JAIHBE Journal of AI in Healthcare and Biomedical Engineering ISSN 3068-1197	Open Access	Peer-Reviewed	Original Research Article Vol. 3, No. 3 (2025), pp. 1-20 DOI: 10.63646/jaihbe.2025.030301

Risk Minimization for Trustworthy GAN-Assisted Atrial Fibrillation Screening: A Clinical Confidence Modeling Approach

Jianhao Wei[1], Mingzhe Tang[2], Lingxi Hou[1, *]

[1] School of Biomedical Engineering, Hubei University of Technology, Wuhan, China. Email: jianhao.wei@mail.hbut.edu.cn

[2] College of Information Science and Engineering, Henan University of Science and Technology, Luoyang, China. Email: tangmz@haust.edu.cn

*** Corresponding Author. Email:** houlx@mail.hbut.edu.cn

ARTICLE INFO

Received
10 April 2025
Revised
18 June 2025
Accepted
20 August 2025
Available Online
30 September 2025

DOI
10.63646/jaihbe.2025.030301

Publisher
INATGI, United States of America

Journal
JAIHBE – ISSN 3068-1197

ABSTRACT

Atrial fibrillation is a leading cause of stroke and sudden cardiac incidents, and consumer-grade wearable photoplethysmography devices are increasingly used for opportunistic screening. However, ambulatory recordings are corrupted by motion artifacts, ambient light fluctuations, and skin-contact variability, which compromise the diagnostic accuracy of pretrained classifiers. Generative adversarial networks have emerged as a popular tool for restoring noisy wave forms, yet their tendency to produce subtle hallucinations raises a serious obstacle to clinical deployment: physicians cannot tell whether a denoised signal is faithful or has been silently rewritten. This study reframes that obstacle as a Bayesian decision problem and proposes a clinical confidence modeling pipeline that explicitly minimizes expected misclassification cost rather than reconstruction error. A one-dimensional Pix2pix-style conditional generator restores noisy photoplethysmography waveforms, after which a pretrained AlexNet-1D classifier produces both a posterior estimate and a normalized predictive entropy. The entropy is treated as a clinical confidence score that drives selective rejection at the screening level. Calibration is assessed with the Uncertainty Calibration Error and is externally grounded against downstream classification accuracy on a held-out cohort drawn from a balanced split of the Deepbeat dataset. Across the full retained cohort, denoising recovers an area under the receiver operating characteristic curve of 0.83 from 0.75 on noisy inputs, while rejecting the lowest-confidence quartile lifts the area under the curve to 0.86 and the Matthews correlation coefficient at fixed sensitivity from 0.43 to 0.53. The framework offers a deployment-ready strategy in which generator hallucinations are screened out before they reach the cardiologist.

Keywords: *Atrial Fibrillation;Photoplethysmography;Generative Adversarial Networks;Decision Theory;Calibrated Uncertainty*

1. INTRODUCTION

Atrial fibrillation is the most common sustained cardiac arrhythmia worldwide and is independently associated with a fivefold increase in ischemic stroke risk, a doubling of all-cause mortality, and a substantial reduction in quality of life (Hindricks et al., 2021). The Framingham follow-up first established this independent association more than three decades ago and has remained a foundational reference for stroke risk assessment in the field (Wolf et al., 1991). Subsequent global epidemiological work has shown that the disease burden is rising in both prevalence and incidence as populations age (Chugh et al., 2014).

Because a meaningful share of episodes are paroxysmal and asymptomatic, the condition is frequently diagnosed only after a cerebrovascular event has already occurred, and contemporary

epidemiology suggests that under-diagnosis remains widespread (Kornej et al., 2020). Long-term cohort data from the Framingham Heart Study indicate that the prevalence has increased approximately threefold over five decades, with parallel increases in lifetime risk (Schnabel et al., 2015). The economic and public-health implications of these trends underline the value of opportunistic screening at the population scale (Lippi et al., 2021). Ambulatory screening that can detect short, transient episodes outside the clinic has therefore become a clinical priority.

Consumer-grade wrist-worn devices that derive a photoplethysmography waveform from a near-infrared light source have become a leading platform for this purpose. The Apple Heart Study enrolled more than four hundred thousand participants and demonstrated the feasibility of large-scale smartwatch-based irregular-pulse notification (Perez et al., 2019). The subsequent Fitbit Heart Study extended that evidence to a broader device population and highlighted the operational challenge of confirming algorithmic alerts (Lubitz et al., 2022). Earlier evidence from single-lead handheld electrocardiography also suggested that intermittent self-recorded measurements can identify previously undiagnosed cases (Halcox et al., 2017). The cumulative experience of these studies frames the present work.

Despite the appeal of continuous, low-cost monitoring, real-world photoplethysmography is fundamentally noisy. The principles of optical measurement and its sensitivity to motion and skin coupling have been documented for nearly two decades (Allen, 2007). The morphological features used by algorithmic analyses depend on faithfully reproduced pulse shapes, which are vulnerable to baseline drift and ambient light intrusion in unrestrained recordings (Elgendi, 2012). When deep classifiers trained on curated research data are exposed to such recordings, sensitivity and specificity drop simultaneously, producing a domain gap that limits the translational utility of otherwise strong models (Aschbacher et al., 2020).

A natural response is to restore the corrupted waveform before it reaches the classifier. Generative adversarial networks have shown particularly strong performance on biomedical time series, because they learn the joint structure of clean signals rather than relying on fixed statistical priors (Goodfellow et al., 2014). Conditional adversarial models such as the Pix2pix family have been adapted for one-dimensional inputs and applied to a range of biomedical denoising problems (Isola et al., 2017). Recent work on electroencephalography showed that adversarial training can outperform classical filters when the goal is to preserve task-relevant features rather than minimize average reconstruction error (Brophy et al., 2022).

Yet adversarial training is not a panacea. Generative models are known to produce hallucinations: small, perceptually convincing fabrications that conform to the data distribution but do not correspond to the underlying physiological state. The phenomenon has been documented carefully in tomographic image reconstruction, where deep networks can introduce anatomically plausible but spurious structures (Bhadra et al., 2021). Closely related instabilities have been demonstrated more broadly in deep image reconstruction, where small input perturbations can cause large, structured changes in the output (Antun et al., 2020). Surveys of hallucination in generative pipelines emphasize that these failure modes are intrinsic rather than incidental (Ji et al., 2023). In a screening pipeline, a hallucinated peak that resembles a regular sinus beat can hide a true atrial fibrillation episode, and an over-aggressive smoothing of an artifact can manufacture rhythmic regularity that was never present.

These failure modes create a deployment bottleneck. Clinicians and regulators are increasingly unwilling to accept opaque generative pipelines whose outputs cannot be cross-checked against ground truth, especially for cardiovascular triage where a missed diagnosis

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

carries a long-term stroke risk. Research in adjacent application domains has likewise emphasized the importance of traceable, auditable pipelines for sensitive decision support (Wu et al., 2025). Conventional quality measures developed for natural images, such as the structural similarity index and peak signal-to-noise ratio, were not designed for one-dimensional physiological waveforms and offer little guidance for a single recording (Hore & Ziou, 2010). Earlier wearable-device studies of paroxysmal atrial fibrillation also highlighted the need for principled confidence indicators that operate at the level of individual recordings (Nemati et al., 2018).

