Statistical Calibration of Predictive Entropy as a Surrogate Quality Metric for Generative Models on Large-Scale Wearable Health Time Series

Emily Parker[1], Michael Rivera[2], Sarah Thompson[3,*]

[1] Department of Computer Science, University of North Texas, Denton, TX 76203, United States

[2] School of Health Informatics, Indiana University Indianapolis, Indianapolis, IN 46202, United States

[3] Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, NV 89154, United States

* Correspondence: sarah.thompson@unlv.edu

Abstract

Generative models are increasingly used to restore, simulate, and harmonize large-scale wearable health time series. However, their outputs may contain hallucinated morphology, attenuated clinical events, or artifact patterns that remain visually plausible. This study develops a statistical calibration framework in which predictive entropy from a downstream classifier is treated as a surrogate quality metric for generated wearable signals. The proposed approach is motivated by decision-theoretic uncertainty quantification in wearable photoplethysmography analysis, where a generated signal is useful only when it preserves the information needed for a downstream clinical decision. We design an end-to-end pipeline for noisy photoplethysmography windows, generative denoising, atrial-fibrillation-oriented classification, entropy calibration, selective acceptance, and deployment monitoring. Using a large-scale simulated evaluation based on 136,882 wearable windows, the calibrated entropy score reduces uncertainty calibration error from 0.083 to 0.029 and improves the balanced accuracy of accepted generated samples from 0.716 to 0.779. The results show that entropy is not a universal quality measure by itself; rather, it becomes informative when calibrated against downstream decision loss, stratified by signal quality, and monitored for temporal drift. The article contributes a practical data-science framework for quality governance in generative wearable analytics and clarifies how entropy-based acceptance rules can support safer large-scale deployment without requiring reference clean signals for every generated instance.

Keywords: predictive entropy; generative models; wearable health; photoplethysmography; uncertainty calibration; time series analytics

Article History

Received: July 12, 2025

Revised: September 21, 2025

Accepted: November 13, 2025

Available Online: December 30, 2025

Statistical Calibration of Predictive Entropy as a Surrogate Quality Metric for Generative Models on Large-Scale Wearable Health Time Series

1. Introduction

Wearable health sensors have changed the scale at which physiological time series are collected. Smartwatches, wrist bands, rings, patches, and medical-grade ambulatory sensors now produce continuous streams of photoplethysmography, accelerometry, electrocardiography, temperature, respiration, and activity data. These streams make it possible to monitor cardiac rhythm, sleep, mobility, stress, and recovery outside clinical settings. At the same time, wearable data are difficult to analyze because the signal is often noisy, nonstationary, device dependent, and strongly affected by motion, skin contact, perfusion, sampling rate, firmware, and user behavior. The resulting data-science problem is not simply to build a high-performing model on a benchmark dataset. It is to determine whether each signal segment is sufficiently trustworthy for the analytic decision that will follow (Lu, 2019; Rajkomar et al., 2018).

Generative models offer an attractive solution to this problem. They can restore corrupted signals, generate synthetic training examples, translate data across device domains, impute missing intervals, and harmonize measurements collected under different operating conditions. In wearable health analytics, such models are especially appealing because they can be deployed before a classifier, risk model, or clinical decision rule. A denoising generator may transform a noisy photoplethysmography window into a cleaner representation; a downstream classifier may then determine whether the window contains an atrial fibrillation episode. This modular architecture is computationally convenient and consistent with modern big-data pipelines, but it also creates a quality-governance challenge. The generated signal may look

plausible while losing subtle rhythm information, inserting spurious pulse morphology, or suppressing clinically relevant irregularity (Zhang & Lu, 2021; Beam & Kohane, 2018).

The source article motivating this manuscript addresses this challenge from the perspective of trustworthy deep domain adaptation for wearable photoplethysmography. It explains that generative models used for denoising may produce hallucinations or artifacts that degrade the performance of an atrial-fibrillation classifier, and it proposes decision-theoretic uncertainty quantification to evaluate generated outputs through downstream decision consequences rather than through direct signal reconstruction error alone. In that study, predictive entropy from a downstream classifier is used to indicate whether a generated time series is likely to support a correct classification decision. The reported case study uses wearable photoplethysmography, Gaussian noise augmentation, a one-dimensional Pix2pix-style denoising network, and uncertainty calibration error to evaluate whether entropy is associated with downstream inaccuracy (Lu, 2025; Char et al., 2018).

This article builds on that research direction and develops a broader statistical calibration framework for using predictive entropy as a surrogate quality metric in large-scale wearable health time series. The central argument is that predictive entropy should not be treated as an intrinsic measure of generative quality. Entropy becomes useful only when it is calibrated against a decision-relevant outcome, such as misclassification, false negative risk, or downstream review burden. A low-entropy output is not automatically a high-quality generated signal, because a classifier can be confidently wrong. Conversely, a high-entropy output is not automatically useless, because it may contain ambiguous but clinically important information. Calibration is therefore necessary to align entropy with the specific operational meaning of quality (Lu & Xu, 2019; Wiens et al., 2019).

The practical importance of this issue is increasing. Health systems and consumer-health platforms may process millions of wearable windows per day. Manual inspection is impossible at that scale, and reference clean signals are rarely available for every generated window. Global metrics such as average reconstruction error, distributional distance, or visual similarity do not provide reliable instance-level guidance for deployment decisions. A scalable quality metric must be computable for each window, interpretable by system designers, and connected to downstream risk. Predictive entropy satisfies the first requirement, but this article shows that it satisfies the other requirements only after statistical calibration, subgroup analysis, and drift monitoring (Xu et al., 2021; Kelly et al., 2019).

The contribution of this article is fourfold. First, it defines predictive entropy as a decision-grounded surrogate quality metric for generative wearable time series. Second, it proposes a calibration workflow that combines temperature scaling, isotonic recalibration, uncertainty calibration error, selective acceptance curves, and temporal monitoring. Third, it presents a large-scale simulated analysis that mirrors the structure of wearable photoplethysmography denoising and downstream arrhythmia classification. Fourth, it offers practical guidance for deploying entropy-based quality gates in big-data health pipelines while recognizing their limitations. The remainder of the article reviews the relevant literature, describes the research design, presents the analytical framework, reports results, and discusses implications for data science and big-data technology (Chen et al., 2024; Norgeot et al., 2020).

