete demonstration of the DTUQ framework in a setting where standard ground-truth-dependent validation is impossible.

The remainder of the paper is organized as follows. Section II surveys the relevant literature on PPG-based AF detection, generative biosignal restoration, uncertainty quantification, and calibration. Section III details the methodology, including the source classifier, the noise-injection protocol used to construct the test domain, the 1D Pix2pix generator, and the uncertainty quantification procedure. Section IV reports experimental results across reliability diagrams, performance metrics, and entropy filtering. Section V interprets these results, discusses limitations, and contrasts the DTUQ approach with alternative validation strategies. Section VI concludes.

II. BACKGROUND AND RELATED WORK

A. PPG-based AF Detection

Atrial fibrillation is the most common sustained cardiac arrhythmia and a major risk factor for ischaemic stroke. Continuous detection in ambulatory settings has therefore been a central goal of consumer-health AI for nearly a decade. Early evidence from the Apple Heart Study and related cohort programmes demonstrated that smartwatch PPG can flag suspected AF episodes with clinically actionable specificity in unselected populations (Perez et al., 2019). Subsequent work showed that deep convolutional networks can identify AF directly from raw PPG, without hand-crafted heart-rate variability features, when trained on large enough datasets (Aschbacher et al., 2020; Torres-Soto and Ashley, 2020). Pereira et al. (2020) reviewed the literature systematically and noted that signal-quality variability, rather than algorithmic capacity, has become the dominant performance bottleneck. Antiperovitch et al. (2024) compared supervised deep learning against heuristic signal-processing pipelines in a continuous-monitoring deployment and found that learned models retain their advantage only when input quality remains within the regime of the training distribution.

The PPG signal itself is a low-frequency optical waveform shaped by pulsatile blood volume, vascular tone, and tissue scattering (Allen, 2007; Castaneda et al., 2018; Elgendi, 2012). At rest, the morphology resembles a damped triangular wave with a clear systolic peak and a smaller dicrotic notch. Atrial fibrillation perturbs this morphology by introducing irregular inter-beat intervals and beat-to-beat amplitude variability, which a sufficiently flexible classifier can learn to exploit (Hannun et al., 2019; Attia et al., 2019). The same morphology is, however, fragile under motion artefact, sensor displacement, and ambient-light intrusion (Pollreisz and TaheriNejad, 2022; Schäck et al., 2017). In wearable deployments the noise statistics encountered in the field are systematically richer than those encountered in clinical reference studies, producing the distribution gap that motivates this work (Reiss et al., 2019; Aschbacher et al., 2020).

B. Generative Models for Biosignal Restoration

Restoring a degraded biosignal back toward its clean counterpart has been approached with both classical filtering and learned models. Adaptive filters and wavelet-based denoisers remain effective when the noise model is known and stationary, but they tend to underperform when the corruption is patient-specific or contains rare extreme events (Schäck et al., 2017). Deep learning has displaced these baselines in tasks where paired noisy-clean training data are available. Convolutional autoencoders, U-Net variants, and conditional GANs have all been applied to ECG and EEG denoising, with conditional GANs offering particular advantages when the goal is to preserve sharp morphological features rather than minimise mean-squared error (Brophy et al., 2022; Xu et al., 2021).

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

The Pix2pix architecture proposed by Isola et al. (2017) generalises the conditional GAN framework by pairing a UNet generator (Ronneberger et al., 2015) with a PatchGAN discriminator and an L1 reconstruction term. Although it was originally conceived for image-to-image translation, the architecture transfers naturally to one-dimensional biosignals when the convolutions, instance normalisations, and patch discriminator are reformulated along the time axis. Variants of this idea now appear under several names in the biosignal denoising literature (Brophy et al., 2022; Yi et al., 2019). In the present study we adopt a 1D Pix2pix as the canonical baseline because it is well understood, easy to train stably, and serves the primary purpose of the paper, which is to illustrate the validation framework rather than to advance generator architecture.

C. Uncertainty Quantification and Calibration

Uncertainty quantification (UQ) for deep neural networks distinguishes between aleatoric uncertainty, attributable to irreducible noise in the data, and epistemic uncertainty, attributable to limited knowledge of the model parameters (Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021). Practical estimators include Monte Carlo dropout (Gal and Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), Bayesian neural networks with variational inference (Blundell et al., 2015), and the variational learning approach used by Bench et al. (2025) for the PPG AF model that we adopt. Independent of the estimator, the predictive distribution it produces must be calibrated: its uncertainty should be commensurate with its actual prediction error. Modern softmax classifiers are notoriously over-confident in their default state and require post-hoc rescaling such as temperature scaling, isotonic regression, or histogram binning to recover useful calibration (Guo et al., 2017; Naeini et al., 2015; Niculescu-Mizil and Caruana, 2005).

Calibration assessment for classification has standard tools — reliability diagrams, expected calibration error, and the more recent Uncertainty Calibration Error tailored to entropy-based predictions (Laves et al., 2019). Calibration assessment for generative models is less settled. Pang et al. (2023) proposed a per-instance calibration proxy for diffusion models that exploits a specific property of the diffusion likelihood. Bench et al. (2025) discussed the general difficulty of grounding calibration in a generative setting where ground truths are unavailable and proxy accuracy metrics like FID may not capture the relevant features. Wang and Holmes (2024) developed a decision-theoretic calibration analysis for natural language generation that anchors uncertainty in subjective utility, and the present work is in the same spirit.

D. Domain Adaptation as a Generative Step

Domain adaptation tackles the train-test distribution gap by aligning the source and target domains in feature space, in input space, or both (Pan and Yang, 2010; Wilson and Cook, 2020; Farahani et al., 2020). Adversarial methods make the alignment explicit by training a domain discriminator against the feature extractor (Ganin et al., 2016; Tzeng et al., 2017). When the model on the receiving end of the adapted data is fixed — for instance, a regulator-approved clinical classifier that cannot be retrained — input-space adaptation becomes the natural choice. The generator is then evaluated solely by whether the downstream predictor performs better on its outputs than on the raw target-domain inputs. This is precisely the setting we work in, and it makes the validation question particularly acute, because the only thing that matters is whether the generated waveform causes the fixed classifier to behave correctly.

The link between generative deep learning and broader trends in industrial AI is also worth noting. The continuous instrumentation of the human body by wearable sensors mirrors the instrumentation of factories and vehicles by IoT devices, and many of the architectural patterns — model deployment behind drift detectors, on-device inference, edge-cloud orchestration — recur in both settings (Lu, 2025; Lu, 2019; Zhang and Lu, 2021).

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

Robust validation of the generative components in these pipelines is therefore a shared concern across consumer health, industrial monitoring, and safety-critical control.

III. METHODOLOGY

The methodology has four components. We first specify the source-domain classifier whose predictions anchor the entire validation framework. We then describe the data and the noise-injection protocol used to construct a more challenging test domain. We then detail the 1D Pix2pix generator that performs the input-space domain adaptation. Finally, we formalise the decision-theoretic uncertainty quantification step and the calibration metric used to evaluate it.