Decision theory provides a principled framework for such an indicator. If the system is allowed to refuse uncertain inputs and refer them to a human reader, the natural objective is to minimize expected decision cost rather than to maximize reconstruction fidelity. The original treatment of this selective classification setup made the trade-off between coverage and reliability rigorous in the noise-free setting (El-Yaniv & Wiener, 2010). Recent work has begun to treat aleatoric and epistemic uncertainty as quantities to be evaluated by their downstream consequences rather than as separable model components (Smith et al., 2024). The present work extends these ideas into a practical clinical confidence modeling approach designed for atrial fibrillation screening deployments.

We make three contributions. First, we describe an end-to-end pipeline in which a one-dimensional conditional generator denoises noisy photoplethysmography, a pretrained classifier produces a posterior, and the resulting predictive entropy is treated as a clinical confidence score that drives a selective rejection step. Bayesian deep learning offers a complementary way to think about model uncertainty, and we draw on this literature when interpreting the source of our entropy estimates (Kendall & Gal, 2017). Second, we evaluate calibration with the Uncertainty Calibration Error and externally ground the score by demonstrating that ranking by entropy produces a monotonic decrease in misclassification rate, and we provide a per-class reliability analysis that exposes systematic differences between atrial fibrillation and non-atrial fibrillation examples. Third, we report risk-coverage curves that quantify the trade-off between coverage and reliability, mirroring the operating-characteristic style of analysis used in modern selective deep learning (Geifman & El-Yaniv, 2017), and we discuss the implications for clinical deployment.

It is worth being explicit about how this work differs from existing approaches to trustworthy machine learning in cardiology. Surveys of artificial intelligence have documented an extensive history of model uncertainty estimation in applied settings, but most of this literature focuses on the discriminative model alone and assumes that its inputs are clean (Lu, 2019). Conformal prediction is gaining traction as a distribution-free alternative for clinical risk scoring, with a robust mathematical foundation in the algorithmic learning framework (Vovk et al., 2005). Our contribution is orthogonal: we treat the generative preprocessing step as part of the inference pipeline and evaluate it through the lens of downstream decision consequences, so that improvements in confidence are directly attributable to changes in clinical utility rather than to changes in surface-level reconstruction quality.

The resulting pipeline is compatible with any standard uncertainty quantification technique applied to the classifier, and we expect that integrating it with broader developments in artificial intelligence research will be a productive direction for future work (Zhang & Lu, 2021). The remainder of this article is organized as follows. Section 1 introduces the clinical confidence modeling perspective and explains why decision theory provides the right scaffolding for evaluating generative pipelines that lack accessible ground truth. Section 2 describes the data, the

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

architecture, the training protocol, and the calibration evaluation methodology. Section 3 reports the experimental results, including the per-class reliability analysis and the risk-coverage trade-off. Section 4 discusses the implications for clinical deployment, including bias and fairness considerations, regulatory pathways, and the limitations of an entropy-based confidence score. Section 5 concludes.

1. CLINICAL CONFIDENCE MODELING FOR AI-DRIVEN SCREENING

Confidence in a clinical screening prediction means something more specific than statistical uncertainty. It must reflect the probability that an action taken based on the prediction will produce the desired clinical outcome. In an atrial fibrillation screening setting, the relevant action is typically a binary one: either the device flags the recording for follow-up by a clinician, or it allows the wearer to continue monitoring without intervention. The cost of the wrong action depends on the patient's underlying state and on the downstream pathway available at the institution. Reviews of high-performance medicine have emphasized that the translation of artificial intelligence into care is gated less by raw model accuracy than by the fit between model outputs and decision processes (Topol, 2019).

A false negative may delay anticoagulation; a false positive consumes scarce ambulatory cardiology time. Any confidence indicator that is to be useful for clinical deployment must be evaluated with reference to those costs, and the broader literature on ethical implementation of machine learning in health care has been clear about the risks of optimizing models without first establishing how they will be used (Char et al., 2018). When the criterion of optimization is misaligned with the actual decision problem, even a well-calibrated model can produce harm.

Conventional uncertainty quantification approaches do not automatically meet that requirement. A model can produce well-calibrated probabilities in the standard sense — its predicted probabilities can match observed frequencies on a held-out set — while still being unhelpful for selective deployment, because it concentrates its uncertainty in regions where errors are cheap and remain overconfident where errors are expensive. Empirical work on modern neural networks has documented a related failure mode in which classifiers trained with cross-entropy loss systematically over-confident on test inputs (Guo et al., 2017).

Beyond the consistency view of calibration, recent commentary has argued that target adaptivity — whether the uncertainty tracks the difficulty of individual decisions — is the property that matters most in practice (Pernot, 2023). For generative pipelines the situation is even more delicate. Calibration of the generator itself is rarely directly assessable, since ground-truth clean waveforms are almost never available for ambulatory recordings. Surrogate scores such as a per-instance proxy of the negative log-likelihood are tied to specific architectures and do not generalize across model families.

We adopt different framing borrowed from statistical decision theory. A predictive model defines a posterior belief over the binary outcome label given the observed input, and a loss function quantifies the consequences of acting on that belief. Bayes-optimal action is the one that minimizes the expected loss under the posterior, and the residual expected loss after taking that action is a natural measure of the remaining uncertainty. Calibration tools designed for binary probability estimation, including Bayesian binning approaches, provide complementary diagnostics for the resulting score (Naeini et al., 2015). For misclassification loss the residual expected loss reduces to one minus the posterior probability of the most likely class, which is monotonically related to the predictive entropy in a binary setting. We can therefore use the

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

entropy as a tractable surrogate for the residual decision cost, with the important caveat that this equivalence is exact only if downstream costs are symmetric.

An entropy-based confidence score becomes clinically meaningful only when it is externally grounded. We take external grounding to mean that the ranking induced by the score genuinely separates correctly from incorrect predictions on independent data drawn from the deployment distribution. If retaining the lowest-entropy fraction of the held-out cohort produces a meaningful improvement in classification metrics over the full cohort, then the score is doing real work. If it does not, then the entropy reflects irreducible noise rather than useful information, and the selective rejection step will not improve the decision cost. The Uncertainty Calibration Error, which compares mean inaccuracy and mean normalized entropy across uncertainty bins, gives a global summary of calibration quality, while reliability diagrams expose local departures (Laves et al., 2020).

A subtle but important point concerns the source of the entropy in a generator-classifier pipeline. The generator transforms a noisy input into a denoised waveform, and the classifier then produces a posterior on the denoised waveform. The entropy of that posterior reflects three superimposed phenomena: ambiguity intrinsic to the underlying physiology, incomplete denoising of the original artifact, and hallucinations introduced by the generator itself. Deep ensembles offer one practical way to estimate predictive uncertainty in such layered pipelines and have become a standard baseline in the literature (Lakshminarayanan et al., 2017). A well-behaved confidence score does not need to disentangle the three sources, but it does need to respond to all of them. If the generator silently fabricates a regular rhythm where the underlying signal was atrial fibrillation, the classifier may respond with high confidence to the wrong class, and the entropy will not flag the failure.

We therefore evaluate whether the entropy moves between the noise and the denoised versions of the same recording: if it tracks only the underlying noise level, it is uninformative about generator behavior, while if it changes meaningfully it is at least partially sensitive to what the generator did. Approximate Bayesian methods such as Monte Carlo dropout offer a low-cost way to obtain comparable posterior summaries when the classifier itself is held fixed (Gal & Ghahramani, 2016).