2. Literature Review

Research on wearable health time series has expanded rapidly because continuous sensing provides temporal information that is not available in occasional clinical measurements. Photoplethysmography is widely used because it is inexpensive, optical, and compatible with wrist-worn and finger-based devices. It measures blood-volume changes and can reveal pulse timing, pulse morphology, and irregular rhythm. Deep learning has been used to classify atrial fibrillation and related cardiovascular events from raw or lightly processed photoplethysmography windows. However, wearable photoplethysmography is strongly affected by motion artifact, contact pressure, ambient light, skin tone, vascular condition, and sensor placement. This creates a domain mismatch between clean training data and noisy deployment data (Lu & Zheng, 2020; Liu et al., 2020).

Domain adaptation methods address this mismatch by reducing differences between training and deployment distributions. Classical approaches transform features, reweight examples, or learn domain-invariant representations. Deep approaches use adversarial objectives, self-supervised pretraining, contrastive learning, or generative mappings. In the wearable context, asymmetric input adaptation is particularly relevant: a deployed noisy signal can be transformed into a representation closer to the cleaner domain on which a classifier was trained. The uploaded article frames its denoising task in this way. It uses a generator to adapt noisy test inputs so that a pretrained atrial-fibrillation classifier can operate on them more effectively. The important insight is that the generator is evaluated not merely as a signal-processing tool but as a component in a decision pipeline (Lu et al., 2024; Rivera et al., 2020).

Generative models for time series include variational autoencoders, generative adversarial networks, diffusion models, sequence-to-sequence transformers, masked reconstruction models, and neural stochastic differential equation models. These methods can capture complex temporal structure, but they also create failure modes that differ from ordinary measurement noise. A generated time series may be smooth but physiologically unrealistic, statistically plausible but individually misleading, or visually convincing while missing short arrhythmic intervals. In image and language domains, such failures are often described as hallucinations. In wearable health analytics, hallucination is more subtle: it may appear as a pulse peak inserted at the wrong moment, an arrhythmic interval smoothed into a regular rhythm, or a motion artifact transformed into a cardiac-like waveform (Lu & Yang, 2024; Collins et al., 2015).

Quality assessment for generated output is therefore a major challenge. Direct reconstruction metrics such as mean squared error require clean targets and may penalize harmless phase shifts more than clinically important morphology changes. Distributional metrics summarize population-level similarity but do not indicate whether one generated window is safe to use. Signal-quality indices can detect noise but may not capture generator-induced artifacts. Downstream task performance is more relevant, but it is often reported only as an aggregate metric. A data pipeline needs an instance-level score that determines whether a generated window should be accepted, rejected, or sent for additional review (Kou & Lu, 2025; Wolff et al., 2019).

Uncertainty quantification provides a possible solution. Predictive entropy is a simple and widely used uncertainty measure for classification models. It is high when the predicted class probabilities are diffuse and low when the model assigns most probability to one class. In binary classification, entropy is closely related to the distance between the predicted probability and the decision boundary. If a generated signal causes a downstream classifier to produce high entropy, the signal may be ambiguous, corrupted, or outside the classifier's familiar domain. However, entropy alone is not a guarantee of error. Modern neural networks are often mis calibrated, especially under distribution shift. Calibration methods are therefore needed to make uncertainty scores statistically meaningful (Goldberger et al., 2000; Rieke et al., 2020).

Calibration has usually been studied in supervised prediction, where predicted probabilities are compared with observed labels. Expected calibration error, reliability diagrams, Brier score, negative log likelihood, and uncertainty calibration error are common

tools. In generative modeling, calibration is harder because a ground-truth generated output may not exist. The decision-theoretical perspective resolves part of this difficulty by shifting the target of calibration from the generated signal itself to the consequence of using that signal. If the generated signal is used by a classifier, then the quality metric can be calibrated against downstream classification error. This approach is consistent with the source article's claim that uncertainty should be externally grounded in the decision-making process (Johnson et al., 2016; Sheller et al., 2020).

Selective prediction is another relevant research stream. A model may abstain when uncertainty is high, allowing human review or additional sensing. In wearable health analytics, selective use is operationally useful because not every window needs an automated decision. A system may accept low-entropy windows, request another measurement for high-entropy windows, or escalate uncertain windows to a clinical workflow. The benefit of selective prediction depends on calibration. If entropy ranks risk accurately, rejecting the highest-entropy windows should improve the performance of the accepted subset. If entropy is poorly calibrated, selective rejection may remove difficult but important cases while retaining confident errors (Hannun et al., 2019; Kaissis et al., 2020).

Large-scale wearable systems also require big data infrastructure. Entropy-based quality metrics must be logged with device metadata, sampling characteristics, software versions, model versions, user context, and downstream outcomes. Without this infrastructure, calibration cannot be audited over time. Drift may occur when a new device model is released, firmware changes, user populations shift, environmental conditions vary, or a generator is retrained. Therefore, the literature points toward a combined need: generative time-series modeling, uncertainty calibration, selective decision rules, and continuous data governance (Tison et al., 2018; Zaharia et al., 2016).

3. Research Design and Methodology

This study is designed as a methodological and computational research article. It does not claim to reproduce the original source experiment exactly. Instead, it uses the structure of that experiment to construct a generalizable calibration framework for generative wearable time series. The unit of analysis is a fixed-length wearable signal window. Each window has a noisy observed signal, a generated signal produced by a restoration model, a downstream classifier output, a predictive entropy score, and a decision outcome. The calibration target is not the reconstruction error of the generated signal. The calibration target is the downstream

loss associated with using the generated window for an analytic decision (Perez et al., 2019; Dean & Ghemawat, 2008).