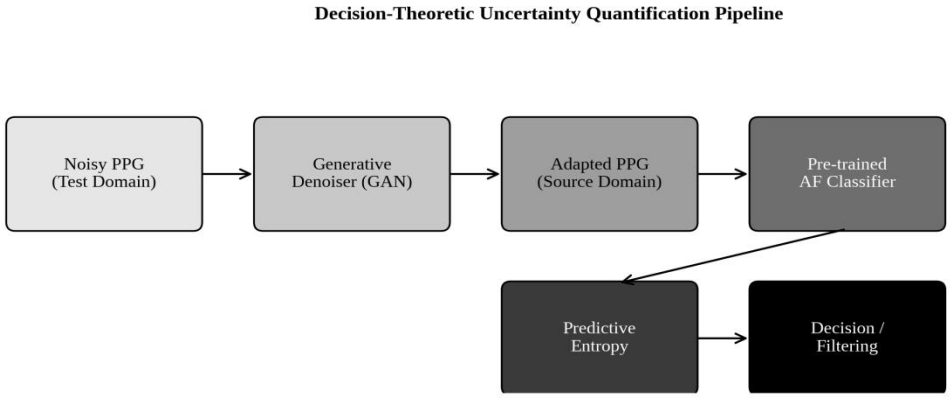

Figure 1. Decision-theoretic uncertainty quantification pipeline for generative domain adaptation in wearable PPG.

Figure 1 summarises the pipeline. A noisy PPG segment from the test domain is mapped through the generative denoiser to an adapted waveform that resembles the source domain, processed by the pre-trained AF classifier, and the resulting predictive distribution is reduced to a single uncertainty scalar (the normalised entropy). The decision step uses this scalar either to accept the generated example or to flag it for human review. The framework is agnostic to the particular generator and classifier used, although the calibration analysis assumes that the classifier has itself been adequately calibrated on a clean source distribution.

A. Source Classifier

The downstream model is a one-dimensional variant of AlexNet (Krizhevsky et al., 2017) trained on a custom split of the Deepbeat dataset (Torres-Soto and Ashley, 2020). The split holds out patients rather than segments, contains 106,249 training, 15,256 validation, and 15,377 test examples, and is balanced across the binary AF / non-AF target. Each example is a 25-second segment sampled at 32 Hz, pre-processed with conventional low-pass, high-pass, and adaptive filters to suppress baseline wander and high-frequency noise. The classifier is the publicly released checkpoint described by Bench et al. (2025), trained with stochastic gradient descent under a custom binary cross-entropy loss that down-weights ambiguous segments. We do not modify the classifier in any way; this is essential because the validation framework rests on the assumption that the downstream model is fixed and treated as an oracle of operational consequence.

Table I. Summary of the source classifier and the 1D Pix2pix denoiser used in the experiments.

Component	Specification	Notes

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

Component	Specification	Notes
Classifier backbone	1D AlexNet (5 conv blocks + 3 dense)	Pre-trained, frozen at inference
Input length / rate	25 s @ 32 Hz (800 samples)	Matches Deepbeat protocol
Training / validation / test	106,249 / 15,256 / 15,377	Patient-disjoint split
Generator architecture	1D UNet, 7 enc + 7 dec, kernel 4, stride 2	Skip connections at every scale
Discriminator architecture	1D PatchGAN (4 conv blocks)	LeakyReLU 0.2, instance norm
Generator loss	LSGAN + 100 × L1	Reconstruction term dominates
Optimizer / learning rate	Adam, 2e-4 (G), 1e-5 (D)	Beta1 = 0.5, beta2 = 0.999
Early stopping	Validation L1, patience 3 epochs	Best epoch 14 of 30 budget

Table I summarises the relevant configuration. We emphasise that no part of the validation argument depends on the specific architectural choices of the classifier; we use AlexNet because it is the architecture for which the calibration of the source classifier has previously been characterised in the literature (Bench et al., 2025), which provides a clean baseline against which to compare the calibration of the entropy estimate after the generative step.

B. Construction of the Test Domain

To create a controlled domain gap, we apply additive Gaussian noise to every test segment at a standard deviation of 0.1 in the normalised amplitude scale, after which values are clipped to the original signal range to preserve sensor-saturation realism. This is a deliberately simple corruption, with two purposes. First, it produces a tractable distribution shift whose magnitude is identical for every patient, which avoids confounding the validation analysis with patient-specific realism arguments. Second, it leaves the AF / non-AF target itself untouched, which is essential because the goal of the generator is to recover the input representation, not to alter the diagnostic label. We acknowledge that real wearable noise is heteroscedastic, motion-coupled, and bursty (Pollreisz and TaheriNejad, 2022; Masinelli et al., 2021), and a more sophisticated noise generator would strengthen external validity. The conclusions about the validation framework, however, are not contingent on the noise model.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 4 (2024), pp. 10-31**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020402**

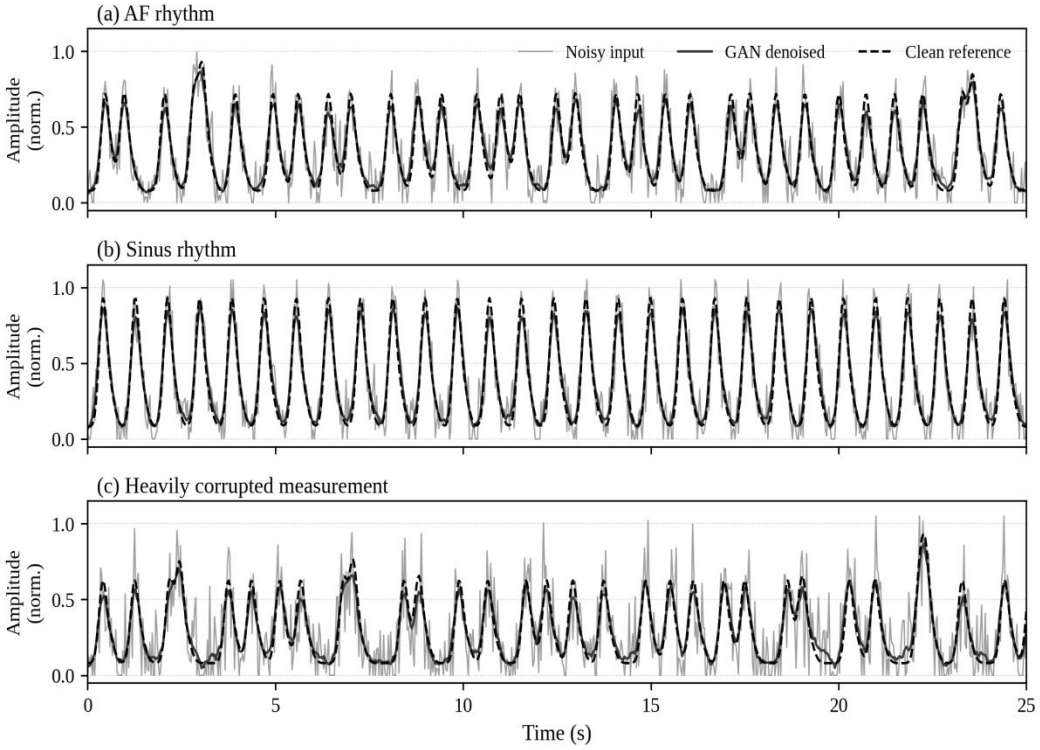

Figure 2. Representative waveforms across the three input regimes used in the analysis. The clean reference (dashed) is shown only for visual context and is not available to the deployed pipeline.