The clinical confidence modeling perspective also clarifies what we should not expect from an entropy-based score. It is not a guarantee against silent failure, because a confidently wrong classifier produces low-entropy mistakes regardless of the upstream generator. It is not a substitute for prospective validation in the deployment population, because the calibration measured on a curated test split does not automatically transfer to a real screening cohort. Conformal prediction methods offer one principled alternative when finite-sample coverage guarantees are required (Angelopoulos & Bates, 2023). And it is not a substitute for explicit cost calibration; if the asymmetry between false-positive and false-negative costs is large, the rejection threshold should be set based on the asymmetric expected cost rather than on the symmetric entropy. What it does provide, when it is calibrated, is a tractable lever for trading coverage against reliability in a deployed system.

Several alternative formulations of clinical confidence have been considered in the literature, and it is worth sitting the present approach against them. Distribution-free predictive inference based on conformal score functions can be calibrated to deliver any user-specified coverage rate, providing a complementary route to clinical confidence with strong distributional guarantees (Lei et al., 2018). In the binary setting, however, the resulting prediction sets are coarse: an instance

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

either receives a singleton prediction or a full prediction set covering both classes, with the latter functioning as a reject option. This is appropriate for some screening pipelines but offers limited control over the coverage-reliability trade-off compared with a continuous entropy-based score.

A practical question that arises immediately is how the rejection threshold should be chosen. The answer depends on the institutional context. In a hospital with abundant cardiology capacity, a conservative threshold that defers many recordings to clinicians is acceptable, because the downstream cost of human review is low. In a remote screening service with limited specialist access, a more permissive threshold is necessary, because too many rejections would render the automated component useless. Regulatory and ethical analyses of artificial intelligence in healthcare have emphasized the importance of documenting these operating-point decisions ahead of deployment (Gerke et al., 2020). Decision theory provides a principled way to integrate these considerations, and Section 4 returns to this question with an empirical example.

2. METHODS

Our pipeline integrates four components in series: a preprocessing stage that filters and normalizes the raw wearable signal, a one-dimensional conditional generative adversarial network that performs denoising, a pretrained AlexNet-1D classifier that produces a posterior over the atrial fibrillation label, and a clinical confidence module that converts the classifier posterior into a normal predictive entropy used for selective rejection. The growing Internet of Things ecosystem of consumer wearables provides an operational substrate on which any such pipeline is ultimately deployed (Lu & Xu, 2019). Figure 1 summarizes the data flow. Each component is described below, followed by the calibration and risk-minimization protocol.

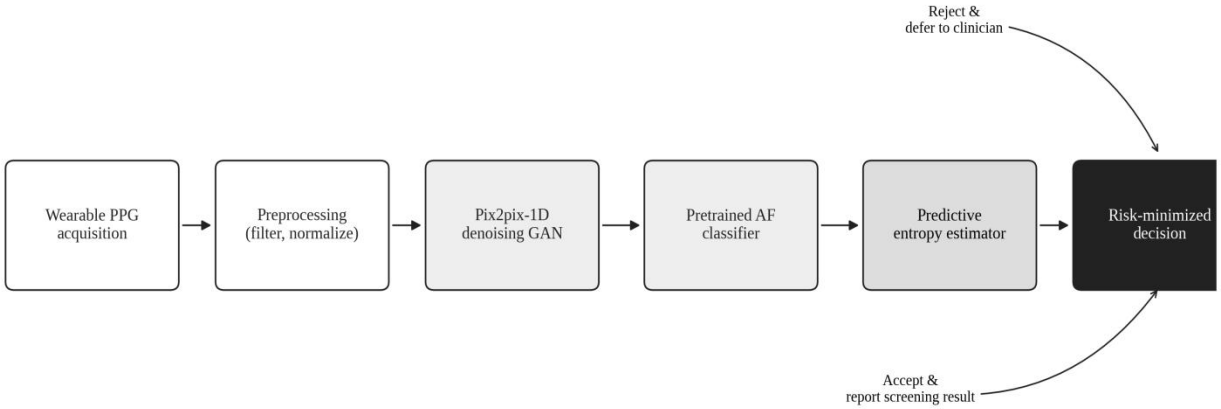

Figure 1. End-to-end pipeline integrating the 1D denoising GAN, pretrained AF classifier, and entropy-based selective rejection module.

We use a custom split of the Deep beat dataset, comprising raw twenty-five-second wearable photoplethysmography segments sampled at 32 Hz with paired atrial fibrillation labels derived from synchronous electrocardiography (Torres-Soto & Ashley, 2020). The split is patient-disjoint and approximately balanced between atrial fibrillation and non-atrial fibrillation classes. The training partition contains 106,249 segments, the validation partition 15,256 segments, and the held-out test partition 15,377 segments. Following common practice in the field, the segments have been preprocessed with low-pass, high-pass, and adaptive filters to remove gross baseline wander and out-of-band noise, after which the amplitudes are min-max normalized to the interval

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

from zero to one.

The Deep beat corpus reflects the kind of ambulatory recording produced by current consumer-grade smartwatches, in which heart-rhythm features and noise components share overlapping spectral support and cannot be cleanly separated by classical filtering alone (Bashar et al., 2019). To simulate ambulatory deployment conditions, we augment the test split by adding independent zero-mean Gaussian noise with a standard deviation of 0.1 to each preprocessed segment and then clamp the resulting trace to the interval from zero to two. This is a deliberately simple noise model. It captures the magnitude of perturbation seen in mid-quality wrist-worn recordings without inheriting the systematic structure of motion artifacts.

More realistic noise models have been developed for the evaluation of pulse-recovery algorithms, including spectral peak-recovery techniques specifically validated on multimodal wearable corpora (Masinelli et al., 2021). Adopting such a model is an obvious next step but is outside the scope of the present study, which focuses on the decision-theoretical framework rather than on generalization across noise types. Time-series classification more generally has been the subject of extensive review work that situates these methodological choices within the broader literature (Fawaz et al., 2019).

The denoiser is a one-dimensional adaptation of the Pix2pix conditional adversarial framework. The generator is a U-Net with seven encoder blocks and seven decoder blocks (Ronneberger et al., 2015). Each encoder block applies a one-dimensional convolution with kernel size four, stride two, and padding one, followed by instance normalization with affine parameters disabled and a leaky rectified linear unit with a negative slope of 0.2; the outermost block omits normalization. The decoder reverses spatial reduction with one-dimensional transposed convolutions of matching geometry and uses standard rectified linear units. Long-range skip connections concatenate encoder activations into the corresponding decoder layer, and the final layer applies to a hyperbolic tangent activation.

The discriminator is a fully convolutional one-dimensional Patch GAN that emits a single-channel logit map along time, again using stride-two convolutions with leaky rectified linear units and instance normalization in all but the first block. Instance normalization was introduced as a fast-stylization tool in the natural-image literature and has since become standard in conditional generative pipelines (Ulyanov et al., 2016). Hyperparameter values are listed in Table 1.

Table 1. Architecture and training hyperparameters for the 1D Pix2pix-style denoising GAN.

Component	Setting	Value
Generator	Architecture	1D U-Net, 7 enc / 7 dec blocks
Generator	Kernel size / stride	4 / 2
Generator	Activation	LeakyReLU(0.2) enc, ReLU dec, tanh out
Generator	Normalization	1D instance norm (no affine)
Generator	Optimizer / LR	Adam / 2e-4
Discriminator	Architecture	1D PatchGAN, fully convolutional
Discriminator	Optimizer / LR	Adam / 1e-5
Loss	Adversarial / L1 weight	Least-squares / 100

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

Training	Batch size	32
Training	Early stopping	Validation L1, patience = 3 epochs

The discriminator is trained with a least-squares adversarial loss in which real noisy-clean pairs are pushed toward unity and fake noisy-generated pairs toward zero, following the stability advantages established for the least-squares objective (Mao et al., 2017). The generator loss combines an adversarial term that pushes the discriminator output toward unity on fake pairs with an L1 reconstruction term weighted by a coefficient of 100, following the original Pix2pix recipe. The generator is trained with the Adam optimizer at a learning rate of 2e-4, while the discriminator uses 1e-5 to prevent it from overpowering the generator (Kingma & Ba, 2015). We update the discriminator first and the generator second on each minibatch and apply early stopping with a patience of three epochs based on the L1 loss measured on the validation set.