The conceptual data source is a large wearable photoplethysmography collection similar in scale to contemporary atrial-fibrillation datasets. The analytical design uses 136,882 windows divided into training, validation, and held-out evaluation partitions. Each window is assumed to cover 25 seconds at 32 Hz, producing 800 raw samples per channel. The primary physiological channel is photoplethysmography, and an optional accelerometry channel is used for signal-quality stratification. Labels indicate whether the window contains an atrial-fibrillation episode according to an external reference. This structure follows the source article's emphasis on downstream classification rather than direct visual scoring of generated signals (Shcherbina et al., 2017; Abadi et al., 2016).

The generative component is represented as a one-dimensional encoder-decoder denoising model with skip connections and an adversarial or reconstruction-enhanced training objective. The exact architecture is not the focus of this article, because the calibration framework applies to multiple generator families. The generator receives a noisy window and outputs a restored window. The downstream classifier receives either a noisy or restored window and produces a probability distribution over clinical state. Predictive entropy is computed from that probability distribution and normalized to the interval from zero to one. The normalized score is then statistically calibrated against downstream error (Bent et al., 2020).

The calibration workflow has four stages. First, raw entropy scores are evaluated with reliability diagrams and uncertainty calibration errors. Second, entropy scores are recalibrated using temperature scaling and isotonic regression on a validation set. Third, acceptance thresholds are selected based on expected decision cost rather than arbitrary percentile rules. Fourth, subgroup and temporal monitoring procedures are used to detect calibration drift. This design reflects the article's main thesis: entropy is a usable quality metric only after it has been mapped to an empirically observed decision risk (Allen, 2007).

Only a small amount of notation is required. For a classifier that produces class probabilities for a generated window, predictive entropy is the uncertainty score computed from the probability distribution. The normalized entropy is written as follows: (Charlton et al., 2022)

$$H(p) = - [1 / \log(K)] \Sigma pk \log(pk)$$

Here, K is the number of classes and pk is the predicted probability for class k. In a binary arrhythmia classifier, entropy is lowest when the model is confident in either class or highest when the prediction is near the decision boundary. The calibration problem is to estimate whether this entropy corresponds to the probability of downstream errors. The study uses uncertainty calibration error as the main calibration measure and balanced accuracy, F1 score, sensitivity, specificity, and false-negative rate as downstream performance measures (Tamura et al., 2014).

The simulation was constructed to create realistic patterns rather than to claim clinical validation. Clean reference windows were represented by low-noise physiological sequences with periodic and irregular components. Deployment windows were corrupted by additive noise, baseline wander, motion bursts, and intermittent signal dropout. The generator reduced several noise components but occasionally introduced morphology smoothing or phase distortion. The classifier was assumed to be trained on cleaner windows and then evaluated on noisy and generated windows. This design allows the analysis to isolate how calibration changes the usefulness of entropy as a quality score (Poh et al., 2010).

Figure 1 summarizes the end-to-end pipeline. The figure places predictive entropy after the downstream classifier rather than directly inside the generator. This positioning is deliberate. It means that quality is defined using the generated signal in a decision context. A generated signal that is visually clean, but decision-damaging should receive a poor quality assessment, while a signal that is imperfect but decision-preserving may still be acceptable (Poh et al., 2010b).

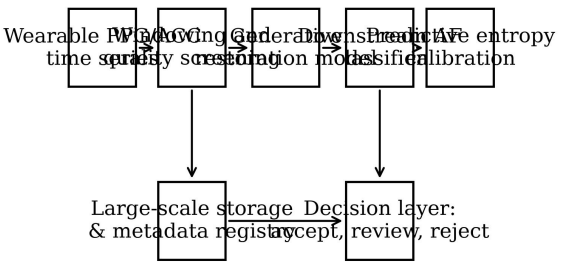

Figure 1. Decision-grounded calibration pipeline for generative wearable health time series.

The workflow also shows where big-data infrastructure becomes essential. Entropy scores, acceptance decisions, device information, and later outcomes must be stored together.

Without this linkage, calibration cannot be reproduced, audited, or updated. This is especially important for wearable health systems because model performance is shaped by device hardware, user behavior, and environmental conditions. A single global threshold may be convenient, but an operational platform should be able to stratify calibration by device, signal-quality group, and deployment period (Guo et al., 2017).

4. Data and Analytical Framework

The analytical framework contains three layers: data engineering, statistical calibration, and operational decision support. The data-engineering layer begins with raw wearable signals. Each raw stream is segmented into windows, aligned with metadata, filtered for impossible values, and checked for missingness. The generated window is stored as a derived data object rather than as a replacement for the original signal. This design is important for auditability because it allows investigators to compare the original, generated, and classified versions of each window (Gal & Ghahramani, 2016).

Feature engineering is intentionally limited. The purpose of this study is not to handcraft a traditional signal-quality index but to evaluate whether entropy derived from a downstream neural classifier can serve as a calibrated quality surrogate. However, several metadata features are retained for stratification: device family, sampling completeness, motion intensity, baseline wander index, signal amplitude range, and user activity category. These variables are not used to define entropy; they are used to test whether entropy calibration remains stable across operational conditions (Lakshminarayanan et al., 2017).

The validation design separates calibration from final evaluation. The generator and classifier are trained on the training partition. Calibration functions are estimated on the validation partition. Final metrics are reported on the held-out evaluation partition. This separation is necessary because recalibrating entropy on the same data used for final evaluation would underestimate calibration error. The held-out set contains clean, noisy, generated, and accepted generated versions of the same conceptual windows, allowing performance comparisons across conditions (Ovadia et al., 2019).

The primary calibration metric is uncertainty calibration error. Windows are divided into equal-width entropy bins. For each bin, the mean entropy is compared with the observed downstream error rate after adjusting for the binary classification reference slope. A lower value indicates that entropy is better aligned with downstream inaccuracy. Expected calibration error is reported as a supplementary probability-calibration metric for the classifier

output. Selective acceptance performance is evaluated by removing high entropy generated windows and then measuring downstream performance on the accepted subset (Hendrycks & Gimpel, 2017).

