Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 4989~5002
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.13968

Uncertainty-aware circular residual U-Net for electrocardiogram denoising under realistic nonstationary noise

Ahmad Rifai[1,2], Siti Nurmaini[3,4], Bhakti Yudho Suprapto[5], Radiyati Umi Partan[6], Firdaus[3,4]
[1]Doctoral Program in Engineering Sciences, Faculty of Engineering, Universitas Sriwijaya, Palembang, Indonesia
[2]Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[3]Department of Computer Systems, Intelligent System Research Group (IsysRG), Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[4]Artificial Intelligence-Medical Center of Excellence (AIMed-COE), Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[5]Department of Electrical Engineering, Faculty of Engineering, Universitas Sriwijaya, Palembang, Indonesia
[6]Division of Rheumatology, Department of Internal Medicine, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia

Article Info

Article history:

Received Jul 14, 2026
Revised Aug 27, 2026
Accepted Sep 12, 2026

Keywords:

Bayesian deep learning
Boundary artifact
Electrocardiogram denoising
Monte Carlo dropout
Morphology preservation
Predictive uncertainty

ABSTRACT

Electrocardiogram (ECG) signals in ambulatory monitoring are frequently degraded by nonstationary baseline wander (BW), muscle artifact (MA), and electrode-motion (EM) artifact. This study proposed an uncertainty-aware circular residual U-Net integrating circular convolution, residual-noise prediction, a morphology-aware boundary loss, and Monte Carlo (MC) dropout. Clean QT database segments were contaminated with Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) noise stress test database noise at input signal-to-noise ratios (SNR) from -6 to 24 dB. Evaluation on 2,560 test segments compared the proposed model with db4 wavelet denoising, convolutional neural network-denoising autoencoder (CNN-DAE), a zero-padded U-Net, and a circular U-Net. The proposed model achieved an output SNR of 11.6789 dB and waveform correlation of 0.9309, with the lowest root mean square error (RMSE), mean absolute error (MAE), percentage root-mean-square difference (PRD), derivative MAE, and boundary MAE of 0.1820, 0.1382, 32.5572, 0.0236, and 0.1678, respectively. Its peak-amplitude error of 0.1998 was comparable to the circular U-Net. Output SNR exceeded CNN-DAE and wavelet denoising by 3.1449 and 2.8925 dB. Predictive uncertainty strongly correlated with reconstruction error (r=0.8239, p<0.001), supporting its use as an indicator of reconstruction difficulty.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Siti Nurmaini
Artificial Intelligence-Medical Center of Excellence (AIMed-COE), Faculty of Computer Science
Universitas Sriwijaya
Palembang, South Sumatra, Indonesia
Email: siti_nurmaini@unsri.ac.id

1. INTRODUCTION

Electrocardiogram (ECG) signals are central to cardiovascular monitoring and diagnosis because they provide information about the electrical activity of the heart. In clinical and ambulatory environments, however, ECG recordings are often contaminated by baseline wander (BW), muscle artifacts (MA), power-line interference, and electrode-motion (EM) artifacts. These disturbances can distort clinically important waveform components, including the P wave, QRS complex, ST segment, and T wave, thereby compromising automated analysis and clinical interpretation. Deep neural networks have achieved

cardiologist-level performance for arrhythmia classification in ambulatory ECG recordings [1], reinforcing the importance of preserving waveform information during preprocessing.

Classical ECG denoising methods [2], including digital filtering and wavelet transforms [3], [4], remain widely used because they are computationally efficient and relatively easy to implement. Beyond wavelet-based approaches, hybrid and adaptive filtering methods have also been developed, including empirical mode decomposition combined with an adaptive switching mean filter [5] and a low-distortion adaptive Savitzky–Golay filter based on discrete curvature estimation [6]. Wavelet-based methods are particularly attractive because they represent ECG signals in both the time and frequency domains. Their performance, however, depends strongly on the selected mother wavelet, decomposition level, and thresholding strategy. Inappropriate parameter choices can over smooth sharp ECG components, attenuate diagnostically relevant features, or distort waveform morphology, especially under nonstationary noise.

Deep learning-based denoising has recently shown promising results by learning nonlinear mappings between noisy and clean ECG signals. Convolutional neural networks, denoising autoencoders, and U-Net-based architectures can learn hierarchical signal representations without manually designed filters. Most convolutional models nevertheless treat each ECG segment as a finite sequence and apply zero padding to preserve its length. This assumption creates artificial discontinuities at the beginning and end of a segment and may therefore produce boundary artifacts. Such artifacts are particularly problematic in continuous ECG monitoring, where independently reconstructed windows are often concatenated into longer signals.

A further limitation of many deep learning ECG denoising models [7]–[9] is that their outputs are deterministic. They produce a single reconstructed waveform without indicating how much confidence should be placed in the prediction. In medical signal processing [10], [11], uncertainty information is valuable because denoising errors can alter clinically meaningful waveform components. A model that estimates predictive uncertainty can identify regions or segments that require more cautious interpretation.

Recent uncertainty-aware ECG studies have primarily addressed diagnostic classification rather than noisy-to-clean waveform reconstruction. Vranken *et al*. [12], Aseeri [13], and Barandas *et al*. [14] evaluated uncertainty in ECG classification under internal, external, or dataset-shift conditions. Dumitru *et al*. [15] introduced probabilistic confidence intervals for beat-wise ECG filtering, while Nakanishi *et al*. [16] applied Monte Carlo (MC) dropout to missing-lead reconstruction. These approaches differ from the present problem because they do not jointly evaluate realistic nonstationary denoising, fixed-window boundary behavior, morphology preservation, and sample-derived uncertainty-error consistency.

This study proposes an uncertainty-aware circular residual U-Net for fixed-window ECG denoising. The framework combines circular convolution, residual-noise prediction, a morphology-aware boundary loss, and bottleneck MC dropout. It is evaluated using QT database-electrocardiogram (QTDB ECG) segments contaminated with NSTDB-derived baseline-wander, muscle-artifact, EM, and mixed noise at input signal-to-noise ratio (SNR) levels from −6 to 24 dB. Predictive uncertainty is assessed through its segment-level correlation with reconstruction error.