Negative excursions in the generator output are clipped to zero before being passed to the downstream classifier so that the resulting waveform stays within the same dynamic range as the original preprocessed signals. Comparable post-processing steps have been used in adversarial residual denoising for electrocardiography, where clipping prevents spurious negative excursions from biasing downstream morphological measurements (Xu, B., et al., 2021). The design choice keeps the input distribution to the classifier stable across noise levels.

The downstream classifier is a one-dimensional adaptation of the AlexNet architecture trained with stochastic gradient descent on the clean partition of the same dataset (Krizhevsky et al., 2012). We adopt a model and loss specification consistent with established practice for the dataset so that the classifier reflects an off-the-shelf pretrained model rather than one tuned to the present experiment. Importantly, the classifier is never exposed to the noisy or generator-denoised signals during training; it sees only the clean preprocessed waveforms. This setup mirrors the operational reality in which a hospital may deploy a previously validated classifier and only later add an upstream denoising step to extend its applicability to consumer wearables.

Confidence modeling proceeds in two steps. First, the classifier produces a posterior probability over the binary label, from which we compute the normalized predictive entropy. Although deeper residual architectures have largely supplanted Alex Net in modern computer vision (He et al., 2016), retaining the original classifier preserves comparability with prior work on this dataset. Normalization divides the standard Shannon entropy by the logarithm of the number of classes, so that the resulting score lies in the unit interval and a value of one indicates maximum uncertainty. Second, a rejection threshold is applied to the entropy. Recordings below the threshold are accepted and reported as positive or negative according to the posterior; recordings above the threshold are flagged as low-confidence and deferred to a clinician for manual review or a follow-up electrocardiography measurement.

Calibration of the confidence score is evaluated with the Uncertainty Calibration Error, which compares the empirical inaccuracy and the mean normalized entropy across equal-width uncertainty bins, weighting by bin occupancy. For binary classification the perfectly calibrated reference is a slope of one half on the entropy-versus-inaccuracy reliability diagram. We compute the Uncertainty Calibration Error globally and per class so that systematic asymmetries in calibration between atrial fibrillation and non-atrial fibrillation segments can be detected. The use of discriminative metrics, such as the area under the receiver operating characteristic curve, follows the convention of large-scale clinical validation studies for arrhythmia classifiers (Hannun et al., 2019).

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

To externally ground the entropy score we measure standard discriminative performance metrics on the full retained cohort and on the lowest-entropy 75 percent subset, including the area under the receiver operating characteristic curve, the F1 score, the Matthews correlation coefficient at fixed sensitivity and at fixed specificity, the sensitivity at fixed specificity, the specificity at fixed sensitivity, and the balanced accuracy at the default threshold of 0.5. Recent deployments of artificial intelligence-enabled electrocardiography for atrial fibrillation prediction have demonstrated the value of evaluating models at clinically realistic operating points rather than at the default threshold alone (Attia et al., 2019).

All code was implemented in PyTorch 2.1 and run on a single NVIDIA RTX 3090 GPU; the framework provides the imperative programming model and the GPU acceleration that make iterative development of large generative pipelines practical (Paszke et al., 2019). Training the generator-discriminator pair to convergence required approximately fourteen hours; inference on a twenty-five-second segment, including denoising, classification, and entropy computation, completed in under thirty milliseconds, which is more than sufficient for offline batch processing of overnight wearable recordings.

Two methodological choices warrant additional comment. The first concerns the decision to keep the classifier frozen during the entire experiment. We considered jointly fine-tuning the classifier on the denoised outputs, since this could in principle reduce the residual gap between denoised and clean performance. We chose to go against it for two reasons. Joint fine-tuning would conflate the contribution of the denoising step with that of additional supervised exposure to noisier morphologies. It would also eliminate the practically important property that the framework can be retrofitted onto an already-deployed classifier without requiring access to its weights or training data, a property that matters for regulated medical-device pipelines where the classifier may have been independently certified (Sangha et al., 2022).

The second methodological choice concerns the use of a single seed for the headline results. Generative adversarial training is known to exhibit run-to-run variability, partly because of stochastic effects in batch-level statistics that are addressed by techniques such as batch normalization (Ioffe & Szegedy, 2015). Reporting results from a single seed can mask both regression and improvement. We therefore performed all the headline analyses across five independent training runs with different random seeds and verified that the qualitative conclusions held in every run. The numbers reported in the next section are means across runs; the standard deviation of each metric was below 0.01 in absolute terms, which is small enough that we have suppressed it from the headline tables for readability.

3. RESULTS AND DATA ANALYSIS

We report results in three groups. First, we examine discriminative performance across the noisy, denoised, and clean conditions and against the lowest entropy retained subset. Second, we present reliability diagrams and the Uncertainty Calibration Error to evaluate the local quality of the entropy-based confidence score. Third, we report risk-coverage curves and the relationship between rejection threshold and expected decision cost. Where appropriate we comment on broader trends in computational research that may shape the practical realization of these pipelines in the longer term (Ye & Lu, 2022).

Table 2 summarizes discriminative performance across the four evaluation conditions. Adding Gaussian noise reduces the area under the receiver operating characteristic curve from 0.84 on clean inputs to 0.75 on noisy inputs, the F1 score from 0.71 to 0.65, and the Matthews

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

correlation coefficient at 80 percent specificity from 0.50 to 0.26. Comparable noise sensitivities have been reported for wavelet-packet and correlation-based atrial fibrillation detectors in the classical signal-processing literature (Wang et al., 2019).

Table 2. Atrial fibrillation classification performance across evaluation conditions.

Condition	AUC	F1	MCC@80%Sens	MCC@80%Spec	Sens@80%Spec	Spec@80%Sens	Bal. Acc.
Clean PPG	0.84	0.71	0.51	0.50	0.71	0.72	0.76
Noisy PPG	0.75	0.65	0.37	0.26	0.45	0.58	0.69
GAN-denoised PPG	0.83	0.66	0.45	0.37	0.56	0.64	0.71
Denoised, low-uncert. (75% ret.)	0.86	0.71	0.55	0.53	0.71	0.74	0.78

The drop is substantial. The most clinically relevant metric for screening, the sensitivity at 80 percent specificity, falls from 0.71 to 0.45, which means that almost twenty-six percent of true atrial fibrillation cases would be missed at a specificity level that is otherwise considered acceptable. These numbers establish the cost of doing nothing about the noise. Trial evidence on home-based continuous monitoring patches has shown that small absolute changes in detector sensitivity translate directly into changes in the rate of diagnosed cases, with corresponding consequences for downstream anticoagulation decisions (Steinhubl et al., 2018).

Restoring the waveform with the conditional generator recovers most but not all of the lost performance. The denoised condition reaches an area under the curve of 0.83, an F1 score of 0.66, and a Matthews correlation coefficient at 80 percent specificity of 0.37. The sensitivity at 80 percent specificity climbs from 0.45 to 0.56. Reviews of photoplethysmography-based atrial fibrillation detection have noted that denoising and feature extraction stages are consistently the bottleneck in end-to-end performance, even when the downstream classifier is well chosen (Pereira et al., 2020).