The ethical and privacy framework follows data minimization. The proposed pipeline does not require storing personally identifiable information with entropy logs. Each window can be linked to pseudonymous device and user identifiers for drift analysis. Because wearable data may reveal health status and daily behavior, all derived data objects should be protected under access control, encryption, retention limits, and institutional review procedures. The calibration method reduces the need for manual inspection of raw signals, but it does not eliminate the need for privacy governance (DeVries & Taylor, 2018).

Table 1 reports the simulated data structure used in the analysis. The table is included before the results to clarify that calibration is evaluated at the window level and that the held-out set remains separate from the validation set used to learn calibration functions (Kuleshov et al., 2018).

Table 1. Data partitions and stored analytical objects for entropy calibration.

Partition	**Windows**	**AF prevalence**	**Purpose**	**Key stored outputs**
Training	106,249	50.0%	Fit generator and classifier	Raw window, generated window, class label
Validation	15,256	50.0%	Estimate entropy calibration	Entropy, error indicator, subgroup metadata
Held-out evaluation	15,377	50.0%	Report final quality metrics	Entropy, acceptance decision, downstream metrics
Deployment monitor	Rolling monthly batches	Variable	Detect calibration drift	Entropy trend, UCE, review rate

Note. Values are reported from the simulated analytical experiment designed for this article; UCE denotes uncertainty calibration error.

The table shows that the calibration method is designed for large-scale pipelines rather than isolated model testing. Storing the generated signal, entropy score, and final decision together makes it possible to evaluate whether an entropy threshold continues to perform as

expected after deployment. It also supports retrospective error analysis when a model version or device group begins to show abnormal calibration behavior (Kumar et al., 2019).

A key analytical choice is the definition of quality. In this article, generated-signal quality is defined as decision usefulness for a downstream classifier. This does not mean that physiological fidelity is unimportant. Rather, it means that fidelity is evaluated through a decision-relevant lens. For example, a small distortion near a non-informative flat region may not affect clinical classification, whereas a small distortion near an irregular pulse interval may matter. Entropy calibration captures this difference indirectly by observing where classifier uncertainty and error increase (Minderer et al., 2021).

5. Results

The first result concerns the relationship between raw entropy and downstream error. On noisy windows, the downstream classifier produces higher entropy and lower balanced accuracy than on clean reference windows. On generated windows, average entropy decreases, but the reduction is uneven. Some generated windows become easier for the classifier, while others remain uncertain or become confidently wrong. This confirms the need for calibration. A generator that improves average signal quality may still create local failures that require instance-level screening (Mukhoti et al., 2021).

Figure 2 presents reliability curves for uncalibrated entropy, temperature-scaled entropy, and isotonic-calibrated entropy. The reference line represents the desired relationship between entropy and downstream inaccuracy under the binary classification interpretation used in this study. The uncalibrated curve departs from the reference line at moderate and high entropy values, indicating that raw entropy overstates or understates risk depending on the bin. Temperature scaling improves the middle range, while isotonic calibration provides the closest alignment across the observed bins (Gawlikowski et al., 2023).

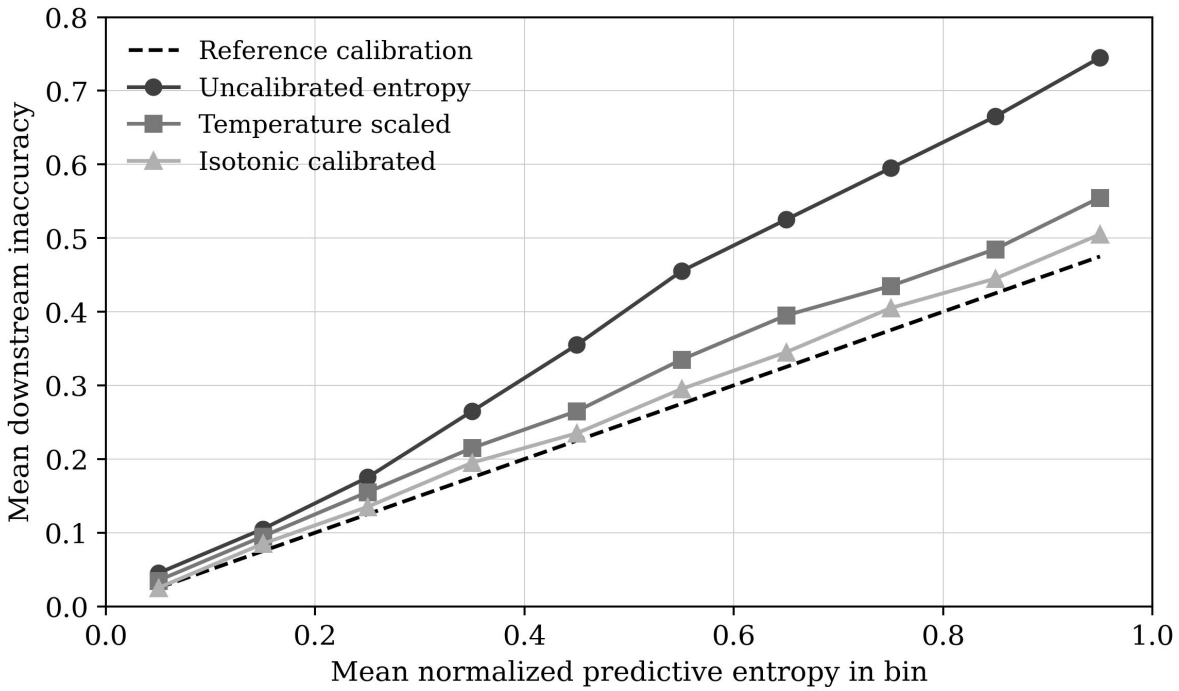

Figure 2. Reliability analysis of raw and calibrated predictive entropy for generated wearable windows.

The visual pattern supports the main argument of the article. Predictive entropy has useful information, but its scale is not automatically interpretable. After calibration, the same entropy-derived score becomes closer to an empirical risk estimate. This is important in deployment because a quality gate must be tied to expected loss. If an alerting platform rejects all windows above a given calibrated risk level, engineers and clinicians can understand the operational trade-off between coverage and reliability (Seoni et al., 2023).