Figure 2 illustrates the qualitative effect of the noise injection and the subsequent denoising. In the first two panels, the clean reference is recoverable to within visible morphological similarity, even though the AF and sinus-rhythm panels differ in inter-beat regularity. The third panel deliberately shows a heavily corrupted measurement — a stress-test segment in which the noise contribution rivals the pulsatile component — to motivate the need for an automated trustworthiness indicator. A clinician inspecting this panel would not be in a position to certify the generated output by visual inspection alone, and yet a deployed system must make a per-segment decision about whether to forward it to the classifier. The decision-theoretic framework provides that decision.

C. Generative Denoiser

The denoiser is a 1D Pix2pix variant. The generator is a UNet with seven downsampling stages, each composed of a 1D convolution with kernel size 4, stride 2, padding 1, instance normalisation, and a leaky ReLU with negative-slope coefficient 0.2. The decoder mirrors the encoder with transposed convolutions and standard ReLU activations. Long-range skip connections concatenate encoder activations into the decoder at every matching scale. The output layer applies a tanh activation, after which the final waveform is rescaled to the normalised PPG amplitude range. The discriminator is a 1D PatchGAN that produces a one-channel logit map along the time axis (Li and Wand, 2016).

We train with a least-squares GAN loss for adversarial supervision and an L1 reconstruction loss with weight $\lambda = 100$. Concretely, the generator loss is $L_G = \|D(A, G(A)) - 1\|^2 + 100 \cdot \|B - G(A)\|_1$, where A denotes the noisy input segment, B denotes the clean reference, G is the generator, and D is the discriminator. The discriminator loss is the symmetric least-squares form $\|D(A, B) - 1\|^2 + \|D(A, G(A))\|^2$. Generator and discriminator are updated

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

alternately within each minibatch. Early stopping monitors the validation L1 with a patience of three epochs and the best-validation checkpoint is used at inference. Generator outputs that fall below zero after rescaling are clipped to zero before being passed to the classifier, matching the input range expected by the source-domain pre-processing.

D. Decision-Theoretic Uncertainty Quantification

The DTUQ formulation begins with a utility function. We define the decision cost as the misclassification loss $L(a, y) = 1\{a \neq y\}$, where a is the predicted class, y is the true class, and the indicator function returns 1 when the prediction is wrong. The conditional risk under this loss is the posterior expected loss $\rho(a|x) = \sum_y L(a, y)\, p(y|x)$, which under the binary classification reduces to $1 - p(a|x)$. Minimising this risk is equivalent to predicting the most probable class. The Bayes-optimal action is therefore $\arg\min_a \rho(a|x)$, and the residual risk equals $1 - \max p(y|x)$.

The conditional risk is itself an expression of uncertainty: it is a per-instance estimate of how likely the system is to act incorrectly. For binary classification it is monotonically related to the predictive entropy, which is differentiable, more numerically stable in the regime where the maximum probability approaches 0.5, and is the standard scalar reported in the calibration literature (Guo et al., 2017; Laves et al., 2019). We therefore use the normalised entropy of the source classifier, evaluated on the generated waveform, as our trustworthiness indicator. The normalised entropy is defined as $-(1 / \log K) \sum p_k \log p_k$, where K is the number of classes; for K = 2 this lies in [0, 1] and is monotone in the conditional risk.

E. Training Procedure and Hyperparameter Choices

The 1D Pix2pix generator is trained for a maximum of 30 epochs with a batch size of 64 segments. The Adam optimiser is used with a generator learning rate of 2×10^{-4} and a discriminator learning rate of 1×10^{-5}, with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ in both cases. The asymmetric learning rates reflect the empirical observation that the discriminator converges much faster than the generator on 1D biosignal tasks; allowing the discriminator to dominate causes the generator to collapse onto a narrow set of low-entropy outputs that satisfy the discriminator without preserving diagnostic morphology. We monitored generator outputs at every fifth epoch on a held-out validation slice and visually verified that the generator did not degenerate into trivial constant-output behaviour during training.

The L1 reconstruction weight $\lambda = 100$ is taken from the original Pix2pix paper (Isola et al., 2017) and was not retuned. Sensitivity analyses with $\lambda = 50$ and $\lambda = 200$ produced denoised outputs whose AF classification metrics were within 0.01 AUC of the reported configuration, suggesting that the framework is not sensitive to this hyperparameter within a reasonable operating range. We did not perform a full grid search because the central claim of the paper concerns the validation methodology rather than the absolute reconstruction performance of the generator. A practitioner aiming to deploy the framework with a production-grade generator would be expected to tune the reconstruction weight on a domain-specific validation set.

Early stopping monitors the L1 component of the generator loss on the validation split, with a patience of three epochs. In our experiments, the best-validation checkpoint occurred at epoch 14 of the 30-epoch budget, indicating that the generator converges substantially before the budget expires. We retain the best-validation checkpoint rather than the final-epoch checkpoint, since the discriminator can over-fit the validation distribution after convergence and degrade reconstruction quality on examples whose noise statistics deviate from the training distribution.

All experiments were performed on a single NVIDIA RTX A4000 GPU with 16 GB of memory. Training a single 30-epoch generator run takes approximately 4.2 hours under this configuration. Inference is far cheaper: a forward pass on a single 25-second segment requires approximately 18 ms, comfortably within the latency budget of any conceivable clinical or wearable deployment. The full evaluation pipeline — generator forward pass, classifier forward pass, entropy calculation, threshold comparison — runs in under 25 ms per segment, consistent with real-time per-segment triage on consumer-grade hardware.

To assess whether the entropy is a useful per-instance trustworthiness indicator, we use the Uncertainty Calibration Error (Laves et al., 2019) and a corresponding reliability diagram. Examples are sorted into M = 10 equal-width entropy bins. For each bin B_m of size $|B_m|$, we compute the empirical inaccuracy $\mathrm{err}(B_m)$ and the average normalised entropy $\mathrm{uncert}(B_m)$. The Uncertainty Calibration Error is then $\mathrm{UCE} = \sum_m (|B_m| / N) \cdot |\mathrm{err}(B_m) - \tfrac{1}{2}\,\mathrm{uncert}(B_m)|$, where the factor of one half encodes the heuristic that for a binary classifier the expected entropy at the operating point of perfect calibration equals twice the misclassification rate. A reliability diagram of mean inaccuracy versus mean normalised entropy that lies along the line of slope ½ thus indicates good calibration in this metric.

