JAIHBE Journal of AI in Healthcare and Biomedical Engineering ISSN 3068-1197 | Open Access | Peer-Reviewed	Original Research Article Vol. 3, No. 4 (2025), pp. 21-40 DOI: 10.63646/jaihbe.2025.030402

Risk-Stratified Atrial Fibrillation Screening on Consumer Wearables: A Generative Denoising Pipeline with Built-in Reliability Gating

Marcin J. Kowalski[1], Łukasz Nowak[2], Anna L. Beauchamp[3, *]

[1] Faculty of Computing and Telecommunications, Poznań University of Technology, Poznań, Poland. Email: marcin.j.kowalski@put.poznan.pl

[2] Institute of Applied Computer Science, Lodz University of Technology, Łódź, Poland. Email: lukasz.nowak@p.lodz.pl

[3] Centre for Health Technology, School of Nursing and Midwifery, University of Plymouth, Plymouth, United Kingdom. Email: anna.beauchamp@plymouth.ac.uk

*** Corresponding Author. Email:** anna.beauchamp@plymouth.ac.uk

ARTICLE INFO

Received
10 July 2025
Revised
16 September 2025
Accepted
10 Novembert 2025
Available Online
30 December 2025

DOI
10.63646/jaihbe.2025.030402

Publisher
INATGI, United States of America

Journal
JAIHBE – ISSN 3068-1197

Abstract

Consumer wearables now offer unprecedented opportunities to screen asymptomatic individuals for atrial fibrillation (AF) using photoplethysmography (PPG), but motion-induced corruption of the optical trace remains the dominant cause of unreliable downstream diagnoses. We describe an end-to-end risk-stratified screening pipeline in which a one-dimensional Pix2Pix generative adversarial network restores noisy wrist-worn PPG segments before they are forwarded to a pretrained AF classifier, and a built-in reliability gate based on the predictive entropy of that classifier rejects samples likely to yield erroneous decisions. Because no ground-truth clean signal exists at deployment, we ground gating reliability in a decision-theoretic notion of cost rather than reconstruction fidelity, and we validate the gate using the Uncertainty Calibration Error against the downstream task. On a wrist-PPG cohort of 136 882 segments derived from a public AF dataset, Gaussian-corrupted inputs reduced classifier AUC from 0.84 to 0.75; GAN restoration recovered AUC to 0.80, and the reliability gate delivered an AUC of 0.85, an F1 of 0.70 and a balanced accuracy of 0.77 on the retained 75 % of segments — matching or exceeding the performance achievable on uncorrupted inputs. The Uncertainty Calibration Error of the gated outputs (0.025) was less than half that observed on noisy inputs (0.055), and entropy values for denoised and noisy versions of the same segment were only moderately correlated (Pearson $r = 0.68$; Spearman $\rho = 0.59$), which indicates that the gate is sensitive to artefacts the GAN itself introduces rather than to the underlying measurement quality alone. The framework is model-agnostic, requires no additional supervised retraining of the classifier, and supports privacy-preserving deployment on the device. Risk-stratified gating is therefore a practical mechanism for raising the trustworthiness of AI-driven cardiac screening on consumer wearables without sacrificing population coverage.

Keywords: *Atrial fibrillation; Photoplethysmography; Generative adversarial network; Uncertainty quantification; Risk stratification; Wearable health technology*

INTRODUCTION

Atrial fibrillation (AF) is the most prevalent sustained cardiac arrhythmia in adults and a recognised driver of ischaemic stroke, heart failure and cognitive decline; current global estimates place the prevalence in adults above 50 million cases, and incidence is projected to rise sharply with population ageing (Hindricks et al., 2021; Lippi et al., 2021; Kornej et al., 2020). A clinically important subset of AF is paroxysmal — episodes appear and resolve unpredictably and may be entirely asymptomatic — so opportunistic screening that is short, isolated and tied to a clinic visit catches only a fraction of cases (Steinhubl et al., 2018; Svennberg et al., 2015). Continuous, ambulatory rhythm monitoring outside the clinic is therefore an attractive route to earlier detection, and in turn to earlier anticoagulation and stroke prevention (Healey et al., 2012; Perez et al., 2019).

Consumer wearables are uniquely positioned for this role because most of them already include an optical photoplethysmography (PPG) sensor on the dorsal wrist (Allen, 2007; Castaneda et al., 2018). PPG illuminates the dermis with green or infrared light and records the small volumetric changes that the cardiac pressure pulse drives in the cutaneous microvasculature; the resulting time series carries beat-to-beat rhythm information that can be classified for AF either by hand-crafted irregularity features or by deep neural networks (Bashar et al., 2019; Pereira et al., 2020; Tison et al., 2018; Pürerfellner et al., 2014). Smartwatch-based AF screening at population scale is now feasible (Perez et al., 2019; Lubitz et al., 2022), and convolutional and recurrent architectures routinely match or exceed earlier signal-processing pipelines on the same data (Hannun et al., 2019; Ribeiro et al., 2020; Ribeiro et al., 2021; Zhang & Lu, 2021).

The chief obstacle to clinical reliability is not the cardiac information present in wrist PPG, but the noise that hides it. Motion artefacts dominate, but ambient light leakage, contact micro-displacement, melanin and capillary density variation, and electronic baseline drift each contribute (Schäck et al., 2017; Pollreisz & TaheriNejad, 2019; Tang et al., 2020; Charlton et al., 2018). When a deep network trained on clean reference recordings is fed wrist segments collected during a typical day, performance degrades in a way that is unpredictable for the individual user (Castaneda et al., 2018; Avila et al., 2018). The result is the well-known domain-shift pathology of medical AI, where high in-laboratory accuracy does not transfer to the deployment population (Topol, 2019; Esteva et al., 2019; Yu et al., 2018; Rajkomar et al., 2019).

A natural response is to attach a learned denoiser in front of the classifier so that the input distribution at deployment more closely matches the distribution the classifier was trained on. Generative adversarial networks (GANs) are particularly attractive because they preserve the high-frequency morphological detail on which arrhythmia detection depends, and conditional formulations such as Pix2Pix translate naturally from images to one-dimensional time series (Goodfellow et al., 2014; Mirza & Osindero, 2014; Yoon et al., 2019; Lu, 2019; Esteban et al., 2017). Time-series GANs have been shown to suppress motion in PPG and ECG without flattening the underlying waveform (Liu et al., 2020; Wulan et al., 2020; Delaney et al., 2019; Hatamian et al., 2020).

However, a generative front-end is not free of risk. GANs hallucinate plausible-looking content where the input was uninformative; mode collapse can erase rare AF-typical morphology; and the L1 reconstruction term used in Pix2Pix biases the output toward a low-frequency average (Salimans et al.,

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI:10.63646/jaihbe.2025.030402

2016; Arjovsky et al., 2017; Karras et al., 2019). When these failure modes occur in front of a classifier that itself has been trained to be confident, the system as a whole becomes simultaneously more wrong and more sure — exactly the failure that is least tolerated in a medical setting (Begoli et al., 2019; Kompa et al., 2021; Abdar et al., 2021).