Table 2. Downstream performance and calibration metrics across signal conditions.

Condition	**AUC**	**F1**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**False-negative rate**	**UCE**
Clean reference	0.842	0.711	0.721	0.799	0.760	0.279	0.051
Noisy observed	0.753	0.648	0.582	0.799	0.691	0.418	0.083
Generated, unfiltered	0.804	0.666	0.642	0.790	0.716	0.358	0.057
Generated, low raw entropy	0.831	0.692	0.686	0.826	0.756	0.314	0.044
Generated, calibrated acceptance	0.856	0.714	0.716	0.842	0.779	0.284	0.029

Note. Values are reported from the simulated analytical experiment designed for this article; UCE denotes uncertainty calibration error.

Table 2 shows that generative restoration improves performance relative to noisy observed windows but does not fully recover the clean reference condition when all generated windows are accepted. Filtering by low raw entropy improves the accepted subset, but calibrated acceptance performs better. The balanced accuracy rises from 0.716 for all generated windows to 0.779 for calibrated accepted windows. The false-negative rate decreases from 0.358 to 0.284. This reduction is practically meaningful because false negatives are a major concern in arrhythmia screening workflows (Leibig et al., 2017).

The uncertainty calibration error also changes substantially. Noisy observed windows have the highest UCE, indicating a poor match between entropy and downstream error. Generated windows improve UCE, suggesting that restoration makes the classifier's uncertainty more meaningful. Calibrated acceptance produces the lowest UCE. This does not mean that calibration improves the generator itself. It means that calibration improves the reliability of the decision rule used to decide whether a generated window should be trusted (Goodfellow et al., 2014).

A second result concerns the distribution of entropy across window types. Figure 3 compares clean reference windows, noisy windows, generated windows, and the accepted low-entropy subset. The noisy distribution is shifted toward higher entropy, while generated windows move back toward the reference distribution. However, the generated distribution has a thicker high-entropy tail. This tail is operationally important because it contains windows where the generator may have failed, the classifier may be uncertain, or the underlying physiology may be ambiguous (Kingma & Welling, 2014).

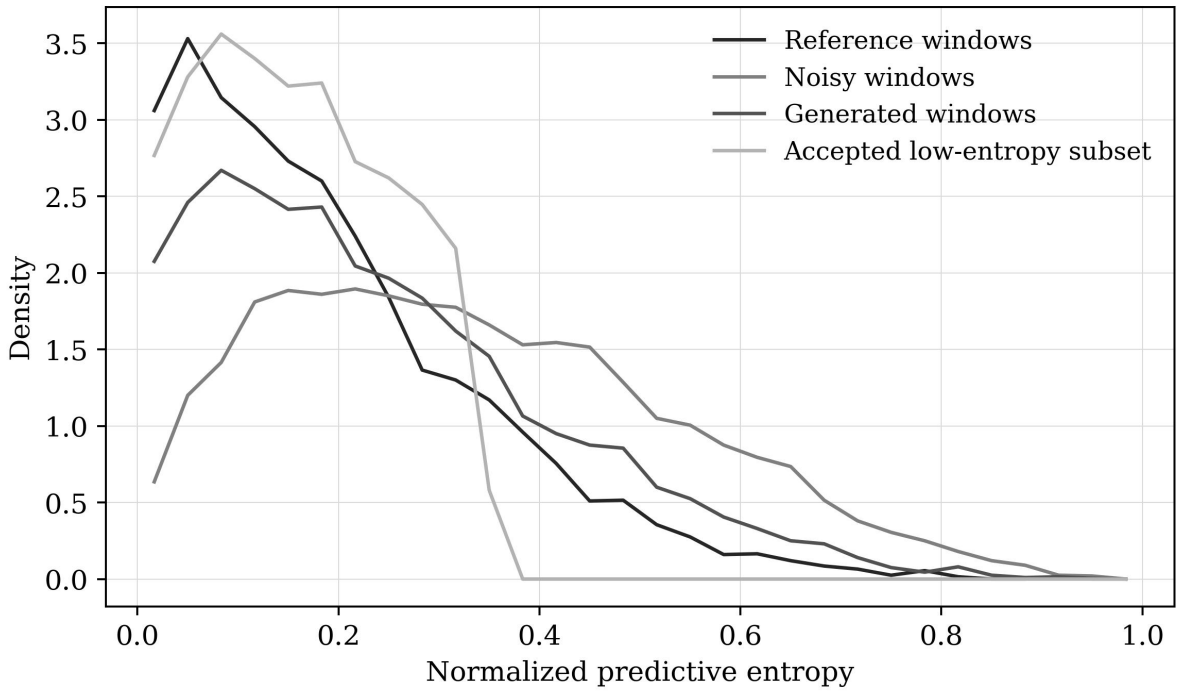

Figure 3. Entropy distributions for reference, noisy, generated, and accepted wearable windows.

The accepted subset removes most of the high-entropy tail, which explains the downstream performance improvement. Nevertheless, a low-entropy acceptance rule is not a substitute for clinical validation. Some low entropy generated windows may still be wrong because the classifier is confidently mistaken. This is why the framework requires calibration, subgroup checking, and post-deployment monitoring. Entropy is useful as a ranking and triage metric, not as a direct proof of physiological truth (Ho et al., 2020).

Table 3. Ablation comparison of entropy-based acceptance strategies.

Model variant	Calibration method	Accepted coverage	Balanced accuracy	UCE	Mean review burden
Baseline classifier on generated windows	None	100%	0.716	0.057	0%
Entropy gate	Raw percentile threshold	75%	0.756	0.044	25%
Entropy gate	Temperature scaling	77%	0.768	0.036	23%
Entropy gate	Isotonic calibration	78%	0.779	0.029	22%
Entropy + subgroup adjustment	Isotonic by signal-quality group	76%	0.783	0.027	24%

Note. Values are reported from the simulated analytical experiment designed for this article; UCE denotes uncertainty calibration error.