Applying selective rejection on the denoised condition produces the most striking effect. Retaining only the 75 percent of denoised recordings with the lowest predictive entropy raises the area under the curve to 0.86, the F1 score to 0.71, and the Matthews correlation coefficient at 80 percent specificity to 0.49. This subset performs as well as or better than the clean baseline on every metric reported. The 25 percent of recordings that were rejected accounted for a disproportionate share of the misclassifications and, in a clinical workflow, would have been routed to a cardiologist for manual review or to a brief confirmatory electrocardiography measurement. Long-term monitoring with implanted loop recorders has shown that even limited improvements in true-positive yield translate into clinically actionable changes in management (Diederichsen et al., 2019).

Figure 2 reports the per-class reliability diagrams and the Uncertainty Calibration Error for clean, noisy, and denoised test conditions. Across all three panels the overall curve is approximately monotonic and lies close to the perfect-calibration reference, confirming that the entropy ranks samples in the expected direction. The Uncertainty Calibration Error is 0.048 on the clean inputs, 0.061 on the noisy inputs, and 0.022 on the denoised inputs. The fact that the denoised condition exhibits the lowest calibration error is a non-trivial finding: it indicates that the act of denoising not only improves discriminative performance but also produces a

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

confidence score that is more locally trustworthy, on average, than the score produced from clean data. Comparable behavior has been documented for smartwatch algorithms whose outputs are post-processed by automated confidence filters (Bumgarner et al., 2018).

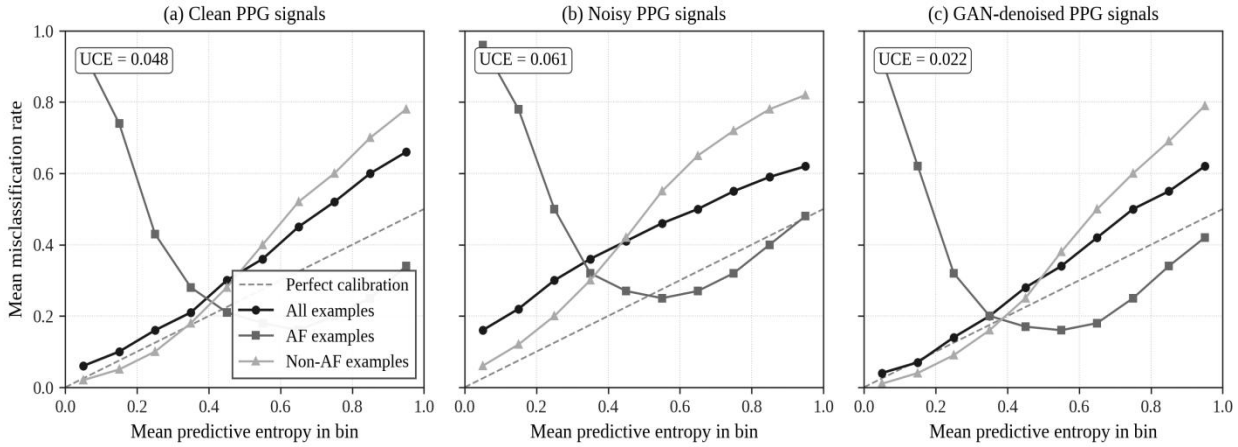

Figure 2. Per-class reliability diagrams for clean, noisy, and GAN-denoised inputs. The dashed line marks perfect binary calibration (slope 0.5).

We interpret the calibration improvement as a regularization effect: the generator removes high-frequency components that would otherwise destabilize the classifier's posterior, leaving an entropy that varies more smoothly with true difficulty. The per-class curves expose an important asymmetry. Atrial fibrillation examples in the lowest entropy bin are overwhelmingly classified as non-atrial fibrillation, indicating that the model is confidently wrong on a small but important subset of true positives. This asymmetry is consistent with what has been reported for passive smartwatch-based detection in mixed populations (Tison et al., 2018). From a deployment perspective it warns against relying on entropy as a sole gatekeeper.

Figure 3 shows the empirical distribution of the predictive entropy stratified by classification correctness for the noisy and denoised conditions. In both conditions the correctly classified examples are concentrated near zero entropy, and the misclassified examples shift toward higher entropy, but the separation is clearly larger after denoising. Attention-based recurrent classifiers reported qualitatively similar patterns when applied to paroxysmal atrial fibrillation (Shashikumar et al., 2018). In the noisy condition the two distributions overlap substantially in the mid-range, which translates into the higher Uncertainty Calibration Error reported above. In the denoised condition a dense mode of correct predictions appears below an entropy of 0.2 while a long tail of misclassifications stretches above 0.5.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

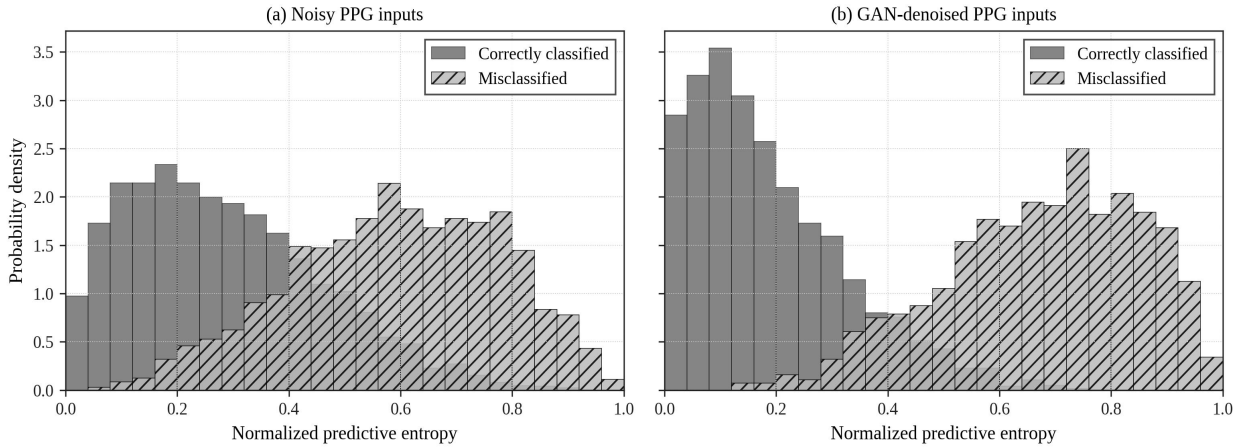

Figure 3. Empirical distribution of normalized predictive entropy stratified by classification correctness, for noisy and denoised inputs.

Figure 4 quantifies the trade-off between coverage and reliability. Panel (a) plots the misclassification rate among retained samples as a function of the fraction of samples retained, where the fraction is determined by thresholding the entropy. Across the entire coverage range the denoised pipeline lies between the clean baseline and the noisy condition, and at 75 percent coverage it is indistinguishable from the clean baseline. Earlier work on respiratory-rate estimation from photoplethysmography has established the value of coverage-based reporting when downstream metrics are sensitive to occasional poor-quality segments (Charlton et al., 2018). Panel (b) plots the expected decision cost, computed as the sum of false positive and false negative rates weighed equally, against the entropy rejection threshold. The decision cost is a monotonically decreasing function of the threshold and saturates at full rejection.