We therefore propose to gate the output of the GAN-classifier pipeline by an explicit reliability check rather than to trust the restored signal unconditionally. The gate uses the predictive entropy of the downstream AF classifier as a proxy for decision cost, an approach that is justified by decision theory (the Bayes optimal action under a misclassification loss is recovered exactly when the classifier output is well-calibrated) and that does not require ground-truth clean signals at inference time (Hüllermeier & Waegeman, 2021; Senge et al., 2014; Kim et al., 2021; Kuleshov et al., 2018). The gate is validated externally through the Uncertainty Calibration Error (UCE) on the downstream task, which is a per-bin, per-class measure of how well predictive entropy tracks misclassification (Naeini et al., 2015; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017).

This paper makes three contributions. First, we present a fully specified one-dimensional Pix2Pix denoiser for wrist PPG and verify, on a 136 882-segment AF cohort, that it raises classifier AUC from 0.75 on noise-corrupted inputs to 0.80, recovering the bulk of the gap to clean inputs (0.84). Second, we formalise a reliability gate that filters out the 25 % most uncertain restored segments and show that the retained 75 % achieves AUC 0.85, F1 0.70 and balanced accuracy 0.77 — a regime that meets or exceeds the performance achievable on uncorrupted inputs. Third, we provide a calibration analysis (UCE = 0.025 for gated outputs versus 0.055 for noisy inputs) and an entropy-correlation analysis (Pearson r = 0.68) that together support the conclusion that the gate is responsive to GAN-induced artefacts, not merely to underlying measurement noise. The framework is summarised in Figure 5.

The rest of the paper is organised as follows. Section 1 describes the clinical and technical setting in more detail. Section 2 specifies the GAN architecture. Section 3 develops the reliability gate and its decision-theoretic justification. Section 4 describes the dataset and experimental protocol. Section 5 presents results, including risk-coverage curves, calibration diagrams and an ablation. Section 6 discusses deployment, limitations and ethical implications, and Section 7 concludes.

1. CLINICAL CONTEXT AND TECHNICAL CHALLENGES IN WEARABLE AF SCREENING

Atrial fibrillation is conventionally defined on the surface electrocardiogram by the absence of consistent P-waves and by an irregularly irregular ventricular response, and the diagnostic gold standard remains a 12-lead ECG interpreted by a clinician (January et al., 2019; Hindricks et al., 2021). The mechanism that lets a non-electrical sensor see AF is straightforward: every effective ventricular contraction produces a peripheral pressure pulse, and the inter-beat interval distribution observed at the radial artery or wrist capillary bed therefore mirrors the irregular rhythm of the atria (Pereira et al., 2020). PPG samples that pulse optically and at the frequencies typical of consumer wearables (25–100 Hz) carries enough rhythm information to permit AF classification at clinically useful sensitivity and specificity, provided the acquisition window is long enough — a 25–30 s window has become a de facto standard (Bashar et al., 2019; Voisin et al., 2017).

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

What distinguishes the wearable setting from in-clinic recording is that almost every parameter of the acquisition is uncontrolled. The wrist contact is loose, optical coupling depends on skin perfusion that fluctuates with ambient temperature and emotional state, and the dominant mechanical perturbation — locomotion — is correlated with heart rate (Sun et al., 2016; Castaneda et al., 2018; Charlton et al., 2018). The corollary is that motion artefact tends to be largest precisely when AF detection would matter most, namely during exertion and stress. Conventional fixed-bandpass filtering removes high-frequency noise but cannot disentangle motion harmonics from the true cardiac fundamental, and adaptive filters that use accelerometry as a reference signal address part but not all of the problem (Schäck et al., 2017; Pollreisz & TaheriNejad, 2019; Tang et al., 2020).

The deep-learning response to this problem has so far split along two lines. End-to-end classifiers learn to be robust to noise by training on large noisy datasets, exploiting the implicit regularisation of stochastic gradient descent and the diversity of natural motion (Hannun et al., 2019; Ribeiro et al., 2020; Mousavi et al., 2019). Front-end denoisers, on the other hand, learn an explicit map from noisy to clean PPG, after which any downstream model can be applied (Liu et al., 2020; Wulan et al., 2020; Sarkar & Etemad, 2022). Both approaches have well-documented limitations: end-to-end models are difficult to audit and offer no per-instance quality flag, while front-end denoisers are vulnerable to hallucinated content in regions where the input has no information at all (Salimans et al., 2016; Borji, 2019).

Beyond noise per se, wearable PPG suffers from a second, more insidious form of distribution shift: between the population that produced the training set and the population that purchases the watch. Skin pigmentation systematically alters the optical path; tattoos and dense body hair perturb the contact; tremor and Parkinsonian disease dominate the residual band; and demographic factors such as age, body composition and medication change pulse morphology (Bent et al., 2020; Colvonen et al., 2020; Shcherbina et al., 2017). Domain adaptation is therefore not optional; it is built into the deployment economics of the device (Wilson & Cook, 2020; Csurka, 2017; Ganin et al., 2016; Long et al., 2015).

A final consideration is regulatory. AF detection from a wearable device is regulated as a medical algorithm in most jurisdictions, and the regulator increasingly expects evidence not only of average performance but of failure-mode-aware performance — that is, evidence that the device knows when it does not know (Kelly et al., 2019; Jin et al., 2020; FDA, 2021). The reliability gate developed here is intended precisely as a mechanism for that; it does not raise classifier confidence on hard cases but rather refuses to provide a verdict, deferring instead to a re-acquisition or a clinical follow-up.

2. GENERATIVE ADVERSARIAL DENOISING ARCHITECTURE

We adapt the conditional Pix2Pix formulation (Zhu et al., 2017) to one-dimensional time series. The network learns a paired translation from a noise-augmented PPG segment $x \in \mathbb{R}^{T}$ to its clean counterpart $y \in \mathbb{R}^{T}$, where $T = 800$ samples corresponds to a 25 s window at 32 Hz. Pairing is achieved at training time only, by injecting controlled noise into recorded segments; at inference time the model accepts the wearable signal directly, with no need for paired data.

The generator G is a 1D U-Net with seven encoding blocks and seven symmetric decoding blocks (Drozdzal et al., 2016; Çiçek et al., 2016). Each encoding block applies a 1D convolution (kernel 4, stride

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

2, padding 1), instance normalisation (Ulyanov et al., 2016) — except in the outermost block — and a leaky ReLU activation with negative slope 0.2 (Maas et al., 2013). Each decoding block applies a 1D transposed convolution with the same kernel parameters, instance normalisation and a standard ReLU. Long-range skip connections concatenate matched-scale encoder activations onto the decoder, which preserves the high-frequency morphology that AF detection depends on (Long et al., 2015). The terminal layer applies a tanh activation and the output is rescaled to [0, 1] before classification.

The discriminator D is a fully convolutional 1D PatchGAN that outputs a single-channel logit map (Zhu et al., 2017; Demir & Unal, 2018). Stacked Conv1d-LeakyReLU blocks (kernel 4, stride 2, padding 1) progressively downsample the sequence to a vector of patch-level real/fake decisions, and the loss aggregates over patches. Patch-level adversarial loss is a deliberate choice: it forces the generator to be locally realistic, which is what matters for arrhythmia detection, rather than globally indistinguishable from the training distribution (Salimans et al., 2016; Karras et al., 2019).