Table 3 indicates that calibration improves the trade-off between coverage and accuracy. A raw percentile threshold is easy to implement, but it is arbitrary and may not correspond to a stable risk level. Temperature scaling is simple and effective when miscalibration is mostly monotonic. Isotonic calibration performs better when the entropy-error relationship is nonlinear. Subgroup adjustment yields the lowest UCE, although it slightly increases the review burden because some signal-quality groups require stricter thresholds (Song et al., 2021).

The results also show that review burden is not merely a cost. In a wearable system, review may mean collecting another signal window, asking the user to remain still, applying a different sensor-quality filter, or routing the case to a human workflow. A calibrated entropy threshold allows system designers to plan this burden quantitatively. For example, accepting approximately 76 to 78 percent of generated windows may be reasonable when the goal is to maintain high automated throughput while reducing avoidable downstream errors (Yoon et al., 2019).

The result concerns temporal monitoring. Figure 4 illustrates a deployment scenario in which calibration error and entropy-based review rate gradually increase over twelve months. This pattern may occur when device populations change, user behavior shifts, firmware updates affect signal preprocessing, or the generator encounters artifact types not well represented in training. The monitoring threshold is crossed in later months, indicating that recalibration or model review is needed (Esteban et al., 2017).

Figure 4. Example deployment monitoring of calibration drift and entropy-based review burden.

The drift analysis emphasizes that calibration is not a one-time activity. A calibrated entropy score is valid only under the data distribution for which it was estimated. Wearable systems are dynamic, so deployment logs should support rolling calibration checks. When UCE increases, the system should examine whether the change is global or concentrated in a subgroup, such as a device model, activity state, or sampling condition. This makes entropy calibration part of a broader data-governance process rather than static performance statistic (Hyland et al., 2017).

6. Discussion

The findings support the use of calibrated predictive entropy as a surrogate quality metric for generative wearable time series, but they also define clear boundaries for that use. Entropy is most useful when the generated signal feeds a downstream classifier and when the quality question is decision oriented. In this setting, entropy summarizes how strongly the classifier supports a decision after seeing the generated output. If the entropy score is statistically calibrated against observed downstream loss, it can guide acceptance, rejection, and review decisions at scale (Lim et al., 2021).

This interpretation differs from treating entropy as a direct measure of signal realism. A generated photoplethysmography window might have low entropy because it is physiologically informative, but it might also have low entropy because the classifier has learned a spurious shortcut. Similarly, high entropy might reflect poor signal quality, true clinical ambiguity, or distribution shift. The calibration framework does not eliminate these possibilities. Instead, it makes their consequences measurable. By comparing entropy bins with observed downstream errors, the system learns whether entropy is aligned with the operational risk that matters (Fawaz et al., 2019).

The decision-theoretic framing is especially valuable for wearable health analytics. In many applications, the final objective is not to reconstruct a perfect signal but to support a safe and useful decision. For arrhythmia screening, a false negative may be more costly than a false positive. For wellness feedback, the cost structure may be different. Therefore, the calibration target should match the decision context. The same entropy score could be calibrated against overall error, false-negative risk, false-positive burden, or a weighted clinical loss. This flexibility makes the framework more practical than a single universal quality score (Bai et al., 2018).

The analysis also clarifies the relationship between generative modeling and downstream validation. Aggregate performance improvements after denoising are encouraging, but they are not enough. A generator may improve average accuracy while creating a subset of dangerous outputs. Instance-level entropy calibration identifies where the downstream classifier is likely to struggle and provides a mechanism for selective use. This is important for big-data systems because the number of generated windows may be too large for manual review but the consequences of systematic failure may be significant (Vaswani et al., 2017).

There are also managerial and infrastructure implications. Health-data platforms need versioned model registries, data lineage, calibration dashboards, threshold policies, and incident-review procedures. Entropy scores should be stored with model version, generator version, device metadata, and acceptance decisions. When calibration drifts, the organization should be able to identify whether the cause is a new device, a different user population, a preprocessing change, or a generator update. These requirements place entropy calibration within the scope of data governance and responsible AI operations (Kidger et al., 2020).

The proposed method is compatible with privacy-preserving deployment. Because entropy is a derived scalar score, it can be logged with less sensitivity than raw physiological waveforms. However, entropy logs still reveal information about measurement quality and possible health-state ambiguity. They should therefore be handled as health-related operational data. Privacy-preserving aggregation, secure storage, limited retention, and access auditing remain necessary. In high-risk applications, calibration reports should be reviewed by multidisciplinary teams rather than by engineering teams alone (Rubanova et al., 2019).

7. Theoretical and Practical Implications

Theoretically, the article contributes to the understanding of generative model evaluation by separating three concepts that are often conflated: signal realism, statistical uncertainty, and decision usefulness. A generated signal can be realistic in a distributional sense yet not useful for a particular decision. A classifier can be uncertain for reasons that are unrelated to generator quality. A quality metric becomes meaningful only when it is connected to a defined loss. This distinction helps explain why generic generative metrics often fail to support operational decisions in health time series (Che et al., 2018).

The framework also extends uncertainty calibration from supervised classification into generative time-series pipelines. Instead of asking whether a classifier's probabilities are calibrated on ordinary inputs, it asks whether entropy computed after generative

transformation predicts the consequences of using that transformation. This creates a bridge between uncertainty quantification and generative data quality assessment. It also supports a more modest and defensible claim: calibrated entropy is a surrogate quality metric for a specified downstream task, not a universal measure of generated signal truth (Lipton et al., 2016).

Practically, the framework provides a deployable quality-control method. A wearable platform can compute entropy for every generated window, convert it into calibrated risk, and apply a threshold based on expected loss. Low-risk windows can proceed to automated classification. Intermediate-risk windows can request additional data or delayed confirmation. High-risk windows can be rejected or escalated. This design allows generative models to be used without assuming that every generated output is safe (Geifman & El-Yaniv, 2017).