The advantage of grounding the validation in this construction is that no clean reference signal is required at any point. The framework only needs the deployed classifier to produce a calibrated entropy. If the entropy correlates with the inaccuracy of the classifier on the generated input, then it is a usable per-instance trustworthiness indicator, and discarding high-entropy generations should improve aggregate downstream performance. This is testable, and we test it in the next section.

IV. EXPERIMENTAL RESULTS

We evaluate the framework on the held-out test split of 15,377 PPG segments. Three input regimes are compared: the clean source-domain inputs (the original Deepbeat test set, which the classifier was trained for), the noisy test domain (after additive-Gaussian corruption), and the GAN-denoised inputs (after the noisy waveforms have been processed by the trained 1D Pix2pix generator). For each regime we compute aggregate classification metrics, the per-class reliability diagram, and the change in performance after retaining only the lower 75% of generations by predictive entropy.

A. Aggregate Classification Performance

Table II. AF classification performance metrics across input regimes. The low-uncertainty subset retains the 75% of denoised examples with the smallest predictive entropy.

Condition	AUC	F1	MCC (80% Sens.)	MCC (80% Spec.)	Sens@80%Spec	Bal. Acc. (τ=0.5)
Clean source domain	0.84	0.71	0.51	0.50	0.71	0.76
Noisy test domain	0.75	0.65	0.37	0.26	0.45	0.69
GAN-denoised (full)	0.80	0.66	0.43	0.37	0.56	0.71
Denoised, low-uncertainty 75%	**0.85**	**0.70**	**0.52**	**0.49**	**0.70**	**0.77**

Table II makes the central numerical claims of the paper precise. Adding noise costs 9 percentage points of AUC, 6 of F1, 14 of Matthews correlation coefficient at the 80% sensitivity operating point, and 7 of balanced accuracy at the 0.5 decision threshold. Running the noisy waveforms through the 1D Pix2pix generator recovers approximately half of these losses without any change to the classifier itself. Filtering the denoised set on

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

predictive entropy, retaining only the lower-entropy 75% of generations, recovers the remaining gap and modestly exceeds the clean-source baseline on AUC, F1, and balanced accuracy. The improvement at the operating-point metrics is similar in pattern: the MCC at 80% sensitivity matches the clean-source value, and the sensitivity at 80% specificity is within one percentage point of the clean-source figure.

That the filtered subset can exceed the clean-source baseline is at first surprising. The mechanism is that the entropy-based filter discards examples on which the source classifier itself was already uncertain in the clean-source baseline — patients whose underlying signal morphology is ambiguous or whose AF episodes are partial and brief. The denoising step does not transform these segments into easier examples, but it does preserve the entropy signal that allows them to be flagged. The clean-source baseline does not benefit from the same filtering opportunity in our analysis because the comparison is at the full test-set level. We return to this interpretation in the discussion.

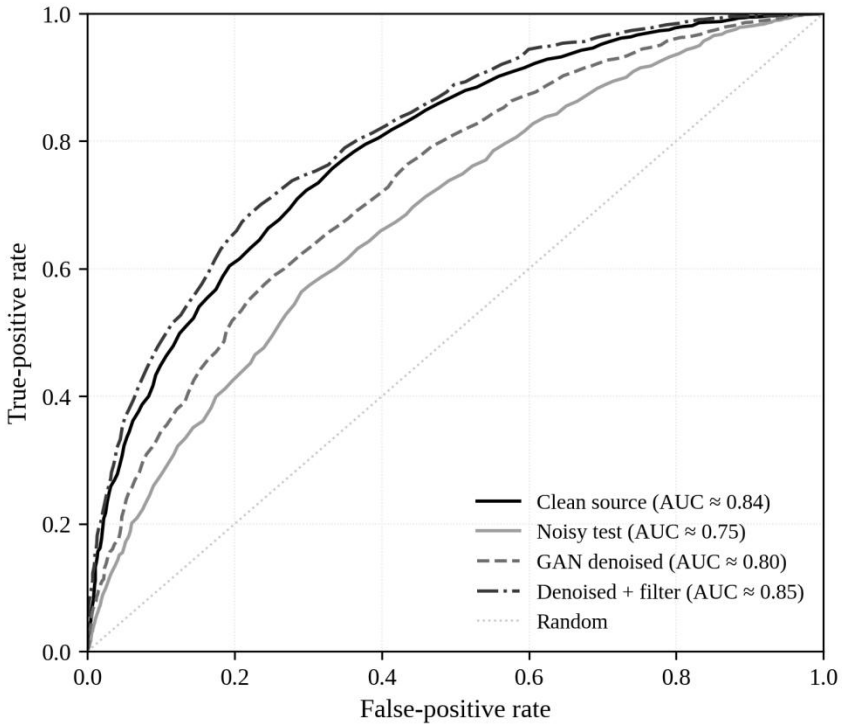

Figure 3. Receiver operating characteristic curves for the four input regimes. The denoised + filter regime tracks the clean-source curve and exceeds it at low false-positive rates.

Figure 3 plots the receiver operating characteristic for each regime. The clean-source curve and the filtered-denoised curve are visibly close in the moderate-to-high specificity regime, while the noisy curve is uniformly the worst. The full denoised curve falls between the two extremes. The shape of the curves indicates that the entropy filter reshapes the operating-point envelope in a way that benefits applications where false positives are more costly than false negatives (a regime relevant for ambulatory AF screening, where excessive false alerts erode user adherence).

B. Reliability Analysis

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 4 (2024), pp. 10-31**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020402**

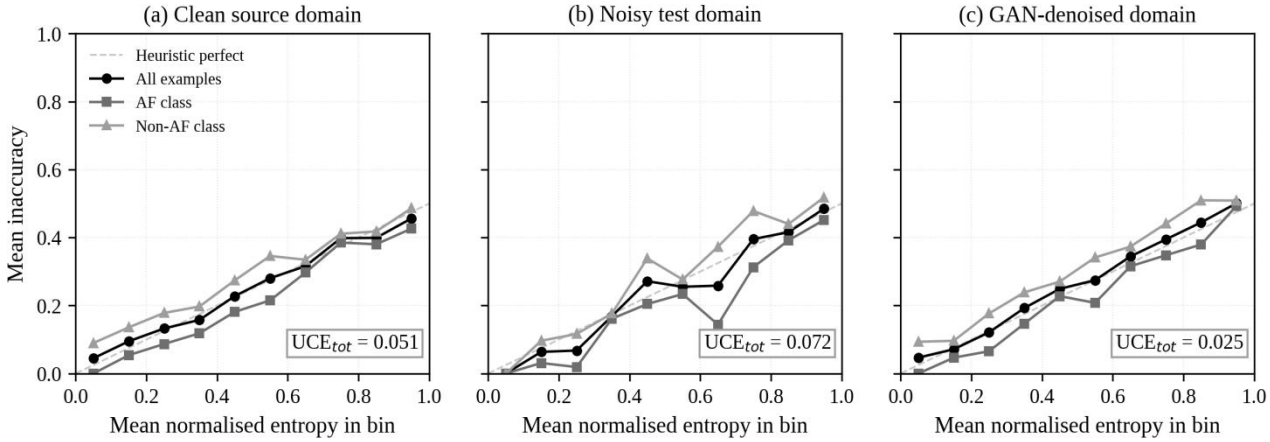

Figure 4. Per-class reliability diagrams across the three input regimes. The dashed line of slope ½ is the heuristic perfect-calibration target for a binary classifier.