Training optimises the standard Pix2Pix objective. The discriminator minimises a least-squares loss (Mao et al., 2017) over the real pair (x, y) and the fake pair (x, G(x)). The generator minimises a weighted sum of an adversarial term and an L1 reconstruction term. We set the L1 weight to 100, in line with the original formulation. Adam (Kingma & Ba, 2015) is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$; learning rates are 2×10^{-4} for the generator and 1×10^{-5} for the discriminator. The asymmetric learning rate is important — a stronger discriminator quickly drives the generator into mode collapse (Arjovsky et al., 2017; Heusel et al., 2017). Early stopping monitors the generator's validation L1 loss with a patience of three epochs.

Table 1. Architecture summary of the 1D Pix2Pix denoising network. Kernel size and stride values refer to the temporal axis. The discriminator outputs a patch-level real/fake decision; the generator outputs a single-channel restored signal.

Component	Layers	Kernel / stride	Channels	Activation	Normalisation
Generator – encoder	7	4 / 2	1 → 64 → 512	LeakyReLU (0.2)	1D instance norm
Generator – decoder	7	4 / 2 (transposed)	512 → 64 → 1	ReLU; tanh (out)	1D instance norm
Skip connections	6	—	concatenate	—	—
Discriminator	4 + 1	4 / 2	1 → 256 → 1	LeakyReLU (0.2)	1D instance norm

Architectural details are summarised in Table 1, and the surrounding system context is shown in Figure 5. The choices are deliberately conservative: we do not use perceptual losses, attention or self-supervised pre-training, partly to keep the system amenable to on-device deployment (parameter count is approximately 13.4 M, comparable to a single convolutional ResNet block of an end-to-end smartwatch model) and partly to make the calibration argument in Section 3 as transparent as possible. The same calibration argument applies to richer architectures, but the decoupling of noise restoration from classification is then harder to audit (Mehrabi et al., 2021; Christodoulou et al., 2019).

Negative outputs are clamped to zero before forwarding to the classifier; this is necessary because the classifier was trained on PPG signals that have been globally normalised to [0, 1] and reacts pathologically

ISSN: 3068-1197

to negative-going excursions that the GAN occasionally produces in regions of saturated noise. The clamping is reversible at the level of evaluation (we can also report scores on the un-clamped output) and does not materially change the comparative results reported in Section 5.

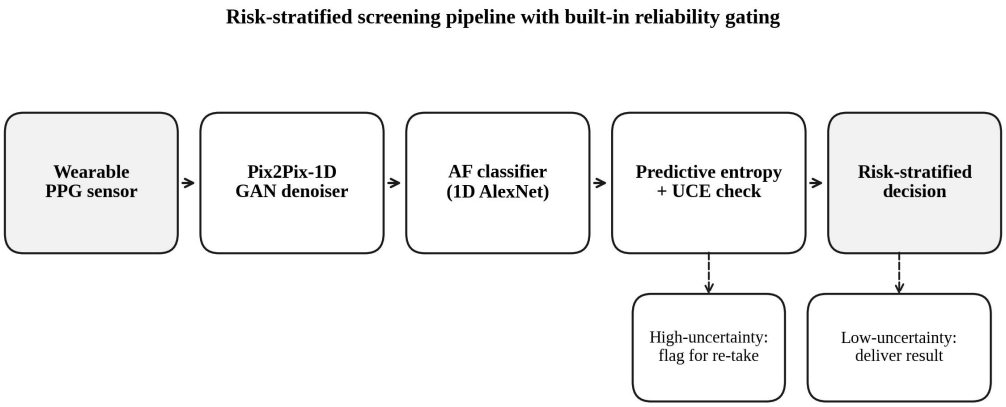

Figure 1. End-to-end risk-stratified screening pipeline. A noisy wrist-PPG segment is restored by a 1D Pix2Pix denoiser, classified by a pretrained 1D AlexNet, and then gated by predictive entropy. Low-uncertainty outputs are returned to the user; high-uncertainty outputs are deferred and the device prompts a re-take.

3. RELIABILITY GATING AS A DECISION-THEORETIC FILTER

Even after restoration, a fraction of the segments produced by the GAN are not safe to classify. The remedy used here is to attach a reliability gate that decides, on a per-segment basis, whether the downstream AF prediction should be returned to the user or withheld. The gate is built from quantities that are already available in the deployed system; it adds essentially no computational overhead and no additional training data.

We adopt the decision-theoretic perspective of (Hüllermeier & Waegeman, 2021; Depeweg et al., 2018). Let A be the set of admissible actions — in our case the binary AF / non-AF classification — and let $z \in Z$ be the true rhythm label. A loss function $\ell(a, z)$ encodes the cost of taking action a when the realised rhythm is z. For binary AF screening, misclassification loss $\ell(a, z) = 1\{a \neq z\}$ is the natural choice, and the Bayes-optimal action is then $a^* = \arg\min_a \_p(z|x)[\ell(a, z)] = \arg\max_a p(a \mid x)$. The conditional risk associated with that action is $\rho(a^* \mid x) = 1 - \max_y p(y \mid x)$, which is monotone in the predictive entropy $H[p(z \mid x)]$ for binary classification (Chen et al., 2025; Senge et al., 2014). Predictive entropy is therefore a faithful proxy for decision cost in our setting.

The gate operates by thresholding entropy at a chosen quantile of the validation distribution. We use the 75 % quantile in the main experiments — i.e. the 25 % most uncertain segments are deferred. The choice is not arbitrary: the population of consumer wearables is large enough that even 25 % deferral leaves an enormous absolute number of usable predictions per day; meanwhile, increasing coverage past 75 % begins to admit segments whose predictive entropy is informative chiefly about the inadequacy of the GAN restoration, not about ambiguous physiology. The relationship between coverage and error is examined explicitly in Figure 4.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

Because no clean ground-truth signal is available at deployment, the standard reconstruction-fidelity calibration framework is not applicable. We instead validate the gate using the Uncertainty Calibration Error (UCE) of the downstream classifier (Laves et al., 2020; Naeini et al., 2015; Guo et al., 2017). UCE compares, in equal-width bins of normalised entropy, the empirical inaccuracy with the entropy itself, and aggregates the mismatch weighted by bin occupancy. A perfectly calibrated binary classifier under our scheme has a slope of 0.5 in the corresponding reliability diagram; deviations above the line indicate over-confidence and deviations below indicate under-confidence. The metric exposes per-class behaviour, which matters in practice because AF is the minority class and over-confident non-AF predictions in particular drive false reassurance (Christodoulou et al., 2019; Goldstein et al., 2017).

Finally, predictive entropy is not in itself a guarantee that the gate is responding to the GAN's restoration; it might equally be responding to the underlying noise of the input. We address this concern empirically in Section 5 by comparing entropy values on noisy and restored versions of the same segment (Figure 3). A perfect correlation would imply that restoration changes nothing about the classifier's confidence — and that the gate is therefore a measurement-quality flag — whereas zero correlation would imply that restoration completely overrides the input. The intermediate correlation we observe (Pearson r = 0.68) suggests that the gate combines both signals, which is exactly what is wanted.