The principal contributions of this study are fourfold. First, a three-level circular residual U-Net is developed for fixed-window ECG denoising, using encoder channels of 32, 64, and 128, a 256-channel bottleneck, and a symmetric decoder. Second, a morphology-aware boundary loss combines global reconstruction, absolute-error, temporal-derivative, and segment-edge terms to promote waveform and boundary preservation. Third, the framework is evaluated under realistic NSTDB-based baseline-wander, muscle-artifact, EM, and mixed-noise conditions at input SNR levels ranging from −6 to 24 dB. Fourth, MC dropout [17], [18] is incorporated at the bottleneck to generate uncertainty-aware reconstructions, and the reliability of the resulting predictive standard deviation is assessed through its correlation with segment-level reconstruction error.

The remainder of this paper is organized as follows. Section 2 describes the datasets, noise simulation procedure, proposed architecture, loss function, uncertainty estimation method, baseline models, and evaluation metrics. Section 3 presents and discusses the experimental results. Section 4 concludes the paper and outlines directions for future work.

2. METHOD

2.1. Dataset and preprocessing

Clean ECG signals were obtained from QTDB [19], [20], while realistic noise was obtained from NSTDB [20], [21]. All signals were processed at 250 Hz, and channel index 0 of each QTDB recording was used. Each ECG segment contained 512 samples, corresponding to 2.048 s. Starting positions were sampled uniformly from valid recording locations using a seeded random-number generator; therefore, no fixed stride or overlap percentage was applied.

Each segment was normalized by subtracting its median and dividing by the 95th percentile of its absolute amplitude, with the standard deviation used when the scale was below 10^{-8}. Recordings were partitioned into 73 training, 16 validation, and 16 test records, producing 5,840, 640, and 2,560 segments, respectively. No recording contributed segments to more than one subset.

2.2. Simulation of nonstationary noise

BW, MA, and EM artifact noise recordings from NSTDB were resampled to 250 Hz when necessary, centered, and standardized. For each ECG segment, a contiguous 512-sample noise window was sampled from a valid random position using seed 42. Mixed noise was generated by combining independently sampled BW, MA, and EM segments with weights drawn from U(0.3,1.0), followed by centering and standardization. Let x_i and n_i denote the normalized clean ECG and selected noise samples. The noisy signal was generated as $y_i = +an_i$, where α was calculated from the clean and noise powers to achieve the specified target SNR. Input SNR values were sampled from −6, 0, 6, 12, 18, and 24 dB.

$$y_i = x_i + \alpha n_i \tag{1}$$

where y_i is the noisy ECG sample and α is the scaling factor used to obtain the specified input SNR. The clean-signal and unscaled-noise powers were calculated as (2):

$$P_x = \frac{1}{L}\sum_{i=1}^{L} x_i^2,\ P_n = \frac{1}{L}\sum_{i=1}^{L} n_i^2 \tag{2}$$

The scaling factor was computed from the target input SNR as (2a):

$$\alpha = \sqrt{\frac{P_x}{(P_n+\varepsilon)10^{\frac{SNR_{target}}{10}}}},\ \varepsilon = 10^{-8} \tag{2a}$$

One noise realization was generated for each clean ECG segment. Training and validation used mixed NSTDB noise, with 80 and 40 segments per record, respectively. Testing was generated separately for BW, MA, EM, and mixed-noise conditions using 40 segments per test record, yielding 640 segments per condition and 2,560 test segments in total. No additional normalization was applied after mixing. The actual input SNR was recalculated for every generated segment; the mean absolute target error was below 8.1×10^{-7} dB and the maximum absolute error was 2.3×10^{-5} dB.

2.3. Proposed circular residual U-Net

The proposed network followed a three-level one-dimensional U-Net encoder–decoder topology [22], with encoder levels containing 32, 64, and 128 channels, a 256-channel bottleneck, and decoder levels containing 128, 64, and 32 channels. Encoder blocks incorporated residual shortcuts following the residual-learning principle [23], with 1×1 projections when the channel dimensions differed. Each convolutional block contained two Conv1D layers with kernel size 3, stride 1, circular padding of 1, batch normalization [24], and rectified linear unit (ReLU) activation. Max pooling with kernel size 2 and stride 2 performed downsampling, while nearest-neighbor upsampling by a factor of 2 and skip-feature concatenation restored temporal resolution. MC dropout with probability 0.20 was applied only at the bottleneck. A final 1×1 convolution estimated the noise residual, which was subtracted from the noisy input to obtain the 512-sample denoised output. The denoised signal is then obtained by subtracting the estimated residual from the noisy input:

$$\hat{x}(t) = y(t) - f\theta(y(t)) \tag{3}$$

where $y(t)$ is the noisy ECG input, $f\theta(\cdot)$ is the learned residual mapping, and $\hat{x}(t)$ is the denoised ECG output. This formulation allows the model to focus on learning noise patterns rather than reconstructing the entire ECG waveform from the beginning.

Figure 1 summarizes the proposed architecture, including the circular convolutional encoder, MC dropout bottleneck, symmetric decoder, skip connections, and residual-noise output. The decoder restores the temporal resolution through upsampling and skip connections. The final convolutional layer estimates the noise residual, which is subtracted from the noisy input to produce the denoised ECG signal.

The circular convolutional blocks allow the filters to access samples from the opposite segment boundary, thereby avoiding the artificial zero-valued context introduced by conventional padding. The skip connections preserve fine temporal information, whereas residual-noise prediction directs the network toward estimating the contaminating component rather than reconstructing the complete ECG waveform.