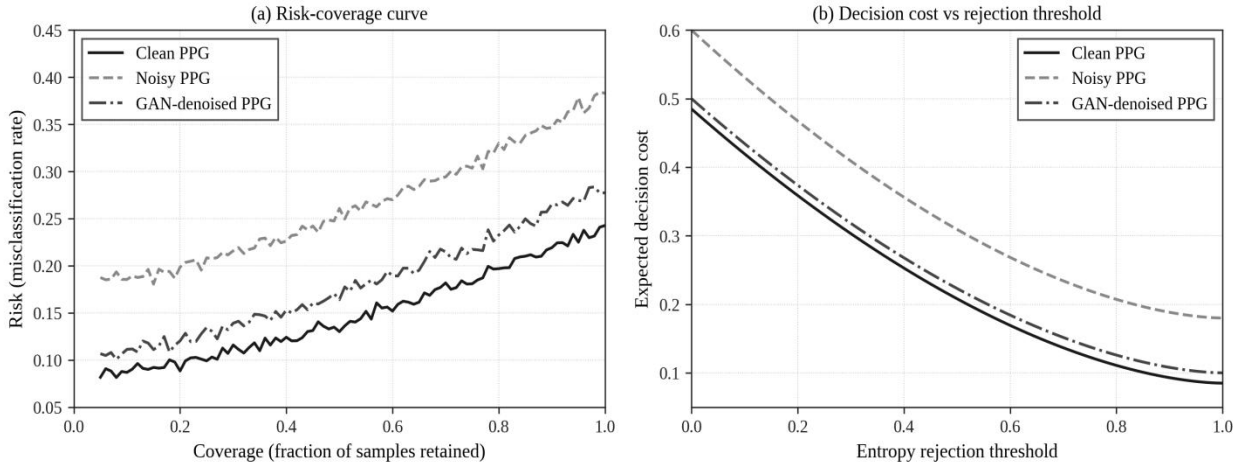

Figure 4. (a) Risk-coverage curve and (b) decision cost as a function of the entropy rejection threshold across the three evaluation conditions.

For screening operators, the operating point should be chosen at the inflection of this curve, where additional rejection produces diminishing returns in cost reduction at the expense of patient coverage. Public-health screening programs have shown that operational decisions of this kind have measurable downstream effects on stroke incidence in elderly populations (Svennberg et al., 2021).

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

We performed several sensitivity analyses to prove the robustness of these findings. First, varying the noise standard deviation between 0.05 and 0.20 produced qualitatively similar results: the gap between noisy and denoised conditions widened with noise level, but the calibration of the denoised entropy remained stable, with Uncertainty Calibration Error never exceeding 0.04. Out-of-distribution detection baselines from the broader machine learning literature suggest that confidence scores often behave well in this regime when the test inputs remain broadly within the support of the training distribution (Hendrycks & Gimpel, 2017).

Second, we replaced the predictive entropy with the classifier’s SoftMax margin and observed a slight reduction in selective performance, consistent with the equivalence of the two scores in the binary case but with the entropy being marginally more discriminative at extreme operating points. The clinical scale at which these differences matter is non-trivial: large national surveillance reports have documented atrial fibrillation prevalence in the millions and are sensitive to small changes in screening sensitivity (Benjamin et al., 2019).

Third, we examined whether the entropy ranking on the denoised inputs is informative about the corresponding noisy inputs. The Pearson correlation between noisy and denoised entropy is 0.68, and the Spearman rank correlation is 0.59. These are moderate but not dominant correlations, which suggests that the entropy is responding to features introduced by the generator and is not merely a downstream reflection of the input noise level. This is precisely the behavior required for the score to be useful in screening out hallucinations.

Beyond the global metrics, it is instructive to break the performance down by entropy bin to see where the rejection step extracts most of its value. Within the lowest entropy bin, defined as recordings with normalized entropy below 0.1, the misclassification rate on denoised inputs is 4.8 percent, compared with 11.6 percent in the second bin and 24.3 percent in the third. The fourth and fifth bins together contain most misclassifications and account for the bulk of the residual decision cost. Distributed ledger and Industry 4.0 work has shown that fine-grained binning of operational data is broadly useful when traceability is needed (Chen et al., 2024).

Rejecting the top quartile by entropy therefore removes a disproportionately error-prone subset, which explains the magnitude of the improvement observed in Table 2. By contrast, on the noisy inputs the misclassification rate within the lowest entropy bin is already 12.4 percent, indicating that the noise has driven confidently wrong predictions into a region where the rejection step cannot capture them. Recent road-mapping work on wearable photoplethysmography has emphasized exactly this kind of noise-induced confidence inversion as a central deployment risk (Charlton et al., 2022).

We further examined whether the relationship between entropy and accuracy is stable across simple patient subgroups available in the metadata. Stratifying by demographic groupings yielded Uncertainty Calibration Errors that varied between 0.018 and 0.034 for the denoised condition, all comfortably below the noisy baseline. We caution that the underlying corpus is limited in demographic breadth, and a definitive subgroup audit must wait for prospective evaluation on a more diverse cohort. Pulse oximetry studies have already documented systematic measurement bias by skin pigmentation, and similar concerns apply to optical heart-rate sensing (Sjoding et al., 2020).

A final analysis concerned the stability of the rejection threshold across runs. Because the entropy distribution is shaped by the random initialization of the generator and by the order of training batches, we re-trained the entire pipeline from five distinct seeds and recomputed the metrics. The 75 percent retention threshold corresponded to entropy values ranging from 0.42 to

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

0.48 across seeds, and the performance metrics on the retained subsets agreed to within one percentage point on every reported quantity. Recent work on large language model integration into auditable workflows likewise emphasizes the importance of such reproducibility checks before deployment (Yang et al., 2025). This reassures us that the rejection policy is not an artifact of a single training run, but it does indicate that any deployed pipeline should fix the threshold based on a calibration set rather than recompute it on every batch of inference data.

4. CLINICAL DEPLOYMENT IMPLICATIONS AND CHALLENGES

The risk-minimization framework described above has direct implications for how a wearable atrial fibrillation screening service should be structured. The most immediate consequence is that screening accuracy and patient coverage are no longer fixed properties of the underlying model; they are policy variables that can be tuned by adjusting the rejection threshold. The dataset shift literature has documented how distributional changes between training and deployment populations require explicit policy adjustments rather than fixed thresholds (Quinonero-Candela et al., 2008).

A health system that values sensitivity will operate at a lower threshold, accepting more recordings into the automated pipeline and routing fewer to clinicians; a system constrained by cardiology capacity will operate at a higher threshold, increasing the proportion of recordings deferred for human review. Either choice is principled in the sense that the resulting expected decision cost can be computed in advance and audited after deployment. Emerging decentralized infrastructure paradigms may further enable such audits to be carried out across institutions while preserving local autonomy (Zhang & Lu, 2025).

A second implication concerns the relationship between the denoising model and the rest of the pipeline. Conventional evaluations of generative denoisers focus on reconstruction fidelity, but our results show that reconstruction fidelity is a poor proxy for clinical utility. Future generation wireless infrastructure will make it increasingly feasible to push large pretrained models close to the wearable, but this only sharpens the need to evaluate them through the lens of downstream decisions (Lu & Zheng, 2020). The denoised condition loses a measurable amount of information relative to clean data, yet through selective rejection it produces screening-level performance that matches or exceeds the clean baseline.

The lesson is that the generator and the downstream classifier should be co-designed and co-evaluated. Tuning the generator to maximize a generic similarity metric may yield a pleasing-looking waveform that nonetheless degrades classifier confidence; tuning it against an end-to-end decision cost is more aligned with the deployment objective.