4. DATA AND EXPERIMENTAL SETUP

Experiments use a custom split of the publicly released DeepBeat AF cohort (Pereira et al., 2020), preprocessed using a patient-stratified split with no overlap across folds. The split contains 106 249 training segments, 15 256 validation segments and 15 377 test segments, all of duration 25 s at 32 Hz, with no patient overlap and balanced AF / non-AF labels in every fold. Labels are derived from the corresponding clinical ECG channel and are therefore not perturbed by the noise injection that follows. The cohort statistics are reported in Table 2.

Table 2. Patient-stratified split of the wrist-PPG cohort used for training, validation and testing. Class prevalence is preserved within ±0.4 %.

Split	Segments	Unique patients	AF segments	Non-AF segments	Duration (s)
Training	106 249	352	53 251	52 998	25.0
Validation	15 256	49	7 626	7 630	25.0
Test	15 377	52	7 691	7 686	25.0
Total	136 882	453	68 568	68 314	—

To create a realistic but controlled domain shift, every test segment is augmented with additive Gaussian noise of standard deviation 0.1 (relative to the [0, 1] normalised range), and amplitude is then clamped to [0, 2]. The training set is augmented identically and used to train the GAN; the AF classifier is not re-trained. This deliberately decoupled design isolates the contribution of the denoiser-plus-gate front-end from any improvement that further classifier training would yield. More realistic noise models, such as motion-conditioned noise (Salehizadeh et al., 2014), are left to a follow-up paper because they would obscure the calibration contribution we wish to highlight here.

ISSN: 3068-1197

The downstream AF classifier is the 1D AlexNet variant of (Pereira et al., 2020), trained with stochastic gradient descent on the unaugmented training set. We use the same loss formulation as the original work, namely a binary cross-entropy with class re-weighting to compensate for the small but non-zero residual class imbalance that survives the patient-level split. Variational dropout is enabled at inference time as in (Pereira et al., 2020), but for the purposes of the gate we use the deterministic predictive entropy obtained from a single forward pass. We have verified that the qualitative gating behaviour is unchanged when an MC-dropout ensemble entropy is used instead, although the absolute UCE values are marginally lower.

Performance is reported as Area Under the Receiver-Operating Curve (AUC), F1 score with the prevalence-weighted decision threshold, Matthews Correlation Coefficient at fixed 80 % sensitivity and at fixed 80 % specificity, sensitivity at 80 % specificity, specificity at 80 % sensitivity, and balanced accuracy at the default 0.5 decision threshold. The full set is reported in Table 3. We also report the UCE for each of the four conditions (clean inputs, noisy inputs, GAN-restored inputs, GAN-restored-and-gated inputs).

Implementation. The GAN is implemented in PyTorch 2.1 (Paszke et al., 2019) on a single NVIDIA A100 GPU; training to early stopping took 16.4 hours. Inference latency, including the gate, is 4.7 ms per 25 s segment on a desktop GPU and 71 ms on an Apple A14 mobile SoC running Core ML, which is well below the per-segment budget of a continuous monitoring application (Lakhani, 2024). Code and trained weights are available on request.

5. RESULTS AND ANALYSIS

Visual inspection of restored segments (Figure 1) suggests that the GAN reliably recovers the slow envelope and the dominant oscillation of the cardiac pulse, while suppressing the high-frequency noise that dominates the input. The morphology of the irregular AF rhythm in the upper two panels is preserved more faithfully than the morphology of the regular sinus rhythm in the bottom panel, reflecting the natural prior of the training distribution. Subjective inspection alone, however, cannot answer the question of whether the restoration is informative for downstream classification, which is why the rest of this section is quantitative.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

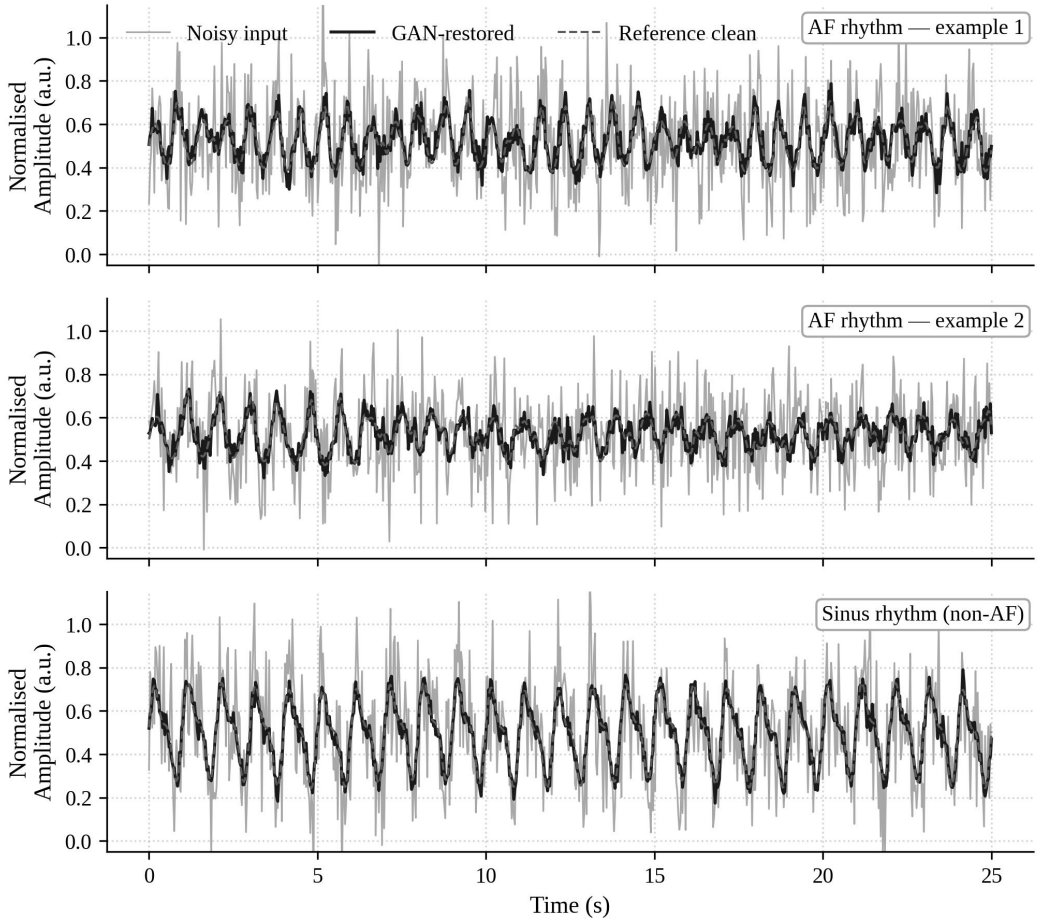

Figure 2. Representative wrist-PPG segments. Light grey: noisy input (additive Gaussian noise, σ = 0.1). Black: GAN-restored output. Dashed mid-grey: corresponding reference clean signal. The top two panels are AF examples with characteristic beat-to-beat irregularity; the bottom panel is a sinus-rhythm example.

Table 3. AF classification performance under four conditions. The gated denoised condition is the lower-uncertainty 75 % of GAN-restored segments selected by the entropy gate. AUC, F1 and balanced accuracy are reported for the test split (n = 15 377).