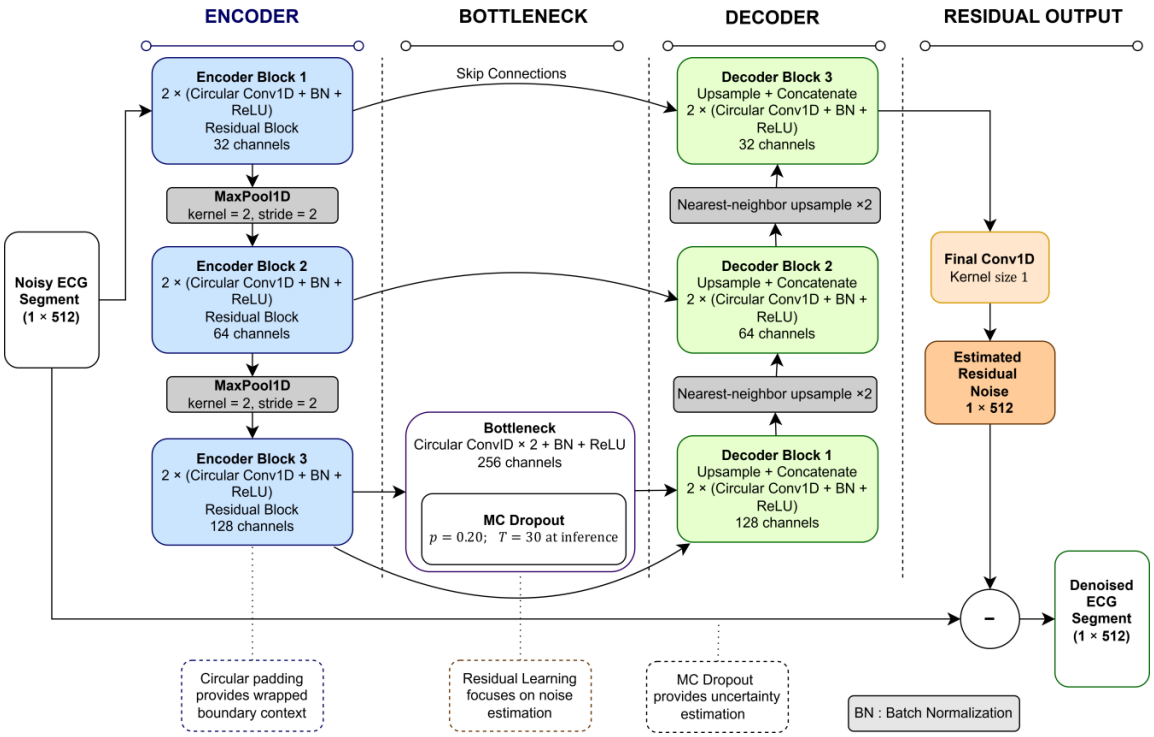


Figure 1. Proposed circular residual U-Net architecture with circular convolution, residual-noise prediction, skip connections, and MC dropout for uncertainty estimation

2.4. Morphology-aware boundary loss

To improve ECG morphology preservation, the model was trained with a morphology-aware boundary loss. The total loss combines mean-square reconstruction loss, absolute-error loss, derivative loss, and boundary loss:

$$L_{total} = \lambda_{MSE}L_{MSE} + \lambda_{MAE}L_{MAE} + \lambda_{der}L_{der} + \lambda_{bnd}L_{bnd} \quad (4)$$

In (4), the training objective defines as a weighted combination of global squared reconstruction error, global absolute reconstruction error, temporal-derivative error, and segment-boundary error. The final weights were λ_{MSE}=1.0, λ_{MAE}=0.5, λ_{der}=0.2, and λ_{bnd}=0.3.

The mean square error (MSE) and mean absolute error (MAE) terms measure the global reconstruction difference between the clean ECG x and the denoised signal $\hat{x}$:

$$L_{MSE} = \frac{1}{L}\sum_{i=1}^{L}(x_i - \hat{x}_i)^2 \quad (5)$$

$$L_{MAE} = \frac{1}{L}\sum_{i=1}^{L}|x_i - \hat{x}_i| \quad (6)$$

In (5) and (6), x_i and $\hat{x}_i$ denote the clean and denoised ECG samples, respectively, and L=512 is the segment length. During mini-batch optimization, the losses were averaged across all samples, channels, and segments in the batch.

$$L_{der} = \frac{1}{L-1}\sum_{i=1}^{L-1}|(x_{i+1} - x_i) - (\hat{x}_{i+1} - \hat{x}_i)| \quad (7)$$

In (7) compares the first-order temporal differences of the clean and denoised signals. The derivative term penalizes discrepancies in rapid waveform transitions and therefore discourages excessive smoothing around QRS complexes and R peaks.

$$L_{bnd} = \frac{1}{2k}\left[\sum_{i=1}^{k}|x_i - \hat{x}_i| + \sum_{i=L-k+1}^{L}|x_i - \hat{x}_i|\right] \quad (8)$$

where k=16 is the number of boundary samples selected from each edge and L=512 is the segment length. In (9), the mean absolute reconstruction error calculates over the concatenated first and last k samples of each segment. This targeted term discourages large reconstruction deviations near segment boundaries.

2.5. Training configuration and hyperparameter settings

All trainable models were implemented in PyTorch 2.5.1 and optimized using Kingma and Ba [25] with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-5}, batch size 64, and a maximum of 80 epochs. ReduceLROnPlateau reduced the learning rate by a factor of 0.5 after four epochs without validation improvement. Gradient clipping used a maximum norm of 1.0.

Early stopping used a patience of 12 epochs, and the checkpoint with the lowest validation loss was retained. PyTorch default initialization and float32 training were used without automatic mixed precision. Seed 42 controlled recording partitioning, segment and noise sampling, data-loader randomization, and parameter initialization; strict deterministic execution was not enforced.

All trainable models used the same morphology-aware loss. Convolutional neural network-denoising autoencoder (CNN-DAE) used dropout 0.10, the deterministic U-Net baselines used no dropout, and the proposed model used bottleneck dropout 0.20 with 30 stochastic inference passes. Experiments were conducted on an RTX 3050 Ti Laptop GPU with 8 GB VRAM under Windows 10, Python 3.10.20, PyTorch 2.5.1, CUDA 12.1, and cuDNN 9.1.

2.6. Bayesian uncertainty estimation using Monte Carlo dropout

Predictive uncertainty was estimated using MC dropout following the approximate Bayesian interpretation of dropout introduced by Gal and Ghahramani [18]. A dropout probability of 0.20 was applied at the bottleneck of the proposed model. During inference, the model was placed in evaluation mode so that the batch-normalization statistics remained fixed, while the bottleneck dropout layer was reactivated to generate stochastic predictions. Each noisy ECG segment was processed through T=30 independent stochastic forward passes. The number of MC passes was fixed before final test-set evaluation and was applied consistently to all test segments. This procedure generated a set of stochastic denoised predictions:

$$\{\hat{x}^{(t)}\}_{t=1}^{T} \tag{9}$$

where T=30 is the number of MC dropout forward passes and $\hat{x}^{(t)}$ is the denoised ECG generated during the t-th pass. Each forward pass uses a different dropout mask and therefore produces a slightly different reconstruction. The variation among these predictions is used as an approximation of epistemic uncertainty associated with the learned model parameters. The final denoised ECG was computed as the predictive mean:

$$\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{x}_i^{(t)} \tag{10}$$

The predictive mean in (10) was used as the final denoised ECG output for all performance calculations reported for the proposed model. Consequently, the reconstruction metrics in Tables 1 and 2 were calculated from the mean of the 30 stochastic predictions rather than from a deterministic single forward pass. Averaging the stochastic reconstructions reduces the influence of an individual dropout realization and provides a more stable waveform estimate. Predictive uncertainty was estimated from the standard deviation of the stochastic predictions:

$$\hat{\sigma_i} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{x}_i^{(t)} - \bar{x}_i\right)^2} \tag{11}$$