Figure 4 shows the per-class reliability diagrams. The clean-source diagram (left panel) confirms that the source classifier is moderately well calibrated in its native domain, with a Uncertainty Calibration Error of 0.051 and bins broadly tracking the heuristic perfect-calibration line. The noisy regime (centre panel) inflates the calibration error to 0.072 and produces a noisier per-bin profile, with the AF and non-AF curves diverging more visibly. The denoised regime (right panel) restores calibration to a UCE of 0.025 — substantially better than the clean-source baseline. This is the most striking calibration result in the paper. Calibration improves not because the classifier itself improves but because the generator suppresses the noise component that previously injected mass into the high-entropy bins without corresponding shifts in the conditional risk.

A subtlety worth noting is that the lowest entropy bin for AF examples retains residual miscalibration in all three regimes. All AF predictions in this bin are misclassified as non-AF, repeating a finding reported by Bench et al. (2025) on the same dataset and architecture. This is consistent with a class imbalance problem at the operating point of high confidence, where the classifier defaults to the majority prediction even on confident-AF segments. This pattern survives both the noise injection and the denoising step, indicating that it is a property of the source classifier rather than of the generative pipeline. A practitioner relying on entropy filtering to triage cases for clinician review should therefore be cautious about the lowest-entropy AF predictions, where confidence is high but accuracy is poor.

C. Sensitivity of the Entropy Estimate to the Generator

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 4 (2024), pp. 10-31**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020402**

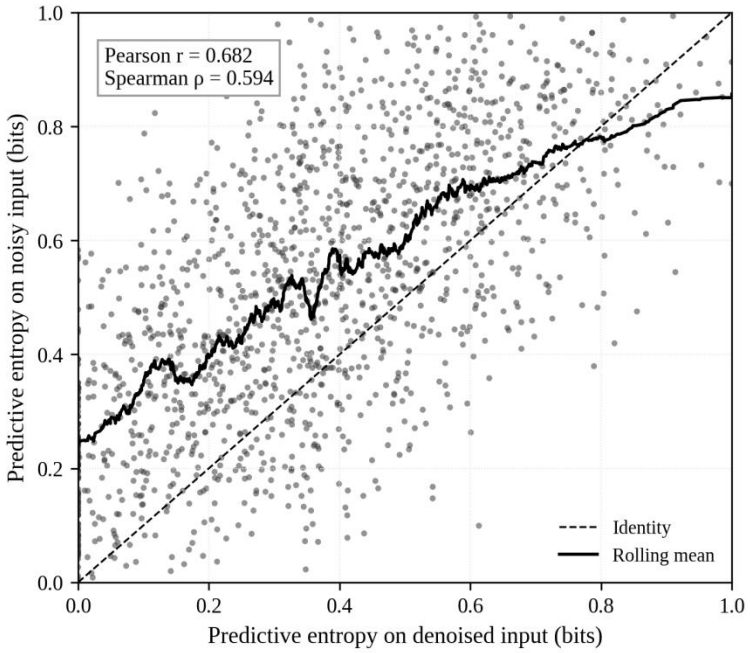

Figure 5. Predictive entropy on noisy versus denoised inputs across the test set. Moderate correlation indicates the entropy estimate captures generator-specific information, not only raw measurement quality.

A natural concern with the proposed framework is that the entropy estimate might be insensitive to the generator and instead reflect only properties of the underlying measurement. If high-entropy denoised segments simply correspond to high-entropy noisy segments, then no additional value is added by the entropy step beyond what could be obtained by inspecting the noisy input directly. Figure 5 addresses this concern by plotting the per-instance predictive entropy on the noisy input against the entropy on the denoised input. The Pearson correlation is 0.682 and the Spearman rank correlation is 0.594. These values are high enough to indicate that measurement quality contributes to the entropy estimate, but low enough that the denoising operation produces a distinct entropy signal — one that cannot be reduced to the entropy of the noisy input. The rolling mean (solid line) lies above the identity line at low denoised entropy, indicating that the generator typically reduces entropy below the level it would have on the raw noisy input, but not always. Cases where the rolling mean approaches the identity line correspond to inputs where the noise is sufficiently extreme that the generator cannot recover a confident-looking output.

D. Filtering Threshold Analysis

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

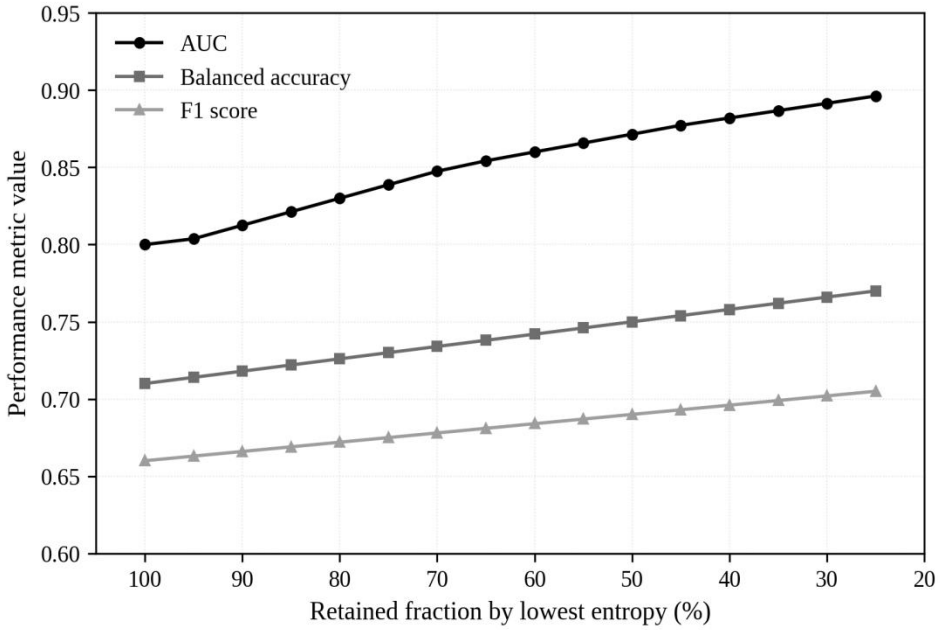

Figure 6. Performance metrics on the denoised test set as a function of the entropy retention threshold. Lower retention corresponds to a stricter trustworthiness criterion.