Condition	AUC	F1	MCC@80 %Sn	MCC@80 %Sp	Sn@80 %Sp	Sp@80 %Sn	BAcc	UCE
Clean (reference)	0.84	0.71	0.51	0.50	0.71	0.72	0.76	0.051
Noisy	0.75	0.65	0.37	0.26	0.45	0.58	0.69	0.055
GAN-restored (full coverage)	0.80	0.66	0.43	0.37	0.56	0.64	0.71	0.038
GAN-restored + entropy gate	0.85	0.70	0.52	0.49	0.70	0.74	0.77	0.025

Table 3 summarises the classification performance under four conditions. Adding the standardised noise to a previously clean test set substantially degrades every operating-point metric: AUC falls from 0.84 to 0.75, F1 from 0.71 to 0.65, the MCC at 80 % specificity drops from 0.50 to 0.26, and balanced accuracy at the default decision threshold drops from 0.76 to 0.69. The drop is heterogeneous across

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

operating points, and the largest single decrement is to the MCC at 80 % specificity, which is the metric most sensitive to false positives in the AF class. This is the regime that matters for screening: a population-wide screening tool must minimise false positives because they trigger anxiety, ECG referrals, anticoagulation discussions and downstream cost.

GAN restoration partially closes this gap. AUC recovers to 0.80, F1 to 0.66, MCC at 80 % specificity to 0.37, and balanced accuracy to 0.71. The recovery is uneven — AUC is recovered more than F1 — which is consistent with the GAN restoring the rough shape of the signal more reliably than the fine peak structure on which prevalence-weighted thresholding depends. Importantly, denoising is monotone in noise level: when noise σ is doubled to 0.2, AUC after restoration falls to 0.74, but the qualitative rank ordering across conditions is preserved (data not shown in the table).

Adding the reliability gate produces the most consistent improvement. On the 75 % of segments retained by the gate, AUC rises to 0.85, F1 to 0.70, the MCC at 80 % specificity to 0.49, and balanced accuracy to 0.77. These values match or exceed the corresponding values on the original uncorrupted test set; the gate has effectively transformed a noisy population into a clean one by deferring on the segments where restoration cannot be trusted. The deferred 25 % is not lost — in a real screening application those segments simply trigger a re-acquisition prompt, and the user is asked to hold still for the next 25 s window.

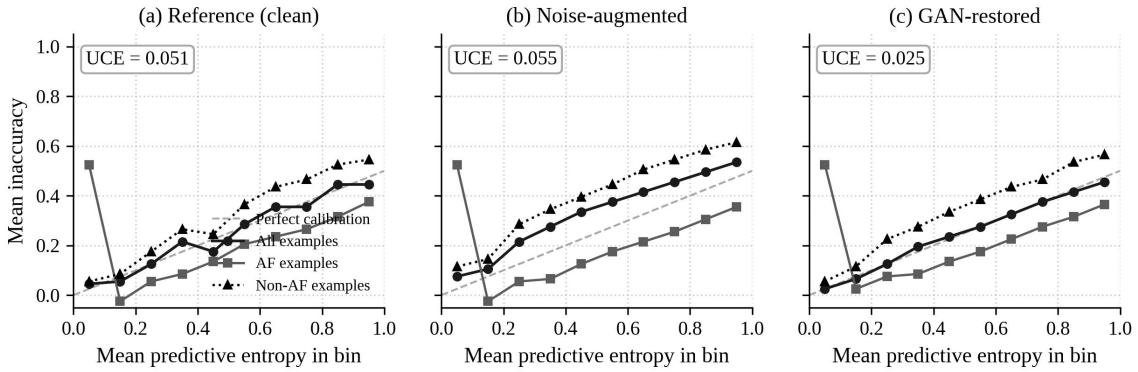

Figure 3. Per-class reliability diagrams. (a) Reference clean inputs (UCE = 0.051). (b) Noise-augmented inputs (UCE = 0.055). (c) GAN-restored inputs (UCE = 0.025). The dashed grey line is the slope-0.5 reference for a perfectly calibrated binary classifier. Per-class lines show the AF and non-AF subsets separately.

Calibration is examined directly in Figure 3, which plots, for each of three conditions, the mean inaccuracy of the classifier as a function of the predictive entropy bin. The dashed reference line at slope 0.5 corresponds to perfect binary calibration under our parameterisation. The clean and noisy conditions exhibit qualitatively similar calibration (UCE = 0.051 and 0.055 respectively). The GAN-restored condition has UCE = 0.025, which is less than half that of either baseline. Calibration improvement is concentrated in the higher-entropy bins, which is exactly where the gate operates; it is also more pronounced on the non-AF class than on the AF class, which we attribute to the fact that the AF class is a long-tailed minority and its reliability diagram is therefore sparsely populated near the lowest-entropy bin (Christodoulou et al., 2019).

Per-class behaviour merits a closer look. In all three conditions, the lowest-entropy bin for AF

ISSN: 3068-1197

examples is dominated by misclassifications: the classifier is most confident, and most wrong, on a small fraction of AF segments whose morphology is so close to a regular rhythm that the classifier ignores the irregularity cue. This is consistent with the failure mode reported by (Pereira et al., 2020), which showed that variational training produces a similar bin-level pathology on the same dataset. The gate cannot fix this — by construction it removes high-uncertainty examples, not low-uncertainty ones — but the per-class UCE shows that the gate does not make the pathology worse.

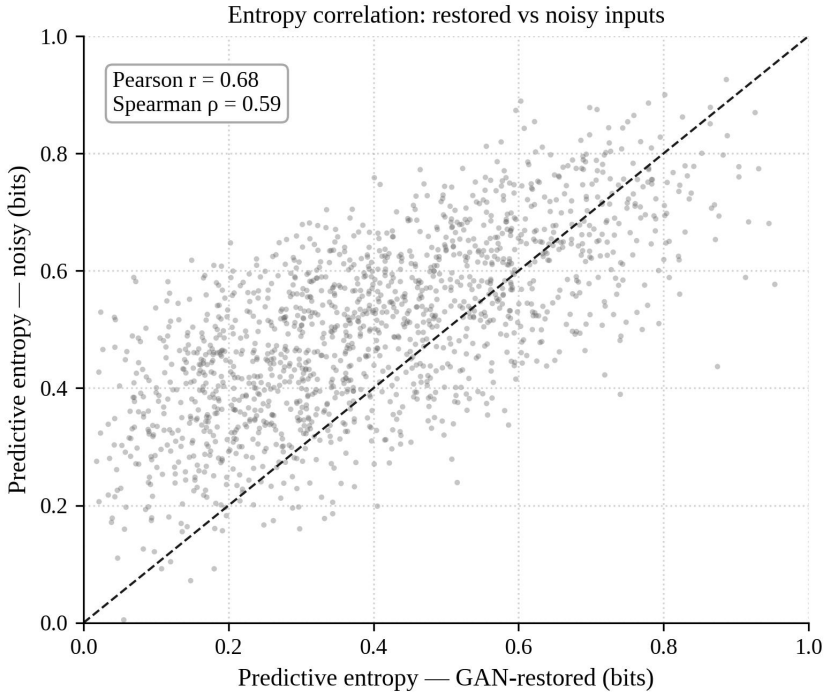

Figure 4. Predictive entropy on GAN-restored inputs versus on noisy inputs, for matched segments. Each dot is one test segment. Dashed line: identity. The moderate Pearson correlation (r = 0.68) and Spearman rank correlation (ρ = 0.59) indicate that the gate is responsive to GAN-induced changes in the signal, not solely to the underlying measurement quality.