For each segment, T=30 stochastic reconstructions were generated by reactivating bottleneck dropout while retaining fixed batch-normalization statistics. Their sample-wise mean was used as the final reconstruction, and their sample-wise standard deviation represented approximate epistemic uncertainty. Segment-level uncertainty was the mean sample-wise standard deviation across 512 samples, whereas segment-level reconstruction error was the MAE between the clean ECG and the predictive mean.

$$u_j = \frac{1}{L}\sum_{i=1}^{L}\sigma_{j,i}, \rightarrow\ a_j = \frac{1}{L}\sum_{i=1}^{L}\left|x_{j,i} - \mu_{j,i}\right|, \tag{12}$$

In (12), u_j denotes the mean predictive standard deviation and a_j denotes the mean absolute reconstruction error for segment j. Pearson correlation was calculated across all N=2,560 segment-level pairs, rather than independently at each ECG sample.

The waveform uncertainty band was defined as $\mu_{j,i} \pm 2\sigma_{j,i}$, estimated from T=30 stochastic forward passes with a bottleneck dropout probability of 0.20. The band represents disagreement among MC-dropout reconstructions and was not interpreted as a formally calibrated clinical confidence interval.

Table 1. Overall reconstruction performance

Model	Output SNR (dB)	RMSE	MAE	PRD	Corr.
Circular residual U-Net with MC dropout	11.6789	0.1820	0.1382	32.5572	0.9309
Circular U-Net	11.1456	0.1891	0.1424	33.6794	0.9262
Zero-padded U-Net	11.0350	0.1893	0.1416	33.9153	0.9242
CNN-DAE	8.5340	0.2437	0.1767	43.8503	0.8846
db4 wavelet	8.7864	0.3679	0.3069	64.4812	0.8684

Table 2. Morphology and segment-boundary preservation performance

Model	Derivative MAE	Boundary MAE	Edge-center ratio	Peak error
Circular residual U-Net with MC dropout	0.0236	0.1678	2.0860	0.1998
Circular U-Net	0.0255	0.1756	2.2337	0.1998
Zero-padded U-Net	0.0253	0.1711	2.1680	0.2308
CNN-DAE	0.0411	0.1927	1.8711	0.3474
db4 wavelet	0.0426	0.3370	1.7275	0.3444

2.7. Baseline models

The proposed framework was compared with four baseline methods: i) db4 wavelet denoising; ii) CNN-DAE; iii) a standard three-level U-Net using zero padding; and iv) a three-level circular U-Net without residual-noise output or uncertainty-aware inference. The zero-padded and circular U-Net baselines used identical encoder-decoder channel configurations. The comparisons were designed to assess overall performance differences among the evaluated frameworks. Because a complete factorial ablation study was not conducted, improvements are interpreted as the combined effect of the proposed framework rather than being attributed exclusively to one component.

The zero-padded and circular U-Net baselines used the same three-level channel configuration, convolutional kernel size, max-pooling operations, nearest-neighbor upsampling, skip connections, and direct clean-signal output. They differed principally in their padding strategy and did not use encoder residual connections or dropout. The CNN-DAE used encoder channels of 32, 64, and 128, with max pooling after the first two convolutional stages. Its decoder used nearest-neighbor upsampling followed by 64- and 32-channel convolutional stages and a final 1×1 output convolution. A dropout probability of 0.10 was applied in the CNN-DAE encoder.

To ensure a fair comparison, all trainable deep learning models were reproduced within the same experimental pipeline and trained from scratch using identical recording-level training, validation, and test partitions. The same 512-sample ECG segments, normalization procedure, NSTDB noise-generation protocol, input SNR levels, and clean target signals were used for all models. The trainable models were also optimized using the same optimizer, batch size, maximum number of epochs, early-stopping criterion, and validation-based checkpoint-selection procedure. The zero-padded U-Net and circular U-Net used the same encoder-decoder depth and feature-channel configuration, with the padding strategy being the principal architectural difference. The CNN-DAE retained its original autoencoder structure but was trained and evaluated using the same data partitions and noise realizations. The db4 wavelet method required no model training and was applied to exactly the same noisy test segments used to evaluate the deep learning models. Consequently, performance differences were evaluated using identical clean references, noise conditions, and test samples.

2.8. Evaluation metrics

Reconstruction performance was assessed using output SNR, root mean square error (RMSE), MAE, percentage root-mean-square difference (PRD), and waveform correlation. Morphology and boundary behavior were assessed using derivative MAE, peak-amplitude error, boundary MAE, and edge-to-center ratio. Each metric was calculated per segment and then averaged within each evaluation group.

$$SNR_{out,j} = 10\, log_{10}\left(\frac{\sum_{i=1}^{L} x_{j,1}^2}{\sum_{i=1}^{L} e_{j,i}^2+\varepsilon}\right), \varepsilon = 10^{-12} \quad (13)$$

In (13), the output SNR for segment j is defined as the ratio between clean-signal energy and reconstruction-error energy. A higher output SNR therefore indicates better noise suppression while preserving the clean ECG waveform.

$$RMSE_j = \sqrt{\frac{1}{L}\sum_{i=1}^{L} e_{j,i}^2} \tag{14}$$

In (14), the segment-level RMSE is defined. Because the reconstruction errors are squared before averaging, RMSE gives greater emphasis to larger sample-wise deviations between the reconstructed and clean ECG signals.

$$PRD_j = 100\sqrt{\frac{\sum_{i=1}^{L} e_{j,i}^2}{\sum_{i=1}^{L} x_{j,i}^2+\varepsilon}} \tag{15}$$

In (15), PRD as the reconstruction-error energy normalized is defined by the clean-signal energy and expressed as a percentage. A lower PRD indicates closer agreement between the reconstructed and clean ECG waveforms.