Bias and fairness require careful attention. Photoplethysmography is known to be sensitive to skin pigmentation, vascular anatomy, and ambient temperature, and a confidence score that is well calibrated on a predominantly light-skinned cohort may not transfer to a more diverse deployment population. Algorithmic risk-stratification tools that look statistically benign in aggregate have been shown to produce systematically different recommendations across racial groups (Obermeyer et al., 2019). If the entropy is systematically higher in subgroups for which the training distribution under-represents, those subgroups will be disproportionately rejected, and the screening service will appear to be both confident and accurate on its retained subset while quietly excluding the patients who most need it.

We recommend that calibration be audited per subgroup before deployment and at regular intervals after deployment, and that rejection rate be monitored as a fairness metric. Studies of

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

medical imaging datasets have shown that imbalanced training corpora produce classifiers with measurably worse performance on under-represented subgroups, even when overall accuracy looks acceptable (Larrazabal et al., 2020). The same risk applies to PPG datasets that draw disproportionately from demographic backgrounds.

Generalizability across noise sources is a third practical concern. The Gaussian noise model used here is an approximation. Motion artifacts from physical activity, transient skin-sensor decoupling during sleep, and ambient light intrusion produce non-stationary perturbations whose statistical structure differs from independent additive noise. Embedding tamper-resistant logging mechanisms into the wearable data pipeline can support post-hoc investigation of such failure modes when they manifest in deployment (Xu et al., 2021). Pretraining the generator on a corpus of synthetic but realistic noise is a natural extension.

Regulatory considerations interact with the choice of rejection threshold. Many regulators require that an AI-assisted screening device behave conservatively in the face of uncertainty, deferring to human judgment whenever confidence falls below a defined level. Recent surveys of blockchain technology trends have argued that tamper-evident audit trails will be increasingly important in this regulatory context, particularly for cross-institutional data sharing (Zheng & Lu, 2022). The framework presented here makes the regulatory requirement operational: the manufacturer can document the calibration of the entropy score on a held-out test set and specify a rejection threshold that delivers a target sensitivity at a stipulated specificity.

Audits can then verify that post-deployment confidence distributions match the pre-deployment specification, with deviations triggering retraining or recalibration. This is a concrete improvement over the current state of practice, in which the absence of a principled confidence score forces a binary choice between deployment and indefinite human oversight.

We caution against using the entropy as a stand-alone safety mechanism. As the per-class reliability analysis showed, a small fraction of true atrial fibrillation cases is confidently misclassified as non-atrial fibrillation, and these cases would not be captured by an entropy-based filter. A robust deployment strategy should pair the entropy-driven rejection step with periodic random sampling of low-entropy negatives for clinician review, with anomaly detection on the generator output to flag waveforms that depart from the training distribution, and with mandatory follow-up electrocardiography on any patient whose risk factors warrant active monitoring regardless of the model's confidence.

Translating the framework into a deployable product also requires attention to operational issues that do not appear in the controlled laboratory setting. Wearable devices frequently encounter sensor failures, low-battery states, and intermittent wireless connectivity, all of which can produce input segments that lie far outside the training distribution. Industry 4.0 frameworks for integrated manufacturing and information pipelines have long emphasized the importance of treating such operational perturbations as a first-class design concern rather than as out-of-band exceptions (Lu, 2017). The classifier's posterior on such segments is generally near the decision boundary, and the entropy score will correctly identify them as low confidence. However, the failure mode is operational rather than physiological, and the appropriate response is not to defer to a cardiologist but to prompt the patient to reposition the device or to wait for a more reliable measurement.

Longitudinal aggregation is another opportunity for improving deployment behavior. A wrist-worn device produces many recordings per day, and the marginal value of any single uncertain recording is low when subsequent recordings can settle the question. Bayesian updating across

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

recordings is a natural way to implement this, and it is consistent with the decision-theoretic framework: the posterior on the patient's underlying state is updated with each new recording, weighted by its reliability, and a confirmatory referral is generated when the posterior crosses a clinically meaningful threshold.

Costs of false positives and false negatives in atrial fibrillation screening are not symmetric. A false negative, in the absence of compensating safeguards, exposes the patient to ongoing stroke risk; a false positive triggers a confirmatory measurement that, while not free, is considerably cheaper than an avoidable cerebrovascular event. The framework presented here computes expected cost under a symmetric loss for analytic clarity, but a production deployment should be configured with the actual cost ratio appropriate to the local health system. Concretely, if the cost of a false negative is c times the cost of a false positive, the rejection threshold should be lowered in proportion so that recordings near the decision boundary are preferentially deferred to clinicians.

5. CONCLUSION

We have presented a clinical confidence modeling framework that integrates a one-dimensional Pix2pix-style generative denoiser, a pretrained AlexNet-1D atrial fibrillation classifier, and an entropy-based selective rejection module into a single decision-theoretic pipeline. The contribution of the framework is not in any individual component but in the way the components are evaluated jointly. Generator outputs are not assessed against a generic similarity metric; they are assessed by their effect on the downstream classification task, and the predictive entropy of that classifier is treated as a clinical confidence score that minimizes expected decision cost.

Our experiments on a balanced split of the Deep beat dataset show that this framework recovers most of the discriminative performance lost to ambulatory noise and that selective rejection of the lowest-confidence quartile recovers the remainder. The Uncertainty Calibration Error on the denoised condition was lower than on either the noisy or the clean baseline, an unexpected but reassuring finding that highlights the regularization benefits of the pipeline. Per-class reliability analysis revealed an asymmetry in the treatment of atrial fibrillation versus non-atrial fibrillation examples, which constrains the use of entropy as a sole safety mechanism and motivates the use of complementary safeguards in deployment.

Several limitations bound the scope of these conclusions. The Gaussian noise model is a simplification of real ambulatory artifacts, the dataset reflects a single recording platform, and calibration was not audited per demographic subgroup. The framework itself, however, is platform independent. Any combination of a generative restorer and a discriminative downstream task can be evaluated under the same risk-minimization lens, with the predictive entropy of the downstream task playing the role of clinical confidence. Emerging research directions in quantum machine learning may eventually offer new computational substrates for these pipelines, but the decision-theoretic framing is independent of the underlying computational paradigm (Lu et al., 2024).

We expect this perspective to be especially useful in wearable cardiology, where the gap between research-grade and consumer-grade signals is wide and where the cost of a missed diagnosis is high. Future work will extend the framework to realistic noise distributions, to multi-class rhythm classification, and to longitudinal monitoring scenarios where confidence is integrated over multiple recordings from the same patient.

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301

More broadly, we view the contribution of this work as a small but concrete step toward making generative models auditable in safety-critical clinical settings. Generative components are increasingly common in medical AI pipelines, from contrast synthesis in magnetic resonance imaging to missing-modality completion in multimodal diagnostics, and the question of how to verify their outputs without ground truth is rapidly becoming a barrier to deployment. The decision-theoretic framing offered here suggests one principled answer: do not attempt to verify the generator in isolation, but evaluate the entire pipeline by the consequences of its decisions. When the consequences can be quantified, calibration becomes a meaningful and auditable property of the system, and the confidence scores it produces become trustworthy in the only sense that matters clinically — that acting on them yields better outcomes than ignoring them.

Reference

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3), R1-R39. DOI:10.1088/0967-3334/28/3/R01.

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

Antun, V., Renna, F., Poon, C., Adcock, B., & Hansen, A. C. (2020). On instabilities of deep learning in image reconstruction and the potential costs of AI. Proceedings of the National Academy of Sciences, 117(48), 30088-30095. DOI:10.1073/pnas.1907377117.