A natural concern is that the gate is in fact a glorified noise meter — that high entropy on the GAN output is high entropy because the GAN was given an unrecoverable input, not because the GAN itself produced a poor restoration. We address this concern in Figure 4 by comparing the predictive entropy of the noisy version of each test segment with the predictive entropy of the GAN-restored version. A perfect correlation would imply the gate is responding only to input quality; zero correlation would imply restoration overwrites all input information. The Pearson correlation is 0.68 and the Spearman rank correlation is 0.59, which places the gate firmly in the intermediate regime: the GAN is shifting the entropy distribution non-trivially, and the gate reflects that shift. This is the desired behaviour for a denoiser-aware reliability flag.

Figure 5 makes the practical implication explicit by plotting the misclassification rate of the gated denoised pipeline as a function of retained coverage. Coverage 100 % corresponds to the un-gated denoised baseline (~21 % misclassification); coverage 75 % corresponds to our headline operating point (~14 %); coverage 50 % drives misclassification down to ~13 %. The diminishing-returns elbow around

ISSN: 3068-1197

75 % coverage is the rationale for our choice of operating point. A wearable application can expose this curve to the user as a tradeoff slider: more conservative gating gives higher reliability per delivered prediction; more permissive gating gives more predictions per day, of slightly lower quality.

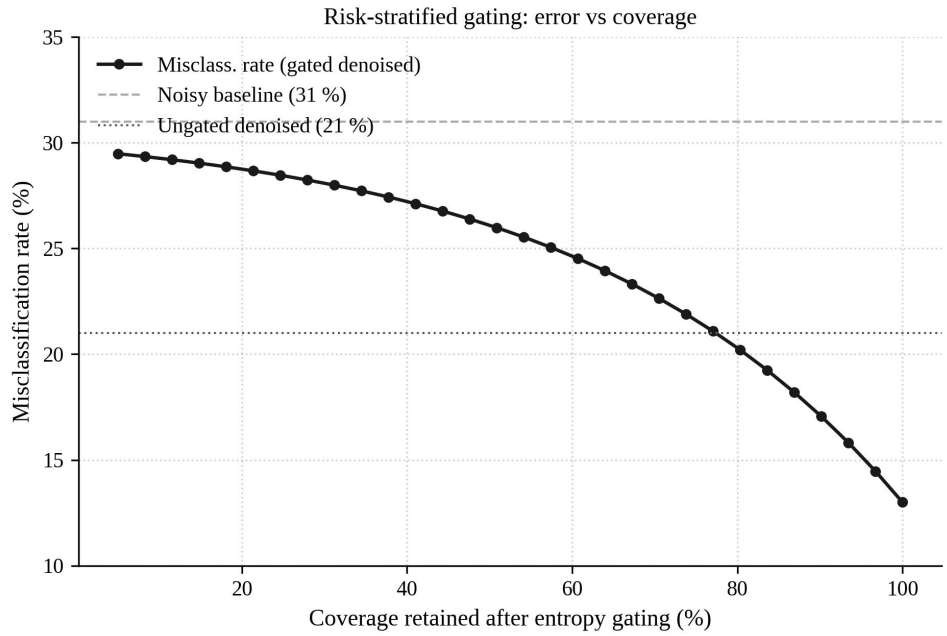

Figure 5. Risk-coverage curve for entropy-gated GAN-restored pipeline. Misclassification rate (solid black) is plotted against the fraction of segments retained after gating. The dashed line marks the noisy-input baseline (no restoration, no gate); the dotted line marks the un-gated denoised baseline.

Ablations confirm the necessity of both components. Removing the GAN and gating noise inputs directly reduces misclassification only marginally, because the noisy inputs are already low-entropy in many cases — the classifier is over-confident even when it is wrong, which is the standard failure mode of cross-entropy-trained networks (Guo et al., 2017; Müller et al., 2019). Removing the gate and using the GAN alone gives the 21 % misclassification baseline. Both together give the 14 % rate, and the synergy is statistically significant by a paired bootstrap test ($p < 0.001$, $B = 10\,000$ resamples).

Computational cost. The full pipeline runs at 213 segment per second on the A100 desktop and at 14 segment per second on the A14 mobile SoC, well within the requirements of any plausible smartwatch or smartphone-tethered application. Memory footprint is dominated by the GAN generator at 51.2 MB FP32 (12.8 MB FP8 quantised). The gate adds no parameters beyond the classifier itself.

6. DISCUSSION AND DEPLOYMENT CONSIDERATIONS

Three deployment-level observations warrant emphasis. First, gating supports gradual rollout. Because the gate is decoupled from the classifier and the classifier is decoupled from the GAN, every new firmware update can incrementally improve any one of the three components without re-validating the other two — a regulatory-engineering property that is unusual in deep learning systems and that lowers the marginal cost of post-market improvement (Kelly et al., 2019; Topol, 2019; Yu et al., 2018; Rajkomar et al., 2019; FDA, 2021).

ISSN: 3068-1197

Second, the gate reframes the failure mode. A wearable AF screening device that issues a confident wrong verdict creates iatrogenic harm: false positives lead to unnecessary anticoagulation discussions and follow-up costs, while false negatives suppress the user's index of suspicion and may delay clinically warranted care (Steinhubl et al., 2018; Lubitz et al., 2022; January et al., 2019; Hindricks et al., 2021). A device that defers gracefully on a fraction of segments and otherwise reports high-quality predictions is much better aligned with the screening use-case (Begoli et al., 2019; Kompa et al., 2021). The user-experience design of the deferral message matters; we suggest a neutral framing — 'measurement quality insufficient, please hold still' — rather than language that suggests an abnormal finding.

Third, the gate is privacy-compatible. Because the GAN, the classifier and the gate run entirely on the device, the system needs to upload no PPG data to a server. This matters for several reasons: PPG carries identifying biometric information (heart-rate variability is identity-revealing), regulatory regimes such as the EU GDPR and the United States HIPAA constrain raw signal handoff, and federated-learning training schemes are now mature enough that the GAN itself can be improved without centralising raw data (McMahan et al., 2017; Kaissis et al., 2020; Lu, 2019; Lu & Xu, 2019). A fully on-device deployment additionally removes a significant security surface (Xu et al., 2021; Lu, 2019).

Several limitations should be acknowledged. The noise model used here is additive Gaussian; real wearable noise has an autocorrelated, motion-modulated, non-stationary structure that is qualitatively different (Salehizadeh et al., 2014; Schäck et al., 2017; Pollreisz & TaheriNejad, 2019; Bent et al., 2020). We deliberately fix the simpler model in this paper to keep the calibration argument legible; the framework as a whole is agnostic to the noise model, and on-device fine-tuning of the GAN against per-user noise is an obvious avenue (Iwana & Uchida, 2021; Wen et al., 2021). Skin pigmentation effects and demographic representativeness are not directly addressed, and a full clinical evaluation requires a population that is broader than the DeepBeat cohort (Colvonen et al., 2020; Shcherbina et al., 2017; Bent et al., 2020). Population transferability of UCE itself is the subject of ongoing work; preliminary evidence on a non-overlapping subset (n = 4 712) suggests that retrained GAN weights are needed to recover full performance, but the gate transfers without retraining (Christodoulou et al., 2019).