$$MAE_j = \frac{1}{L}\sum_{i=1}^{L}\left|e_{j,i}\right|,$$

$$\rho_j = \frac{\sum_{i=1}^{L}(x_{j,i}-\bar{x}_j)(\hat{x}_{j,i}-\bar{\hat{x}}_j)}{\sqrt{\sum_{i=1}^{L}(x_{j,i}-\bar{x}_j)^2\sum_{i=1}^{L}(\hat{x}_{j,i}-\bar{\hat{x}}_j)^2}},$$

$$D-MAE_j = \frac{1}{L-1}\sum_{i=1}^{L-1}\left|\left(x_{j,i+1}-x_{j,i}\right)-\left(\hat{x}_{j,i+1}-\hat{x}_{j,i}\right)\right|,$$

$$B-MAE_j = \frac{1}{2k}\left[\sum_{i=1}^{k}\left|e_{j,i}\right|+\sum_{i=L-k+1}^{L}\left|e_{j,i}\right|\right],$$

$$ECR_j = \frac{B-MAE_j}{C-MAE_j+\varepsilon},$$

$$E_{peak,j} = \left|\max_i\left(x_{j,i}\right)-\max_i\left(\hat{x}_{j,i}\right)\right|. \tag{16}$$

where,

$$C-MAE_j = \left|\frac{1}{2k}\sum_{i=\frac{L}{2}-k+1}^{\frac{L}{2}+k}\left|e_{j,i}\right|\right|,\rightarrow k=16$$

In (16), the complementary evaluation measures is defined. MAE quantifies the mean absolute sample-wise reconstruction error, while ρ measures waveform correlation between the clean and reconstructed signals. Derivative MAE evaluates discrepancies in first-order temporal changes, boundary MAE measures reconstruction error over the first and last k=16 samples, and the edge-to-center ratio compares boundary error with error in the central portion of the segment. Peak-amplitude error measures the absolute difference between the maximum amplitudes of the clean and reconstructed ECG signals and does not represent annotated R-peak detection or timing accuracy.

Lower RMSE, MAE, PRD, derivative MAE, boundary MAE, and peak-amplitude error indicate better reconstruction, whereas higher output SNR and waveform correlation indicate closer agreement with the clean ECG. The edge-to-center ratio should be interpreted together with absolute boundary MAE because it reflects the relative distribution of reconstruction error.

3. RESULTS AND DISCUSSION

3.1. Overall denoising performance

Table 1 summarizes the reconstruction performance across all 2,560 test segments. The proposed circular residual U-Net with MC dropout achieved the highest mean output SNR of 11.6789 dB and

waveform correlation coefficient of 0.9309. It also obtained the lowest RMSE of 0.1820, MAE of 0.1382, and PRD of 32.5572 among the evaluated methods. These results indicate that the predictive mean derived from 30 stochastic forward passes provided the closest overall agreement with the clean ECG references.

As shown in Figure 2, the proposed model achieved the highest output SNR among the learning-based methods at input SNR levels of −6, 6, 12, 18, and 24 dB. At 0 dB, the zero-padded U-Net achieved a marginally higher output SNR. The proposed method nevertheless provided the best aggregate reconstruction performance across the complete test set, as indicated by its lower RMSE, MAE, and PRD and its higher waveform correlation. Figure 2 compares output SNR across six input SNR levels under mixed NSTDB noise.

The proposed model achieved the highest output SNR among the learning-based methods at −6, 6, 12, 18, and 24 dB, whereas the zero-padded U-Net was marginally higher at 0 dB. The db4 wavelet method exceeded the learning-based methods at 18 and 24 dB, but its aggregate waveform errors remained substantially higher.

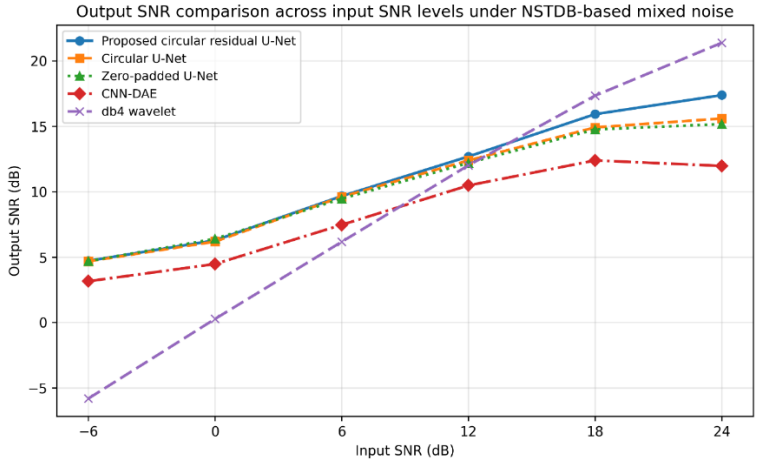


Figure 2. Output SNR comparison across input SNR levels under NSTDB-based mixed noise

3.2. Comparison among U-Net variants

The circular U-Net achieved a slightly higher output SNR than the zero-padded U-Net, with values of 11.1456 and 11.0350 dB, respectively. However, their boundary MAEs were 0.1756 and 0.1711, indicating that circular padding alone did not consistently reduce aggregate boundary error. The proposed model achieved a higher output SNR of 11.6789 dB and a lower boundary MAE of 0.1678, together with lower RMSE, MAE, PRD, and derivative MAE than both deterministic U-Net baselines.

CNN-DAE achieved an output SNR of 8.5340 dB, RMSE of 0.2437, MAE of 0.1767, and peak-amplitude error of 0.3474. For the proposed model, the corresponding values were 11.6789 dB, 0.1820, 0.1382, and 0.1998. The db4 wavelet method produced the highest RMSE, MAE, and PRD values of 0.3679, 0.3069, and 64.4812, respectively.

These findings support the combined circular-padding, residual-noise-prediction, morphology-aware-loss, and MC-predictive-mean framework. Because a complete factorial ablation study was not conducted, the improvement cannot be attributed exclusively to an individual component.

3.3. Boundary-artifact analysis

Table 2 summarizes morphology and segment-boundary preservation. The proposed model achieved the lowest derivative MAE of 0.0236 and boundary MAE of 0.1678. Relative to CNN-DAE and db4 wavelet denoising, the proposed model reduced derivative MAE by 42.59% and 44.64%, and boundary MAE by 12.92% and 50.20%, respectively. Its peak-amplitude error was comparable to that of the circular U-Net, with both values rounded to 0.1998.

The edge-to-center ratio should be interpreted together with absolute boundary MAE. CNN-DAE and db4 wavelet denoising produced lower ratios despite substantially larger boundary errors, indicating that the ratio reflects the relative distribution of error rather than absolute edge-reconstruction quality. Figure 3 compares boundary MAE across input SNR levels under mixed NSTDB noise.

Boundary error generally decreased as input SNR increased. At −6 and 0 dB, the zero-padded U-Net obtained slightly lower boundary MAE than the proposed model. From 6 to 24 dB, the proposed model achieved the lowest boundary MAE among the learning-based methods. When aggregated across all noise types and SNR levels, the proposed model obtained the lowest overall boundary MAE.