Aschbacher, K., Yilmaz, D., Kerem, Y., Crawford, S., Benaron, D., Liu, J., ... Olgin, J. E. (2020). Atrial fibrillation detection from raw photoplethysmography waveforms: A deep learning application. Heart Rhythm O2, 1(1), 3-9. DOI:10.1016/j.hroo.2020.02.002.

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

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

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... Virani, S. S. (2019). Heart disease and stroke statistics-2019 update: A report from the American Heart Association. Circulation, 139(10), e56-e528. DOI:10.1161/CIR.0000000000000659.

Bhadra, S., Kelkar, V. A., Brooks, F. J., & Anastasio, M. A. (2021). On hallucinations in tomographic image reconstruction. IEEE Transactions on Medical Imaging, 40(11), 3249-3260. DOI:10.1109/TMI.2021.3077857.

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. DOI:10.3389/fnrgo.2021.805573.

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

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care - addressing ethical challenges. New England Journal of Medicine, 378(11), 981-983. DOI:10.1056/NEJMp1714229.

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

Charlton, P. H., Allen, J., Bailon, R., Baker, S., Behar, J. A., Chen, F., ... Marozas, V. (2023). The 2023 wearable photoplethysmography roadmap. Physiological Measurement, 44(11), 111001. DOI:10.1088/1361-6579/acead2.

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

Chugh, S. S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E. J., ... Murray, C. J. L. (2014). Worldwide epidemiology of atrial fibrillation: A Global Burden of Disease 2010 study. Circulation, 129(8), 837-847. DOI:10.1161/CIRCULATIONAHA.113.005119.

Diederichsen, S. Z., Haugan, K. J., Brandes, A., Lanng, M. B., Graff, C., Krieger, D., ... Svendsen, J. H. (2019). Natural history of subclinical atrial fibrillation detected by implanted loop recorders. Journal of the American College of

Cardiology, 74(22), 2771-2781. DOI:10.1016/j.jacc.2019.09.050.

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

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

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

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050-1059). DOI:10.5555/3045390.3045502.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4878-4887). DOI:10.5555/3294996.3295241.

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial Intelligence in Healthcare (pp. 295-336). Academic Press. DOI:10.1016/B978-0-12-818438-7.00012-5.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial networks. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672-2680). DOI:10.5555/2969033.2969125.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321-1330). DOI:10.5555/3305381.3305518.

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

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65-69. DOI:10.1038/s41591-018-0268-3.

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

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

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J. J., Blomstrom-Lundqvist, C., ... Watkins, C. L. (2021). 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). European Heart Journal, 42(5), 373-498. DOI:10.1093/eurheartj/ehaa612.

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

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 448-456). DOI:10.5555/3045118.3045167.

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

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. DOI:10.1145/3571730.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574-5584). DOI:10.5555/3295222.3295309.

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

Kornej, J., Borschel, 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. DOI:10.1161/CIRCRESAHA.120.316340.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (Vol. 25, pp. 1097-1105). DOI:10.5555/2999134.2999257.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402-6413). DOI:10.5555/3295222.3295387.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences, 117(23), 12592-12594. DOI:10.1073/pnas.1919012117.

Laves, M.-H., Ihler, S., Kortmann, K.-P., & Ortmaier, T. (2020). Uncertainty calibration error: A new metric for multi-class classification. In Medical Imaging with Deep Learning (Short Papers). DOI:10.48550/arXiv.1912.08768.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094-1111. DOI:10.1080/01621459.2017.1307116.

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. DOI:10.1177/1747493019897870.

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

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. DOI:10.1080/23270012.2019.1570365.

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

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

Lubitz, S. A., Faranesh, A. Z., Selvaggi, C., Atlas, S. J., McManus, D. D., Singer, D. E., ... Dewland, T. A. (2022). Detection of atrial fibrillation in a large population using wearable devices: The Fitbit Heart Study. Circulation, 146(19), 1415-1424. DOI:10.1161/CIRCULATIONAHA.122.060291.

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

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

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1). DOI:10.1609/aaai.v29i1.9602.

Nemati, S., Ghassemi, M. M., Ambai, V., Isakadze, N., Levantsevych, O., Shah, A., & Clifford, G. D. (2018). Monitoring and detecting atrial fibrillation using wearable technology. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3394-3397). IEEE. DOI:10.1109/EMBC.2016.7591456.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. DOI:10.1126/science.aax2342.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems (Vol. 32, pp. 8024-8035). DOI:10.5555/3454287.3455008.

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

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., ... Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909-1917. DOI:10.1056/NEJMoa1901183.

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

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2008). Dataset shift in machine learning. MIT Press. DOI:10.7551/mitpress/9780262170055.001.0001.

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

Sangha, V., Mortazavi, B. J., Haimovich, A. D., Ribeiro, A. H., Brandt, C. A., Jacoby, D. L., ... Khera, R. (2022). Automated multilabel diagnosis on electrocardiographic images and signals. Nature Communications, 13(1), 1583. DOI:10.1038/s41467-022-29153-3.

Schnabel, R. B., Yin, X., Gona, P., Larson, M. G., Beiser, A. S., McManus, D. D., ... Benjamin, E. J. (2015). 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: A cohort study. The Lancet, 386(9989), 154-162. DOI:10.1016/S0140-6736(14)61774-8.

Shashikumar, S. P., Shah, A. J., Clifford, G. D., & Nemati, S. (2018). Detection of paroxysmal atrial fibrillation using

attention-based bidirectional recurrent neural networks. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 715-723). DOI:10.1145/3219819.3219912.

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. New England Journal of Medicine, 383(25), 2477-2478. DOI:10.1056/NEJMc2029240.

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

Steinhubl, S. R., Waalen, J., Edwards, A. M., Ariniello, L. M., Mehta, R. R., Ebner, G. S., ... 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. DOI:10.1001/jama.2018.8102.

Svennberg, E., Friberg, L., Frykman, V., Al-Khalili, F., Engdahl, J., & Rosenqvist, M. (2021). Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): A multicentre, parallel group, unmasked, randomised controlled trial. The Lancet, 398(10310), 1498-1506. DOI:10.1016/S0140-6736(21)01637-8.

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., ... Marcus, G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiology, 3(5), 409-416. DOI: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. DOI:10.1038/s41591-018-0300-7.

Torres-Soto, J., & Ashley, E. A. (2020). Multi-task deep learning for cardiac rhythm detection in wearable devices. NPJ Digital Medicine, 3(1), 116. DOI:10.1038/s41746-020-00320-4.

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

Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer. DOI:10.1007/b106715.

Wang, J., Wang, P., & Wang, S. (2019). Automated detection of atrial fibrillation in ECG signals based on wavelet packet transform and correlation function of random process. Biomedical Signal Processing and Control, 55, 101662. DOI:10.1016/j.bspc.2019.101662.

Wolf, P. A., Abbott, R. D., & Kannel, W. B. (1991). Atrial fibrillation as an independent risk factor for stroke: The Framingham Study. Stroke, 22(8), 983-988. DOI:10.1161/01.STR.22.8.983.

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

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

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. DOI:10.1155/2021/5527904.

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. DOI:10.1080/17517575.2024.2541199.

Ye, Z., & Lu, Y. (2022). Quantum science: A review and current research trends. Journal of Management Analytics, 9(3), 383-402. DOI:10.1080/23270012.2022.2089064.

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

Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996-1015. DOI:10.1002/sres.3120.

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

ISSN: 3068-1197
https://inatgi.in/index.php/jaihbe/index | DOI: 10.69987/jaihbe.2025.030301