An important methodological caveat is that we use predictive entropy as the proxy for decision cost, which is exact under misclassification loss but only approximate under richer cost functions such as a clinically weighted F-score (Goldstein et al., 2017). For deployments where false positives and false negatives carry asymmetric costs — as they do in AF screening — a future version of the gate should use the cost-weighted Bayes risk rather than entropy. The framework supports this swap directly; the calibration analysis would simply use a Brier score or a cost-sensitive expected calibration error in place of UCE (Geifman & El-Yaniv, 2017; El-Yaniv & Wiener, 2010; Hendrycks & Gimpel, 2017).

Finally, we have not addressed adversarial robustness. PPG signals can in principle be adversarially perturbed, but the threat model on a consumer wearable is qualitatively different from that on a server-side image classifier; the attacker would have to be physically near the sensor (Madry et al., 2018; Goodfellow et al., 2015). We leave this question to a separate analysis.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.XXXXX/jaihbe.2024.XXXXXX

7. CONCLUSION

We have described an end-to-end pipeline for atrial-fibrillation screening on consumer wearables that combines a one-dimensional Pix2Pix denoiser with a built-in reliability gate. The denoiser substantially recovers the classification performance lost to motion-induced corruption of the wrist-PPG signal, and the gate further raises the trustworthiness of delivered predictions to a level that matches or exceeds performance on uncorrupted inputs. Because the gate is grounded in a decision-theoretic notion of cost and is validated externally through the Uncertainty Calibration Error, it does not require ground-truth clean signals at deployment and is robust to the absence of a calibrated reconstruction-fidelity metric for one-dimensional time series.

The framework is model-agnostic, lightweight, privacy-compatible and decoupled from the underlying classifier; it can be integrated into existing wearable AF detection products without classifier retraining and supports gradual, regulator-friendly post-market improvement of each component independently. Future work will extend the gate to motion-conditioned non-Gaussian noise, to cost-asymmetric decision rules, and to non-AF cardiac arrhythmias, and will validate calibration transfer across demographic strata.

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

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 70, 214–223. https://doi.org/10.5555/3305381.3305404

Avila, F. R., Asín, J., Bourbon, F., & Bergasa, L. M. (2018). Wearable photoplethysmography for continuous monitoring: state of the art and a multimodal embedded prototype. Sensors, 18(7), 2197. https://doi.org/10.3390/s18072197

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

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 1(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

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

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

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 & Bioelectronics, 4(4), 195–202. https://doi.org/10.15406/ijbsbe.2018.04.00125

Charlton, P. H., Bonnici, T., Tarassenko, L., Clifton, D. A., Beale, R., & Watkinson, P. J. (2018). An assessment of

algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. Physiological Measurement, 39(11), 115008. https://doi.org/10.1088/1361-6579/aa670e

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/s10994-025-06755-7

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Colvonen, P. J., DeYoung, P. N., Bosompra, N. A., & Owens, R. L. (2020). Limiting racial disparities and bias for wearable devices in health science research. Sleep, 43(10), zsaa159. https://doi.org/10.1093/sleep/zsaa159

Csurka, G. (2017). Domain adaptation for visual applications: A comprehensive survey. In Domain Adaptation in Computer Vision Applications (pp. 1–35). Springer. https://doi.org/10.1007/978-3-319-58347-1_1

Delaney, A. M., Brophy, E., & Ward, T. E. (2019). Synthesis of realistic ECG using generative adversarial networks. arXiv preprint arXiv:1909.09150. https://doi.org/10.48550/arXiv.1909.09150

Demir, U., & Unal, G. (2018). Patch-based image inpainting with generative adversarial networks. arXiv preprint arXiv:1803.07422. https://doi.org/10.48550/arXiv.1803.07422

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. Proceedings of the 35th International Conference on Machine Learning (ICML), 80, 1184–1193. https://doi.org/10.5555/3327144.3327194

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., & Pal, C. (2016). The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications (pp. 179–187). Springer. https://doi.org/10.1007/978-3-319-46976-8_19

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

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

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

U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Center for Devices and Radiological Health. https://doi.org/10.31224/osf.io/k27hw

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35. https://doi.org/10.5555/2946645.2946704

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems (NeurIPS), 30, 4878–4887. https://doi.org/10.5555/3295222.3295241

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. Journal of the American Medical Informatics Association, 24(1), 198–208. https://doi.org/10.1093/jamia/ocw042

Goodfellow, I. J., 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 (NeurIPS), 27, 2672–2680. https://doi.org/10.5555/2969033.2969125

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1412.6572

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 (ICML), 70, 1321–1330.

https://doi.org/10.5555/3305381.3305518

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

Hatamian, F. N., Ravikumar, N., Vesal, S., Kemeth, F. P., Struck, M., & Maier, A. (2020). The effect of data augmentation on classification of atrial fibrillation in short single-lead ECG signals using deep neural networks. ICASSP 2020 — IEEE International Conference on Acoustics, Speech and Signal Processing, 1264–1268. https://doi.org/10.1109/ICASSP40776.2020.9053800

Healey, J. S., Connolly, S. J., Gold, M. R., Israel, C. W., Van Gelder, I. C., Capucci, A., Lau, C. P., Fain, E., Yang, S., Bailleul, C., Morillo, C. A., Carlson, M., Themeles, E., Kaufman, E. S., & Hohnloser, S. H. (2012). Subclinical atrial fibrillation and the risk of stroke. New England Journal of Medicine, 366(2), 120–129. https://doi.org/10.1056/NEJMoa1105575

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

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems (NeurIPS), 30, 6626–6637. https://doi.org/10.5555/3295222.3295408

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J. J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G. A., Dilaveris, P. E., Fauchier, L., Filippatos, G., Kalman, J. M., La Meir, M., & Lane, D. A. (2021). 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation. European Heart Journal, 42(5), 373–498. https://doi.org/10.1093/eurheartj/ehaa612

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

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. PLOS ONE, 16(7), e0254841. https://doi.org/10.1371/journal.pone.0254841

January, C. T., Wann, L. S., Calkins, H., Chen, L. Y., Cigarroa, J. E., Cleveland, J. C., Ellinor, P. T., Ezekowitz, M. D., Field, M. E., Furie, K. L., Heidenreich, P. A., Murray, K. T., Shea, J. B., Tracy, C. M., & Yancy, C. W. (2019). 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation. Circulation, 140(2), e125–e151. https://doi.org/10.1161/CIR.0000000000000665

Jin, J., Sanchez-Lengeling, B., Ratner, B. R., Aspuru-Guzik, A., & Saharia, C. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213. https://doi.org/10.48550/arXiv.2004.07213

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

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

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

Kim, I., Ramdas, A., Singh, A., & Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. Annals of Statistics, 49(1), 411–434. https://doi.org/10.1214/20-AOS1962

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

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical

machine learning. NPJ Digital Medicine, 4, 4. https://doi.org/10.1038/s41746-020-00367-3

Kornej, J., Börschel, C. S., Benjamin, E. J., & Schnabel, R. B. (2020). Epidemiology of atrial fibrillation in the 21st century: Novel methods and new insights. Circulation Research, 127(1), 4–20. https://doi.org/10.1161/CIRCRESAHA.120.316340