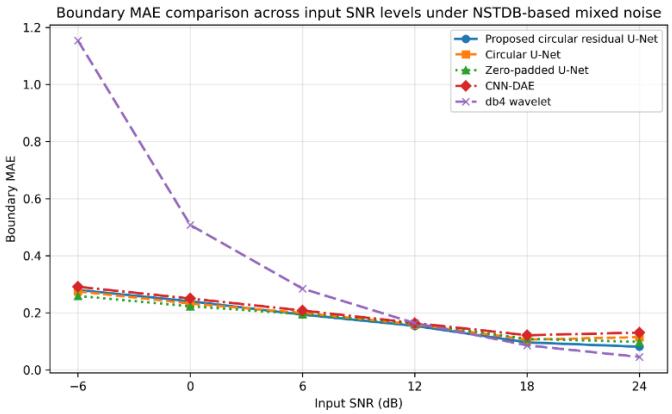


Figure 3. Boundary MAE comparison across input SNR levels under NSTDB-based mixed noise

3.4. Qualitative denoising analysis

Figure 4 presents a representative mixed-noise segment at an input SNR of 6 dB. The U-Net-based methods followed the clean waveform more closely than CNN-DAE and db4 wavelet denoising. The proposed reconstruction preserved the dominant waveform transitions while suppressing the nonstationary disturbance, consistent with its lower derivative MAE and peak-amplitude error. This qualitative comparison does not establish improvement in downstream heartbeat detection, arrhythmia classification, or clinical diagnosis.

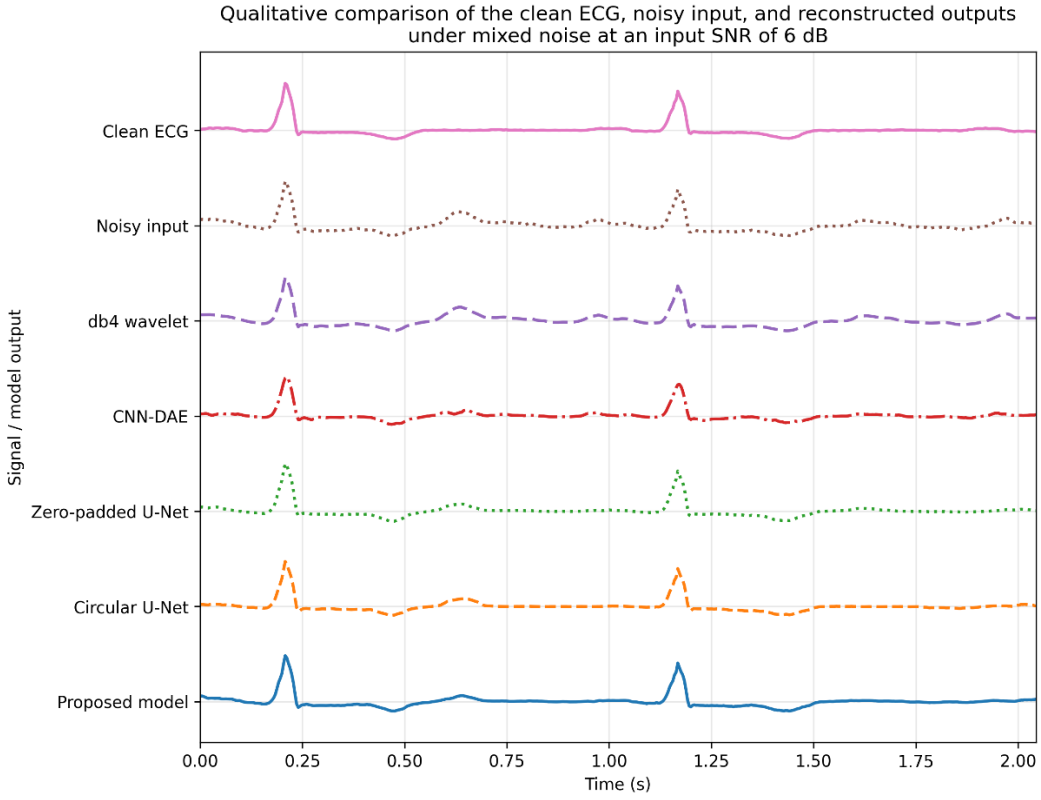


Figure 4. Qualitative comparison of the clean ECG, noisy input, and reconstructed outputs under mixed noise at an input SNR of 6 dB

3.5. Uncertainty-estimation analysis

Table 3 reports the relationship between segment-level predictive uncertainty and reconstruction error across the complete test set and individual noise conditions. Across all 2,560 segments, predictive uncertainty correlated strongly with reconstruction error (r=0.8239, 95% CI 0.8110–0.8360, p<0.001). Positive correlations were also observed for BW (r=0.7793), MA (r=0.8159), EM artifact (r=0.8807), and mixed noise (r=0.8420), all p<0.001.

Table 3. Segment-level uncertainty-error correlation

Evaluation group	Pearson r	p-value	Segments
Overall	0.8239	<0.001	2,560
BW	0.7793	<0.001	640
MA	0.8159	<0.001	640
EM	0.8807	<0.001	640
Mixed noise	0.8420	<0.001	640

Figure 5 illustrates a representative reconstruction with sample-wise uncertainty estimated from 30 stochastic MC dropout passes. The shaded region represents the predictive mean±2 standard deviations. Wider bands indicate greater disagreement among stochastic predictions and therefore higher estimated epistemic uncertainty; they are not clinically calibrated confidence intervals.