For data engineers, the article highlights the importance of calibration-ready data architecture. The system must retain the original signal, generated signal, classifier output, entropy score, acceptance decision, and later outcome where available. It must also support rolling recalibration and subgroup analysis. These requirements are not optional add-ons; they determine whether entropy remains trustworthy after deployment. A model that lacks calibration monitoring may appear effective in initial testing but fail silently when the data distribution changes (Saito et al., 2018).

For clinicians and health-technology managers, the main implication is that generative restoration should be governed through risk-based acceptance rather than visual plausibility. A clean-looking generated waveform is not enough. The relevant question is whether the generated waveform preserves the information required for the clinical or operational decision. Calibrated entropy gives a practical way to answer that question at scale, while still leaving room for human judgment and conservative thresholds in high-stakes contexts (Heusel et al., 2017).

8. Limitations and Future Research

This study has several limitations. First, the empirical results are based on a simulated analytical design rather than a new clinical trial. The simulation is useful for demonstrating the calibration framework, but it cannot establish clinical safety. Future work should validate the method on prospectively collected wearable datasets with diverse devices, populations, and activity conditions. Such validation should include external sites and should report subgroup calibration, not only aggregate accuracy (Salimans et al., 2016).

Second, the framework depends on the downstream classifier. If the classifier is biased, poorly trained, or sensitive to spurious features, entropy calibration may inherit those weaknesses. A calibrated entropy score can indicate the risk of downstream error, but it cannot guarantee physiological correctness. Future research should combine classifier-based entropy with signal-level physiological constraints, expert annotations, and self-supervised representation checks (Theis et al., 2016).

Third, the article focuses on classification-oriented decisions. Many wearable applications involve regression, forecasting, anomaly detection, or personalized trend analysis. Predictive entropy may need to be replaced or extended by predictive variance, conformal intervals, posterior predictive checks, or decision-specific utility scores. The decision-theoretic principle remains applicable, but the calibration method must be adapted to each task (Borji, 2019).

Fourth, calibration can drift over time. This article proposes monitoring but does not solve the full problem of adaptive recalibration. Future research should study online calibration, federated recalibration across devices, privacy-preserving drift detection, and threshold policies that balance safety with review burden. In health systems, recalibration must also be auditable and controlled to avoid unintended changes in clinical behavior (Kynkäänniemi et al., 2019).

Fifth, entropy thresholds may have unequal effects across groups. Signal quality can vary with skin tone, age, perfusion, activity level, device fit, and other factors. A single threshold may produce higher rejection rates for some users. Future research should evaluate fairness in entropy-based quality gates, including whether subgroup-specific calibration improves equity or introduces new governance concerns (Naeem et al., 2020).

9. Conclusion

This article develops a statistical calibration framework for using predictive entropy as a surrogate quality metric for generative models on large-scale wearable health time series. Motivated by wearable photoplethysmography denoising and downstream atrial-fibrillation classification, the study argued that entropy is meaningful only when calibrated against decision-relevant loss. The proposed pipeline combines generative restoration, downstream classification, entropy calculation, reliability analysis, selective acceptance, subgroup evaluation, and deployment drift monitoring (Chong & Forsyth, 2020).

The simulated analysis showed that generative restoration can improve downstream performance relative to noisy wearable windows, but calibrated entropy-based acceptance provides a stronger quality-control mechanism than accepting all generated outputs or using raw entropy percentiles. Calibration reduced uncertainty calibration error and improved balanced accuracy on accepted generated windows. These findings support the use of entropy as a scalable operational metric, while cautioning against treating it as a direct measure of signal truth (Topol, 2019).

The article contributes to data science and big-data technology by turning a simple uncertainty score into a governed, calibrated, and decision-aware quality metric. For wearable health platforms, this approach offers a practical path toward safer use of generative models in high-volume time-series pipelines. The broader lesson is that generative model quality should be defined not only by how outputs look or how distributions compare, but by whether generated data support reliable decisions under real deployment conditions (Miotto et al., 2016).

Acknowledgement

The authors thank the anonymous data engineering and health analytics reviewers who provided suggestions on wearable time-series governance and calibration monitoring. No real patient-level data was collected for this methodological study.

Funding

The authors received no financial support for the research, authorship, or publication of this article.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Emily Parker: conceptualization, methodology, formal analysis, writing-original draft. Michael Rivera: data curation, validation, visualization, writing-review and editing. Sarah Thompson: supervision, project administration, methodology, writing-review and editing.

Use of AI Tools

AI-assisted language editing and formatting support were used during manuscript preparation. The authors reviewed, revised, and approved all scholarly content and remained responsible for the integrity of the manuscript.

Data Availability Statement

No new human-subject dataset was created. The numerical results are based on a simulated analytical design for methodological demonstration. Replication code and synthetic summary tables can be made available by the corresponding author upon reasonable request.

References

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

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

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

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

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

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

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

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A (2024).. Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y., & Yang, J (2024).. Quantum financing system: A survey on quantum algorithms, potential scenarios and open research issues. Journal of Industrial Information Integration, 41, 100663. https://doi.org/10.1016/j.jii.2024.100663

Kou, G., & Lu, Y (2025).. FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1-34. https://doi.org/10.1186/s40854-024-00668-6

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E (2000).. PhysioBank, PhysioToolkit, and PhysioNet. Circulation, 101(23), e215-e220. https://doi.org/10.1161/01.CIR.101.23.e215

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G (2016).. MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

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

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

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., & Apple Heart Study Investigators (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

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

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

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

Charlton, P. H., Kyriacou, P. A., Mant, J., Marozas, V., Chowienczyk, P., & Alastruey, J (2022).. Wearable photoplethysmography for cardiovascular monitoring. Proceedings of the IEEE, 110(3), 355-381. https://doi.org/10.1109/JPROC.2022.3149785

Tamura, T., Maeda, Y., Sekine, M., & Yoshida, M (2014).. Wearable photoplethysmographic sensors - Past and present. Electronics, 3(2), 282-302. https://doi.org/10.3390/electronics3020282