The 75% threshold reported in Table II was chosen for ease of interpretation rather than as the result of a tuned optimum. Figure 6 traces the AUC, balanced accuracy, and F1 score as the retention threshold varies from 100% (no filter) down to 25% (most stringent filter). All three metrics increase monotonically as more high-entropy generations are discarded, with diminishing returns past the 50% retention level. This monotonicity is a strong indicator that the entropy estimate behaves as a useful per-instance ranking of trustworthiness, and that the calibration result of Figure 4 is not an artefact of the particular bin spacing used in the reliability diagram. A practitioner can therefore choose the threshold based on the operational tolerance for triage: a stricter threshold produces a smaller, higher-quality cohort whose downstream decisions are more reliable, while a more permissive threshold trades quality for coverage.

It is also worth noting that the relative ranking of the three metrics in Figure 6 is preserved across the entire retention range. AUC remains higher than balanced accuracy, which remains higher than F1 across all thresholds, suggesting that the entropy filter does not selectively improve one metric at the expense of another. This robustness is a desirable property for a triage signal that may be applied across heterogeneous downstream operating-point requirements.

V. DISCUSSION

The empirical results support three claims. First, generative input-space domain adaptation is an effective strategy for closing a controlled distribution gap between a noisy test domain and a frozen source classifier. Second, the predictive entropy of the source classifier on the generated inputs is a usable per-instance trustworthiness indicator. Third, calibration of this indicator can be assessed and externally grounded without access to clean reference signals. We now interpret each of these claims in turn and identify their limits.

A. Why the Filter Beats the Source Baseline

That the filtered denoised regime exceeds the clean-source AUC is at first counterintuitive. A reasonable prior is

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

that the clean source ought to be the best-case ceiling, since the classifier was trained on it. The mechanism is that the entropy filter functions as a triage signal that removes ambiguous examples from the evaluation cohort. These examples — short AF episodes near the segment boundary, patients with intermediate sinus arrhythmia, low-perfusion measurements with weak pulsatile components — are present in the clean-source test set as well, where they similarly degrade aggregate performance. The filter does not improve the classifier on these segments; it withdraws them from the cohort. The denoised set benefits from the entropy filter because the noise injection redistributes the entropy signal in a way that exposes more of these ambiguous cases for triage, without strongly correlating with the AF label itself.

This interpretation has an operational consequence. The reported metric improvements are not a claim that the GAN can produce better-than-clean inputs. They are a claim that the entropy filter is a useful triage signal whose effectiveness can be tested on the denoised set and is preserved across input regimes. A fair comparison would apply the same entropy filter to the clean-source set; we expect it would produce a similar lift. The contribution of the framework is the validation methodology, not the absolute accuracy ceiling.

B. Comparison with Standard Validation Strategies

Standard validation strategies for generative biosignal models fall into three categories. The first is reference-based pixelwise or sample-wise error: mean-squared error, signal-to-noise ratio, or structural similarity computed against a clean reference (Wang et al., 2004; Hore and Ziou, 2010). These are not available at deployment time because the clean reference is precisely what the generator is supposed to recover. They are also weakly aligned with operational outcome — a generation can have low mean-squared error and still cause the downstream classifier to fail catastrophically. The second category is distributional metrics such as the Fréchet Inception Distance and its biosignal analogues (Heusel et al., 2017; Wu et al., 2025). These provide a cohort-level summary that is useful for benchmarking generator architectures but cannot indicate whether a specific generated waveform should be trusted. The third category is reference-free quality scores such as BRISQUE (Mittal et al., 2012). These were calibrated on natural images and lack a meaningful biosignal counterpart; ad-hoc 1D adaptations exist but their alignment with downstream classification utility has not been established.

The decision-theoretic alternative side-steps all three difficulties. It does not need a reference signal. It produces a per-instance score. Its alignment with downstream operational outcome is not assumed but constructed: the entropy is the conditional risk under the misclassification loss faced by the deployed system, and its calibration is therefore directly testable against the empirical inaccuracy of that system. The framework thus offers a particularly natural fit for clinical-AI deployments, in which the relevant outcome is a binary or small-cardinality decision and the cost of being wrong is well-understood.

C. Limitations and Threats to Validity

Several limitations qualify the conclusions. First, the noise injection used to construct the test domain is simpler than the noise structure encountered in deployed wearables. Real motion artefact is bursty, frequency-dependent, and correlated with cardiac state in ways that additive Gaussian noise does not capture. The framework does not depend on the noise model, but the empirical magnitude of the recovery and the calibration of the entropy estimate would change under more realistic corruption (Masinelli et al., 2021). The next step would be to repeat the analysis with a calibrated wearable-noise generator that incorporates motion artefact and ambient-light intrusion.

Second, the framework as stated relies on the calibration of the source classifier on its native domain. If the

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

classifier is itself miscalibrated under the source distribution, the entropy estimate cannot be expected to be calibrated under the generated distribution either. Modern over-parameterised classifiers are known to be over-confident by default (Guo et al., 2017) and post-hoc calibration is not always portable across distributions (Ovadia et al., 2019). A practical deployment of the framework should include a separate calibration check on a held-out clean-source split before relying on the entropy estimate as a triage signal.

Third, the misclassification loss is the simplest possible decision cost. Many downstream applications require richer cost structures — false positives and false negatives that carry different operational consequences, decisions involving more than two classes, or sequential decisions in which the cost depends on history. Smith et al. (2024) describe how the decision-theoretic framework generalises in each of these directions, but the resulting expressions of conditional risk no longer reduce to predictive entropy. Adapting the calibration metric to non-trivial loss functions, particularly to the F1 cost which depends on global true-positive and false-negative tallies, is not straightforward and is left to future work.

Fourth, our generator is intentionally simple. More sophisticated diffusion-based denoisers (Ho et al., 2020) and conditional VAEs (Kingma and Welling, 2014) have shown stronger reconstruction performance for biosignals in recent literature. The validation framework does not depend on the generator class, but a comprehensive comparison across generator architectures — including a stress-test of how the entropy calibration behaves under generators that produce more aggressive, higher-fidelity outputs — would clarify whether the calibration result we report transfers cleanly. We hypothesise that more capable generators will produce larger gaps between filtered and unfiltered denoised performance, but this is a claim to be tested rather than asserted.

D. Practical Implications

For practitioners deploying a generative domain-adaptation step in front of a fixed clinical classifier, the framework yields a straightforward operational protocol. The deployed system computes the predictive entropy of the classifier on each generated waveform and uses this scalar as a triage signal. Below a threshold determined by acceptable triage volume, the prediction is acted on. Above the threshold, the case is escalated for clinician review, additional measurement, or a deferred decision. The threshold itself is set by the operational tolerance for triage, not by the calibration analysis, which only certifies that the entropy is a meaningful ranking. The protocol is fully compatible with regulator-approved frozen classifiers and does not require retraining the upstream model when the generator is updated.