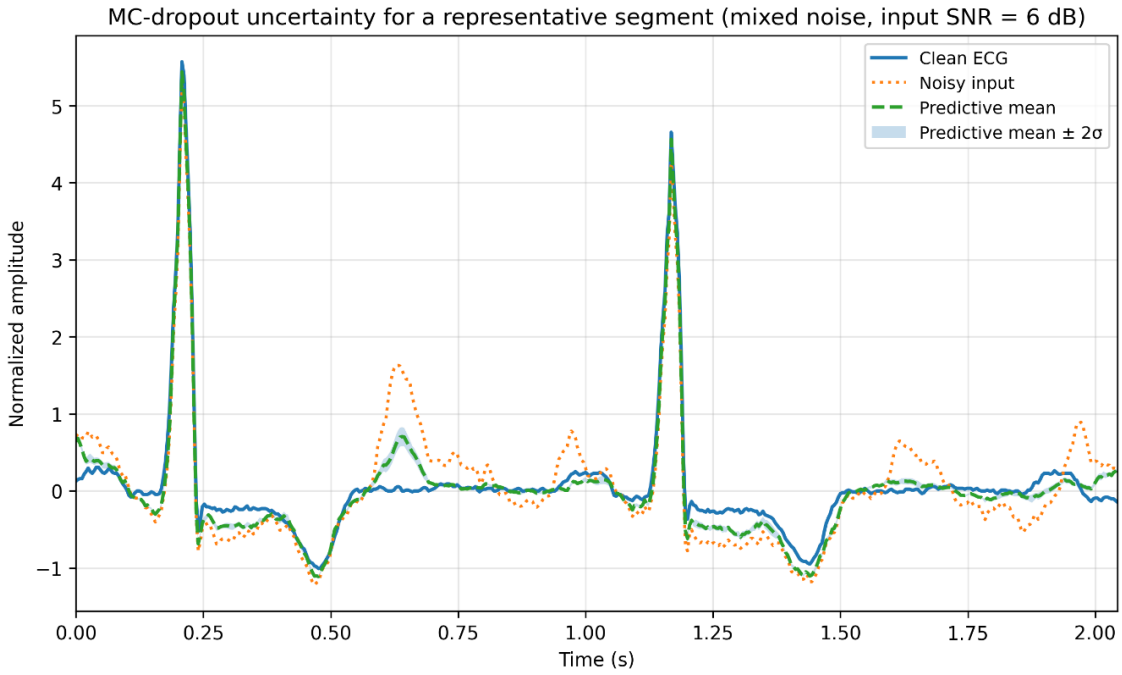


Figure 5. Representative uncertainty-aware ECG reconstruction under mixed noise at an input SNR of 6 dB

Figure 6 visualizes the segment-level relationship between mean predictive standard deviation and MAE. Segments with greater predictive uncertainty generally exhibited higher reconstruction errors. However, this association does not establish formal probabilistic or clinical calibration.

3.6. Computational cost and inference efficiency

Table 4 summarizes trainable parameter counts and inference latency. The proposed model contained 663,041 trainable parameters, approximately 1.61% more than the zero-padded and circular U-Net baselines. The proposed model required 6.28±0.71 ms for deterministic single-pass inference and 203.54±7.17 ms for 30-pass MC inference. The latter was approximately 32.4 times slower but represented only 9.94% of the 2.048-s segment duration on the reported GPU. These measurements do not establish real-time suitability on embedded or wearable hardware, where memory, power consumption, and throughput remain to be evaluated.

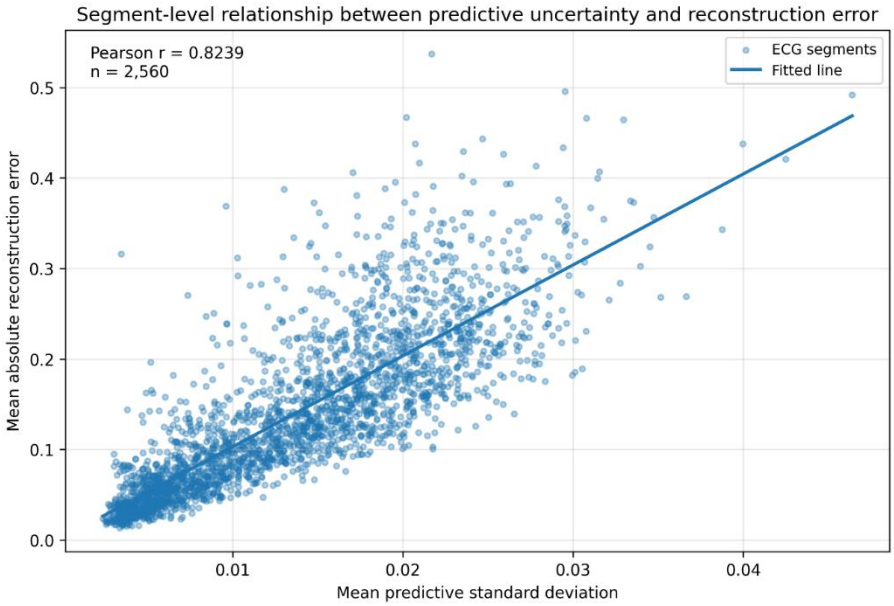


Figure 6. Relationship between predictive uncertainty and reconstruction error across 2,560 test segments

This study has several limitations. The evaluation used normalized single-channel QTDB segments contaminated with NSTDB-derived noise, without external validation on independent patients, devices, or acquisition environments. Only one recording-level split and principal training seed were evaluated, and no complete factorial ablation was conducted. Predictive uncertainty was assessed through error correlation rather than formal calibration or coverage. Downstream arrhythmia classification, R-peak detection, physician interpretation, prospective clinical workflows, and embedded-device deployment were not evaluated.

Table 4. Parameter count and inference latency

Model	Trainable parameters	Single-pass latency (ms)	MC inference latency (ms)
CNN-DAE	62,529	1.96±0.62	Not applicable
Zero-padded U-Net	652,545	4.60±0.37	Not applicable
Circular U-Net	652,545	6.04±0.95	Not applicable
Proposed, single pass	663,041	6.28±0.71	Not applicable
Proposed, $T = 30$	663,041	Not applicable	203.54±7.17

4. CONCLUSION

This study presented an uncertainty-aware circular residual U-Net integrating circular convolution, residual-noise prediction, morphology-aware boundary loss, and MC dropout for ECG denoising under realistic nonstationary noise. Across 2,560 test segments, the model achieved an output SNR of 11.6789 dB, RMSE of 0.1820, MAE of 0.1382, PRD of 32.5572, and waveform correlation of 0.9309. It also obtained the lowest derivative MAE of 0.0236 and boundary MAE of 0.1678, while its peak-amplitude error of 0.1998 was comparable to the circular U-Net.

Predictive uncertainty correlated strongly with reconstruction error (r=0.8239, p<0.001), supporting its use as an indicator of reconstruction difficulty rather than a clinically calibrated confidence measure. Thirty-pass MC inference required 203.54 ms for each 2.048-s segment on the reported GPU. Future work should include external validation, repeated-seed evaluation, formal uncertainty calibration, component-wise ablation, downstream ECG analysis, and embedded-device testing.