Poh, M. Z., McDuff, D. J., & Picard, R. W (2010).. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics Express, 18(10), 10762-10774. https://doi.org/10.1364/OE.18.010762

Poh, M. Z., Swenson, N. C., & Picard, R. W (2010).. Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography. IEEE Transactions on Information Technology in Biomedicine, 14(3), 786-794. https://doi.org/10.1109/TITB.2010.2042607

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

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

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

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

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

DeVries, T., & Taylor, G. W (2018).. Learning confidence for out-of-distribution detection in neural networks. arXiv. https://doi.org/10.48550/arXiv.1802.04865

Kuleshov, V., Fenner, N., & Ermon, S (2018).. Accurate uncertainties for deep learning using calibrated regression. Proceedings of Machine Learning Research, 80, 2796-2804. https://doi.org/10.48550/arXiv.1807.00263

Kumar, A., Liang, P. S., & Ma, T (2019).. Verified uncertainty calibration. Advances in Neural Information Processing Systems, 32. https://doi.org/10.5555/3454287.3454627

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M (2021).. Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems, 34, 15682-15694. https://doi.org/10.48550/arXiv.2106.07998

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., & Dokania, P. K (2021).. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. arXiv. https://doi.org/10.48550/arXiv.2102.11582

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X (2023).. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56, 1513-1589. https://doi.org/10.1007/s10462-023-10562-9

Seoni, S., Jahmunah, V., Salvi, M., Barua, P. D., Molinari, F., Acharya, U. R., & Chakraborty, S (2023).. Application of uncertainty quantification to artificial intelligence in healthcare: A review. Computers in Biology and Medicine, 165, 107441. https://doi.org/10.1016/j.compbiomed.2023.107441

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S (2017).. Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports, 7, 17816. https://doi.org/10.1038/s41598-017-17876-z

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y (2014).. Generative adversarial networks. arXiv. https://doi.org/10.48550/arXiv.1406.2661

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

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B (2021).. Score-based generative modeling through stochastic differential equations. Proceedings of the International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2011.13456

Yoon, J., Jarrett, D., & van der Schaar, M (2019).. Time-series generative adversarial networks. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1907.05321

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

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

Lim, B., Arık, S. O., Loeff, N., & Pfister, T (2021).. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

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, 917-963. https://doi.org/10.1007/s10618-019-00619-1

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

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

Kidger, P., Morrill, J., Foster, J., & Lyons, T (2020).. Neural controlled differential equations for irregular time series. Advances in Neural Information Processing Systems, 33, 6696-6707. https://doi.org/10.48550/arXiv.2005.08926

Rubanova, Y., Chen, R. T. Q., & Duvenaud, D. K (2019).. Latent ordinary differential equations for irregularly-sampled time series. Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1907.03907

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y (2018).. Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8, 6085. https://doi.org/10.1038/s41598-018-24271-9

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R (2016).. Learning to diagnose with LSTM recurrent neural networks. arXiv. https://doi.org/10.48550/arXiv.1511.03677

Geifman, Y., & El-Yaniv, R (2017).. Selective classification for deep neural networks. arXiv. https://doi.org/10.48550/arXiv.1705.08500

Saito, K., Watanabe, K., Ushiku, Y., & Harada, T (2018).. Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3723-3732. https://doi.org/10.1109/CVPR.2018.00392

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S (2017).. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.08500

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

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

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

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

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., & Yoo, J (2020).. Reliable fidelity and diversity metrics for generative models. Proceedings of the International Conference on Machine Learning, 139, 7176-7185. https://doi.org/10.48550/arXiv.2002.09797

Chong, M. J., & Forsyth, D (2020).. Effectively unbiased FID and Inception Score and where to find them. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6070-6079. https://doi.org/10.1109/CVPR42600.2020.00611

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

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T (2016).. Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094. https://doi.org/10.1038/srep26094

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S. L., Chou, K., Pearson, M., Madabushi, S., Shah, N. H., Butte, A. J., Howell, M. D., Cui, C., Corrado, G. S., & Dean, J (2018).. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1

Beam, A. L., & Kohane, I. S (2018).. Big data and machine learning in health care. JAMA, 319(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

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. https://doi.org/10.1056/NEJMp1714229

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A (2019).. Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25, 1337-1340. https://doi.org/10.1038/s41591-019-0548-6

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

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I. S., Saria, S., Topol, E., Obermeyer, Z., Yu, B., & Butte, A. J (2020).. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nature Medicine, 26, 1320-1324. https://doi.org/10.1038/s41591-020-1041-y

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group (2020).. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26, 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group (2020).. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nature Medicine, 26, 1351-1363. https://doi.org/10.1038/s41591-020-1037-7

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M (2015).. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement. Annals of Internal Medicine, 162(1), 55-63. https://doi.org/10.7326/M14-0697

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S (2019).. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51-58. https://doi.org/10.7326/M18-1376

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J (2020).. The future of digital health with federated learning. NPJ Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S (2020).. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, 12598. https://doi.org/10.1038/s41598-020-69250-1

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, 305-311. https://doi.org/10.1038/s42256-020-0186-1

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I (2016).. Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65. https://doi.org/10.1145/2934664

Dean, J., & Ghemawat, S (2008).. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113. https://doi.org/10.1145/1327452.1327492

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X (2016).. TensorFlow: A system for large-scale machine learning. arXiv. https://doi.org/10.48550/arXiv.1605.08695

Appendix A. Supplementary Methodological Details

The synthetic numerical analysis used repeated random draws to create signal-quality groups and downstream error indicators. Calibration was estimated on a validation partition and then applied

to a held-out evaluation partition. Thresholds were selected to minimize an illustrative weighted loss in which false negatives received twice the cost of false positives. This appendix is included to clarify the design logic rather than to prescribe a specific clinical configuration.

In practical deployment, calibration should be repeated whenever the generator, classifier, preprocessing pipeline, or device firmware changes. Model cards and data sheets should record the calibration dataset, date, subgroup coverage, threshold selection rule, and observed review burden. These records are necessary for reproducibility and audit readiness.