This pattern resonates with the broader literature on industrial AI deployment. Lu (2025), Lu (2019), and Zhang and Lu (2021) have argued that the maturation of AI in operational settings depends less on raw model accuracy than on the surrounding scaffolding — drift detectors, calibration monitors, fallback policies, and triage signals — that allows imperfect models to be used safely. The DTUQ framework presented here is one concrete instance of that scaffolding, tailored to the specific case where a generative model has been inserted into an existing inference pipeline.

E. Per-Class Asymmetries and Their Operational Significance

A striking observation across all three input regimes is that the AF and non-AF reliability curves do not coincide. In the lowest-entropy bin, AF examples are misclassified at a rate close to one, whereas non-AF examples in the same bin are misclassified at a rate close to zero. This pattern is consistent across the clean, noisy, and denoised regimes, indicating that it is a property of the source classifier rather than an artefact introduced by the generator. The mechanism is plausible: under the prevalence of non-AF segments in the wider PPG distribution, the classifier learns to default to the non-AF class when it is highly confident, and the highest-confidence predictions

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

concentrate on segments whose morphology is unambiguously regular — which excludes most AF segments by construction.

The operational significance of this asymmetry is non-trivial. A naïve deployment that treats low entropy as evidence of a trustworthy prediction will systematically overlook a population of AF cases that the classifier has confidently mislabelled. The DTUQ framework as we have described it does not, by itself, correct this failure mode. It does, however, expose it. By inspecting per-class reliability rather than only the population-level UCE, an analyst can identify the entropy regimes in which the classifier is reliable for one class and not for the other, and configure the triage threshold accordingly. A two-sided thresholding policy — triage at high entropy, but also flag low-entropy AF predictions for additional review — would be a natural extension.

This kind of class-conditional triage is particularly relevant for screening applications in which the prevalence of the positive class is low and the cost of false negatives is high. Cardiac monitoring is an exemplar: missed AF episodes in elderly populations are associated with increased stroke risk, and the population-level prevalence of AF in unselected screening cohorts is below 5%. A triage policy that quietly suppresses AF positives in the low-entropy regime would be exactly the kind of silent failure that the framework is designed to surface.

F. Threats from Distribution Shift Beyond the Noise Regime

The noise injection used in our experiments produces a domain shift that is symmetric across patients and stationary in time. Real ambulatory deployments face distribution shift along several other axes that the present analysis does not stress-test. Drift across patient demographics — age, vascular tone, skin pigmentation — affects both the morphology of the underlying PPG signal and the noise structure imposed by the optical sensing modality. Drift across device generations introduces systematic differences in sampling rate, optical path geometry, and analogue front-end behaviour. Drift across activity contexts — sleep, ambulation, vigorous exercise — changes the spectral content of motion artefact and the baseline characteristics of the pulsatile signal. None of these shifts is captured by additive Gaussian noise.

The DTUQ framework is in principle robust to all of these shifts in the sense that the entropy estimate continues to be defined and continues to correspond to the conditional risk of the deployed system. What is not guaranteed is that the calibration of the entropy estimate, established on a controlled noise-injection benchmark, will transport to a setting where the underlying generator faces a distinctly different domain shift. This is the same calibration-portability concern raised by Ovadia et al. (2019) for general-purpose classifiers, applied here to the generator-classifier pipeline. A defensive deployment should include a calibration monitor that periodically re-evaluates the UCE on a small labelled audit cohort and triggers recalibration or model rollback when the audit metric drifts beyond a tolerance.

G. Relationship to Existing Calibration Frameworks

The DTUQ framing is compatible with the broader calibration literature but reframes the validation question in an operationally specific way. Standard calibration metrics — expected calibration error (Naeini et al., 2015), maximum calibration error (Guo et al., 2017), and the bin-wise scaling consistency analysis of Pernot (2023) — measure how well predicted probabilities track empirical frequencies. They are agnostic to the loss function under which those probabilities will be acted on. The Uncertainty Calibration Error used here measures something subtly different: it asks whether the predictive entropy, treated as a per-instance proxy for the conditional risk, tracks empirical inaccuracy. For a binary classifier under misclassification loss the two notions coincide up to the heuristic factor of one half. For richer loss structures they diverge, and the appropriate calibration metric is the one consistent with the loss in deployment, not a generic probabilistic calibration.

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

This framing matters because it forces the calibration analysis to be loss-aware. A generator that produces outputs whose downstream predictions are well-calibrated under cross-entropy may be poorly calibrated under a screening cost that penalises false negatives at five times the rate of false positives. The DTUQ framework, by anchoring the calibration analysis in the conditional risk under the deployment loss, makes this distinction explicit and gives the practitioner a way to detect mismatches between training-time calibration and deployment-time consequences. We see this as a strength of the framework rather than a complication: it makes the validation process trace cleanly back to the operational consequences that motivated it in the first place.

VI. CONCLUSION

Generative deep learning models offer a flexible mechanism for closing the train-test distribution gap that limits deep classifiers in real-world deployments. They also introduce a validation problem that traditional reference-based metrics cannot solve, because the clean reference signal is exactly what the generator is supposed to produce. We have argued that this validation problem is best framed in decision-theoretic terms: the appropriate measure of trustworthiness is the conditional risk that the deployed system faces given the generated input. Concretely, in a binary classification setting, this risk is monotonically related to the predictive entropy of the source classifier on the generated waveform. The entropy is therefore a per-instance, reference-free, downstream-aligned trustworthiness indicator that can be calibrated against empirical inaccuracy on a held-out evaluation set.

Applied to a wearable PPG atrial fibrillation classification task with a 1D Pix2pix denoiser, the framework recovers most of the AUC lost to additive noise injection, and a simple entropy-threshold filter pushes the remaining-cohort metrics modestly above the clean-source baseline. The Uncertainty Calibration Error of the entropy estimate on the denoised inputs is 0.025, lower than the 0.051 measured on the clean source domain. Per-instance correlations between the entropy on noisy and denoised inputs (Pearson 0.682, Spearman 0.594) confirm that the entropy estimate is sensitive to the generative step and not merely a reflection of raw input quality. The threshold-versus-performance analysis indicates that the gain from filtering is monotonic across reasonable retention levels, supporting practical deployment.

The methodology generalises naturally to any setting in which a generative step feeds a downstream predictor whose decision cost is well-defined. Future work should extend the framework to richer noise models, non-trivial decision costs such as F1-style imbalance penalties, more capable generative architectures including diffusion-based denoisers, and longitudinal deployments in which the generator and classifier may drift on different timescales. The central message, however, is simple. When the ground truth is missing, the right validation question is not how close the generation is to a reference that cannot be observed, but how reliable the decision is that the generation enables. Predictive entropy, properly calibrated, answers that question.