ACKNOWLEDGMENTS

During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.6 Thinking) for English-language refinement, editorial organization, consistency checking, and limited coding assistance. All AI-assisted outputs and code were critically reviewed, tested, and validated by the authors. The tool was not used to create or manipulate original research data or final scientific figures. The authors independently

verified the experimental procedures, numerical results, statistical analyses, references, interpretations, and conclusions and take full responsibility for the accuracy, originality, and integrity of the manuscript.

FUNDING INFORMATION

This research was supported by a research grant from the Artificial Intelligence-Medical Center of Excellence (AIMed-COE), Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia. No separate research grant or contract number was assigned.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ahmad Rifai	✓	✓	✓	✓	✓	✓		✓	✓		✓			
Siti Nurmaini	✓	✓		✓						✓		✓	✓	✓
Bhakti Yudho Suprapto		✓		✓						✓		✓		
Radiyati Umi Partan				✓		✓				✓				
Firdaus		✓		✓						✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY

The source ECG and noise recordings used in this study are publicly available through PhysioNet in the QT Database and the Massachusetts Institute of Technology - Beth Israel Hospital (MIT-BIH) Noise Stress Test Database. Additional materials supporting the findings of this study may be made available by the corresponding author upon reasonable request.

REFERENCES

[1] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

[2] S. Chatterjee, R. S. Thakur, R. N. Yadav, L. Gupta, and D. K. Raghuvanshi, "Review of noise removal techniques in ECG signals," *IET Signal Processing*, vol. 14, no. 9, pp. 569–590, Dec. 2020, doi: 10.1049/iet-spr.2020.0104.

[3] D. L. Donoho, "De-Noising by Soft-Thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995, doi: 10.1109/18.382009.

[4] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Burlington, MA, USA: Academic Press, 2008, doi: 10.1016/B978-0-12-374370-1.X0001-8.

[5] M. Rakshit and S. Das, "An efficient ECG denoising methodology using empirical mode decomposition and adaptive switching mean filter," *Biomedical Signal Processing and Control*, vol. 40, pp. 140–148, Feb. 2018, doi: 10.1016/j.bspc.2017.09.020.

[6] H. Huang, S. Hu, and Y. Sun, "A discrete curvature estimation based low-distortion adaptive savitzky–golay filter for ECG denoising," *Sensors (Switzerland)*, vol. 19, no. 7, p. 1617, Apr. 2019, doi: 10.3390/s19071617.

[7] L. Hu, W. Cai, Z. Chen, and M. Wang, "A lightweight U-Net model for denoising and noise localization of ECG signals," *Biomedical Signal Processing and Control*, vol. 88, p. 105504, Feb. 2024, doi: 10.1016/j.bspc.2023.105504.

[8] H. Peng, X. Chang, Z. Yao, D. Shi, and Y. Chen, "A deep learning framework for ECG denoising and classification," *Biomedical Signal Processing and Control*, vol. 94, p. 106441, Aug. 2024, doi: 10.1016/j.bspc.2024.106441.

[9] C. T. C. Arsene, R. Hankins, and H. Yin, "Deep learning models for denoising ECG signals," in *European Signal Processing Conference*, 2019, doi: 10.23919/EUSIPCO.2019.8902833.

[10] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5580-5590, 2017.

[11] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific Reports*, vol. 7, no. 1, p. 17816, Dec. 2017, doi: 10.1038/s41598-017-17876-z.

[12] J. F. Vranken *et al.*, "Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms," *European Heart Journal - Digital Health*, vol. 2, no. 3, pp. 401–415, Sep. 2021, doi: 10.1093/ehjdh/ztab045.
[13] A. O. Aseeri, "Uncertainty-aware deep learning-based cardiac arrhythmias classification model of electrocardiogram signals," *Computers*, vol. 10, no. 6, p. 82, Jun. 2021, doi: 10.3390/COMPUTERS10060082.
[14] M. Barandas *et al.*, "Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram," *Information Fusion*, vol. 101, p. 101978, Jan. 2024, doi: 10.1016/j.inffus.2023.101978.
[15] M. Dumitru *et al.*, "A data-driven Gaussian process filter for electrocardiogram denoising," *Journal of Electrocardiology*, vol. 73, p. 28, Jul. 2022, doi: 10.1016/j.jelectrocard.2022.07.062.
[16] R. Nakanishi, A. Hirata, and Y. Kubota, "Deep Learning-Based Multi-Lead ECG Reconstruction from Lead I with Metadata Integration and Uncertainty Estimation," *Sensors*, vol. 26, no. 1, p. 212, Dec. 2026, doi: 10.3390/s26010212.
[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[18] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *33rd International Conference on Machine Learning, ICML 2016*, 2016, vol. 3, pp. 1050–1059.
[19] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A Database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Computers in Cardiology*, IEEE, 1997, pp. 673–676, doi: 10.1109/cic.1997.648140.
[20] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.
[21] G. B. Moody, W. K. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Computers in Cardiology*, vol. 11, no. 1, pp. 381–384, 1984.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*, vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015, vol. 1, pp. 448–456.
[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

BIOGRAPHIES OF AUTHORS



Ahmad Rifai is a doctoral student at the Faculty of Engineering, Universitas Sriwijaya, Indonesia. He is also a lecturer and researcher at Universitas Sriwijaya. His research interests include biomedical signal processing, electrocardiogram denoising, deep learning, and uncertainty quantification. He can be contacted at email: ahmadrifai@unsri.ac.id.



Siti Nurmaini is a professor and researcher with the Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. She received her master's degree in Control Systems from Institut Teknologi Bandung, Indonesia, in 1998 and her Ph.D. degree in Computer Science from Universiti Teknologi Malaysia in 2011. Her research interests include biomedical engineering, deep learning, machine learning, image processing, control systems, and robotics. She can be contacted at email: siti_nurmaini@unsri.ac.id and sitinurmaini@gmail.com.



Bhakti Yudho Suprapto is a lecturer and researcher at the Faculty of Engineering, Universitas Sriwijaya, Indonesia. His research interests include control systems, signal processing, artificial intelligence, and electrical engineering. He can be contacted at email: bhakti@ft.unsri.ac.id.



Radiyati Umi Partan is a rheumatologist and staff member in the Department of Internal Medicine, Faculty of Medicine, Universitas Sriwijaya, Palembang, South Sumatra, Indonesia. Her research interests include rheumatology, internal medicine, inflammation, and clinical medicine. She can be contacted at email: radiyati.u.p@fk.unsri.ac.id.



Firdaus is a lecturer and researcher with the Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. His research interests include text processing, deep learning, and machine learning. He can be contacted at email: firdaus@unsri.ac.id.