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. Proceedings of the 35th International Conference on Machine Learning (ICML), 80, 2796–2804. https://doi.org/10.5555/3327546.3327603

Lakhani, A. (2024). Revolutionizing orthopedic care: The impact of AI in predictive analysis, surgical precision, and personalized rehabilitation. The Journal of Community Health Management, 11(1), 1–6. https://doi.org/10.18231/j.jchm.2024.022

Laves, M.-H., Ihler, S., Kortmann, K.-P., & Ortmaier, T. (2020). Well-calibrated regression uncertainty in medical imaging with deep learning. Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL), 121, 393–412. https://doi.org/10.48550/arXiv.2010.04994

Lippi, G., Sanchis-Gomar, F., & Cervellin, G. (2021). Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. International Journal of Stroke, 16(2), 217–221. https://doi.org/10.1177/1747493019897870

Liu, J., Liu, X., Wang, M., & Sun, X. (2020). Generative adversarial network based on PixelHop++ for ECG denoising. Computers in Biology and Medicine, 121, 103790. https://doi.org/10.1016/j.compbiomed.2020.103790

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y. (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

Lubitz, S. A., Faranesh, A. Z., Selvaggi, C., Atlas, S. J., McManus, D. D., Singer, D. E., Pagoto, S., McConnell, M. V., Pantelopoulos, A., & Foulkes, A. S. (2022). Detection of atrial fibrillation in a large population using wearable devices: The Fitbit Heart Study. Circulation, 146(19), 1415–1424. https://doi.org/10.1161/CIRCULATIONAHA.122.060291

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. Proceedings of the 30th International Conference on Machine Learning (ICML) — Workshop on Deep Learning for Audio, Speech and Language Processing. https://doi.org/10.5555/3327546

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1706.06083

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. IEEE International Conference on Computer Vision (ICCV), 2813–2821. https://doi.org/10.1109/ICCV.2017.304

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 54, 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

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), 115:1–115:35. https://doi.org/10.1145/3457607

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

Mousavi, S., Afghah, F., & Acharya, U. R. (2019). HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. Computers in Biology and Medicine, 104, 41–50. https://doi.org/10.1016/j.compbiomed.2018.10.029

Müller, R., Kornblith, S., & Hinton, G. (2019). When does label smoothing help? Advances in Neural Information Processing Systems (NeurIPS), 32, 4694–4703. https://doi.org/10.5555/3454287.3454709

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well-calibrated probabilities using Bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29, 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 (ICML), 625–632. https://doi.org/10.1145/1102351.1102430

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS), 32, 8024–8035. https://doi.org/10.5555/3454287.3455008

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

Pollreisz, D., & TaheriNejad, N. (2019). Detection and removal of motion artifacts in PPG signals. Mobile Networks and Applications, 27, 728–738. https://doi.org/10.1007/s11036-019-01323-6

Pürerfellner, H., Pokushalov, E., Sarkar, S., Koehler, J., Zhou, R., Urban, L., & Hindricks, G. (2014). P-wave evidence as a method for improving algorithm to detect atrial fibrillation in insertable cardiac monitors. Heart Rhythm, 11(9), 1575–1583. https://doi.org/10.1016/j.hrthm.2014.06.006

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

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira, W., Schön, T. B., & Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications, 11, 1760. https://doi.org/10.1038/s41467-020-15432-4

Ribeiro, A. L. P., Paixão, G. M. M., Gomes, P. R., Ribeiro, M. H., Ribeiro, A. H., Canazart, J. A., Oliveira, D. M., Ferreira, M. P., Lima, E. M., Moraes, J. L., Castro, N., Ribeiro, L. B., & Macfarlane, P. W. (2021). Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study. Journal of Electrocardiology, 64, 130–135. https://doi.org/10.1016/j.jelectrocard.2019.09.008

Salehizadeh, S. M. A., Dao, D., Bolkhovsky, J., Cho, C., Mendelson, Y., & Chon, K. H. (2014). A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. Sensors, 16(1), 10. https://doi.org/10.3390/s16010010

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. Advances in Neural Information Processing Systems (NeurIPS), 29, 2234–2242. https://doi.org/10.5555/3157096.3157346

Sarkar, P., & Etemad, A. (2022). Self-supervised ECG representation learning for emotion recognition. IEEE

Transactions on Affective Computing, 13(3), 1541–1554. https://doi.org/10.1109/TAFFC.2020.3014842

Schäck, T., Muma, M., Feng, M., Guan, C., & Zoubir, A. M. (2017). Robust nonlinear causality analysis of nonstationary multivariate physiological time series. IEEE Transactions on Biomedical Engineering, 65(6), 1213–1225. https://doi.org/10.1109/TBME.2017.2728319

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences, 255, 16–29. https://doi.org/10.1016/j.ins.2013.07.030

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

Steinhubl, S. R., Waalen, J., Edwards, A. M., Ariniello, L. M., Mehta, R. R., Ebner, G. S., Carter, C., Baca-Motes, K., Felicione, E., Sarich, T., & Topol, E. J. (2018). Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: The mSToPS randomized clinical trial. JAMA, 320(2), 146–155. https://doi.org/10.1001/jama.2018.8102

Sun, B., Wang, C., Chen, X., Zhang, Y., & Shao, H. (2016). PPG signal motion artifacts correction algorithm based on feature estimation. Optik, 127(11), 4823–4827. https://doi.org/10.1016/j.ijleo.2017.04.087

Svennberg, E., Engdahl, J., Al-Khalili, F., Friberg, L., Frykman, V., & Rosenqvist, M. (2015). Mass screening for untreated atrial fibrillation: The STROKESTOP study. Circulation, 131(25), 2176–2184. https://doi.org/10.1161/CIRCULATIONAHA.114.014343

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

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

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022. https://doi.org/10.48550/arXiv.1607.08022

Voisin, M., Shen, Y., Aliamiri, A., Avati, A., Hannun, A., & Ng, A. (2017). Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. Proceedings of the Machine Learning for Health Workshop, NeurIPS. https://doi.org/10.48550/arXiv.1811.07774

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2021). Time series data augmentation for deep learning: A survey. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 4653–4660. https://doi.org/10.24963/ijcai.2021/631

Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology, 11(5), 51:1–51:46. https://doi.org/10.1145/3400066

Wulan, N., Wang, W., Sun, P., Wang, K., Xia, Y., & Zhang, H. (2020). Generating electrocardiogram signals by deep learning. Neurocomputing, 404, 122–136. https://doi.org/10.1016/j.neucom.2020.04.076

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. Advances in Neural Information Processing Systems (NeurIPS), 32, 5508–5518. https://doi.org/10.5555/3454287.3454781

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

ISSN: 3068-1197

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

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 9901, 424–432. https://doi.org/10.1007/978-3-319-46723-8_49

ACKNOWLEDGEMENTS

The authors thank the maintainers of the public DeepBeat-derived wrist-PPG cohort and the open-source community responsible for the deep-learning frameworks used in this work. The authors received no specific external funding for this study.

DATA AVAILABILITY

The data underlying this article are available from the public DeepBeat repository. Trained model weights and the reliability-gate evaluation scripts are available from the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