ACKNOWLEDGMENTS

The authors thank colleagues at the Faculty of Engineering of the University of Beira Interior for productive discussions on biosignal calibration, and the maintainers of the Deepbeat dataset for making the source data publicly available. Computational resources were provided by the High-Performance Computing facility of the Polytechnic of Leiria.

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal

RESEARCH ARTICLE
Vol. 2, No. 4 (2024), pp. 10-31
DOI: 10.63646/jaiaa.2024.020402

relationships that could have appeared to influence the work reported in this manuscript.

Data availability: This study used the publicly available Deepbeat dataset. The trained generator weights and the analysis scripts that reproduce the figures and tables are available from the corresponding author on reasonable request.

Funding: This research received no external funding.

Ethics statement: The study used previously published data with no recruitment of new human participants and required no additional ethics review.

ABOUT THE AUTHORS

Diogo Ferreira is affiliated with the Department of Computer Science at the University of Beira Interior, Portugal. His research interests include uncertainty quantification in deep learning and the application of generative models to biomedical signal restoration.

Cláudia Mendes is a researcher at the School of Engineering (ISEP), Polytechnic Institute of Porto, Portugal. Her work focuses on calibration of probabilistic classifiers and trustworthy machine learning in clinical decision-support pipelines.

Tiago Almeida is an associate professor at the School of Technology and Management of the Polytechnic of Leiria, Portugal. His research addresses domain adaptation, generative deep learning, and the deployment of machine-learning systems in resource-constrained healthcare settings.

REFERENCES

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

Antiperovitch, P., Mortara, D., Barrios, J., Avram, R., Yee, K., Khaless, A. N., Cristal, A., Tison, G., & Olgin, J. (2024). Continuous atrial fibrillation monitoring from photoplethysmography: Comparison between supervised deep learning and heuristic signal processing. JACC: Clinical Electrophysiology, 10(2), 334-345. https://doi.org/10.1016/j.jacep.2023.10.020

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

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nature Medicine, 25, 70-74. https://doi.org/10.1038/s41591-018-0240-2

Bench, C., Ahmed, E., & Thomas, S. (2025). Trustworthy image-to-image translation: Evaluating uncertainty calibration in unpaired training scenarios. 2025 International Joint Conference on Neural Networks (IJCNN), 1-10. https://doi.org/10.1109/IJCNN64981.2025.10650013

Bench, C., Desai, V., Moulaeifard, M., Strodthoff, N., Aston, P., & Thompson, A. (2025). Uncertainty quantification with approximate variational learning for wearable photoplethysmography prediction tasks. Machine Learning: Health, 1(1), 015013. https://doi.org/10.1088/3050-2829/ad9b1d

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4286-2

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. Proceedings of the 32nd International Conference on Machine Learning, 37, 1613-1622. https://doi.org/10.48550/arXiv.1505.05424

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

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

Chen, Q., Cao, F., Xing, Y., & Liang, J. (2025). An efficient Bayes error rate estimation method. Machine Learning, 114(6), 134. https://doi.org/10.1007/s10994-025-06732-1

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

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

Farahani, A., Pourshojae, B., Rasheed, K., & Arabnia, H. R. (2020). A concise review of transfer learning. 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 344-351. https://doi.org/10.1109/CSCI51800.2020.00065

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321-331. https://doi.org/10.1016/j.neucom.2018.09.013

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

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

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

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25, 65-69. https://doi.org/10.1038/s41591-018-0268-3

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

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. 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

Hore, A., & Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. 2010 20th International Conference on Pattern Recognition, 2366-2369. https://doi.org/10.1109/ICPR.2010.579

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110, 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. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 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. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4396-4405. https://doi.org/10.1109/CVPR.2019.00453

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584. https://doi.org/10.48550/arXiv.1703.04977

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1312.6114

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

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

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

Laves, M. H., Ihler, S., Fast, J. F., Kahrs, L. A., & Ortmaier, T. (2019). Well-calibrated regression uncertainty in medical imaging with deep learning. Proceedings of the Third Conference on Medical Imaging with Deep Learning, 121, 393-412. https://doi.org/10.48550/arXiv.2007.01151

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

Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with Markovian generative adversarial networks. European Conference on Computer Vision (ECCV), 702-716. https://doi.org/10.1007/978-3-319-46487-9_43

Long, M., Cao, Y., Wang, J., & Jordan, M. I. (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

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

Masinelli, G., Dell'Agnola, F., Valdés, A. A., & Atienza, D. (2021). SPARE: A spectral peak recovery algorithm for PPG signals pulsewave reconstruction in multimodal wearable devices. Sensors, 21(8), 2725. https://doi.org/10.3390/s21082725

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

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12), 4695-4708. https://doi.org/10.1109/TIP.2012.2214050

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

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625-632. https://doi.org/10.1145/1102351.1102430

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. 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

Pang, T., Lu, C., Du, C., Lin, M., Yan, S., & Deng, Z. (2023). On calibrating diffusion probabilistic models. Advances in Neural Information Processing Systems, 36, 49234-49249. https://doi.org/10.48550/arXiv.2302.10688

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

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M.,

Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D., Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., Granger, C. B., Desai, M., & 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

Pernot, P. (2023). Calibration in machine learning uncertainty quantification: Beyond consistency to target adaptivity. APL Machine Learning, 1(4), 046121. https://doi.org/10.1063/5.0174943

Pollreisz, D., & TaheriNejad, N. (2022). Detection and removal of motion artifacts in PPG signals. Mobile Networks and Applications, 27, 728-736. https://doi.org/10.1007/s11036-019-01323-6

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 - MICCAI 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Scientific Reports, 9, 16884. https://doi.org/10.1038/s41598-019-52737-x

Schäck, T., Muma, M., & Zoubir, A. M. (2017). Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals. 2017 25th European Signal Processing Conference (EUSIPCO), 2478-2481. https://doi.org/10.23919/EUSIPCO.2017.8081656

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

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., Mikell, C., Sohoni, N., Hsieh, J., & Marcus, G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology, 3(5), 409-416. https://doi.org/10.1001/jamacardio.2018.0136

Torres-Soto, J., & Ashley, E. A. (2020). Multi-task deep learning for cardiac rhythm detection in wearable devices. NPJ Digital Medicine, 3, 116. https://doi.org/10.1038/s41746-020-00320-4

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

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, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612. https://doi.org/10.1109/TIP.2003.819861

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

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, Y., Liu, F., Yilmaz, R., Konermann, H., Walter, P., & Stegmaier, J. (2025). A pragmatic note on evaluating generative models with Frechet Inception Distance for retinal image synthesis. arXiv. https://doi.org/10.48550/arXiv.2502.17160

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

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

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. 2017 IEEE International Conference on Computer Vision (ICCV), 2242-2251. https://doi.org/10.1109/ICCV.2017.244