An Experimental Study of Atrial Fibrillation Detection Using Deep Neural Networks Under Noisy Conditions

Dalia Vektarytė, Zigmantas Kęstutis Juškevičius, Jolita Bernatavičienė

Vilnius University, Faculty of Mathematics and Informatics,
Institute of Data Science and Digital Technologies,
Akademijos g. 4, LT-08412 Vilnius, Lithuania
dalia.vektaryte@mif.stud.vu.lt

Abstract. Atrial fibrillation, characterized by rapid and irregular electrical activity in the atria, is the most common sustained cardiac arrhythmia and a major risk factor for stroke. Early detection of AFib is critical for timely diagnosis and effective treatment. However, AFib detection from electrocardiogram recordings remains challenging due to noise contamination, particularly in wearable device recordings, where motion-related artifacts are frequently present. Numerous deep learning approaches have been proposed for AFib detection, but direct comparison between existing methods is difficult due to different experimental settings. Also, the effect of real-world noise on performance is underexplored. In response to this challenge, we performed a cross-dataset test on deep learning models CTRhythm, MFEGNet, and MGCNet under clean and noisy conditions. CTRhythm achieved the strongest performance on clean signals, but it was the most vulnerable to noise. MFEGNet demonstrated the greatest noise resilience, consistent with its architecture designed for noise suppression. Despite this, all models' performance degraded substantially when tested on noisy signals. These results highlight the importance of standardised cross-dataset evaluation under real-life conditions for assessing the true utility of AFib detection models and the development of robust pipelines that incorporate signal denoising.

Keywords: atrial fibrillation, deep learning, noise robustness, ECG signals, wearable monitoring.

1 Introduction

Atrial fibrillation (AFib) is one of the most common cardiac arrhythmias and can be asymptomatic and occur in transient episodes, which makes it more difficult to detect. AFib that is not managed in time can significantly increase the risk of stroke and heart failure. In Lithuania, AFib was identified as a

DOI: https://doi.org/10.15388/LMITT.2026.33

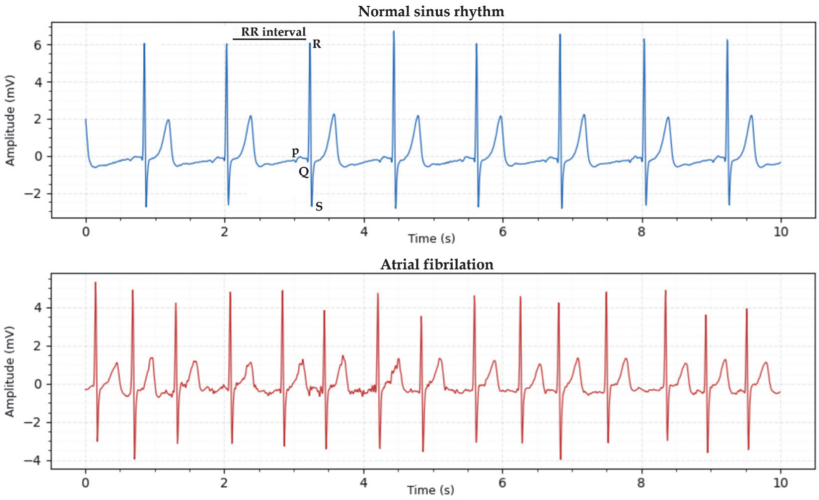

Fig. 1. Electrocardiogram windows illustrating normal sinus rhythm (top), and atrial fibrillation (bottom).

Early automated AFib detection methods relied on manually extracted features, for example, RR interval variability, however, such approaches risk overlooking other diagnostically relevant signal characteristics. Therefore, current research has shifted toward deep learning (DL) models and automatic feature extraction directly from raw ECG signals. In current research, most AFib detection models are based on convolutional neural networks (CNNs) which are effective at extracting local morphological patterns of the ECG. Since AFib exhibits characteristic feature changes in both the time domain and the frequency domain, CNN-based methods have been developed for both representations [2]. One limitation of CNNs is that they have a limited receptive field and are better at extracting local features within a heartbeat. However, AFib is a rhythm disorder that manifests across multiple beats and requires analysis of longer ECG signals for accurate detection. This limitation has been addressed by utilizing sequential architectures, as their hidden state allows rhythm patterns to be preserved across time steps. Long short-term memory networks and gated recurrent units (GRUs) are the most commonly used variants, due to their ability to mitigate the vanishing-gradient problem. Hybrid architectures with CNN can extract both local and rhythm patterns and have been shown to outperform pure CNN baselines. More recently, Transformer-based architectures have been investigated, as

multi-head self-attention allows the model to learn global patterns across the entire input sequence in a single step [3].

Beyond architecture choice, input window length is also an important design decision. ECG signal length used for AFib detection with DL varies across studies significantly depending on the objectives of the analysis and the practical constraints of the algorithm. Reported segment lengths range from single beat intervals to 60-second windows, with 30-second and 10s windows being the most common choice. This choice also has clinical considerations. Earlier European Society of Cardiology guidelines defined AFib as an arrhythmia episode lasting at least 30 seconds [9], however, this threshold could limit detection of shorter paroxysmal episodes that fall below it. Also, a routine clinical ECG recording is typically 10 seconds long, which makes it a clinically appropriate choice for evaluation.

However, regardless of window length, detection accuracy is also highly sensitive to signal quality, as noise can obscure the clinically relevant features of the ECG. ECG signals are frequently corrupted with baseline wander (BW), caused by patient breathing, which shifts the isoelectric line, but it can be addressed through lowcut filtering. One of the most challenging noise sources in ECG recordings is muscle artifact (MA), caused by muscle contractions. MA causes significant challenges in AFib detection because it overlaps with the full frequency spectrum of the ECG signal and therefore cannot be removed through simple frequency-domain filtering. In a clinical setting, MA can be minimized by instructing the patient to remain still. However, in wearable devices this type of noise is practically impossible to avoid. Electrode motion artifact (EM) is caused by poor skin-to-electrode contact and presents a similar challenge in ECG analysis tasks as MA.

DL models can learn to detect subtle patterns in ECG data with high accuracy, but exposure to noise at inference time can cause feature representations to shift away from what the model learned on clean signals and cause classification results to degrade. Despite the wearable recording susceptibility to noise, DL models are rarely evaluated on real-world noisy signals, and the extent to which realistic noise will degrade classification performance is underreported. This paper addresses this gap by evaluating three architectures under a unified cross-dataset protocol that includes noise conditions consistent with wearable device recordings.

2 Data

Two publicly available and physician-annotated Holter monitor ECG databases were used for AFib detection. SHDB-AF [1] contains 98 annotated 24-hour two-lead recordings sampled at 200 Hz, and the MIT-BIH Atrial Fibrillation Database (AFDB) [5] contains 23 10-hour two-lead recordings sampled at 250 Hz. In both datasets, only a single ECG channel was used. To ensure a unified sampling rate, AFDB records were downsampled from 250 Hz to 200 Hz.

Records were then segmented into non-overlapping fixed-length windows and labeled using the majority rule. Each window was normalised using z-score normalization. To prevent the model from learning patient-specific patterns, an inter-patient data split was used during training. SHDB-AF patient records were divided into 5-fold inter-patient training and validation splits, stratified by the number of AFib windows. AFDB was used exclusively as a cross-dataset test set. MIT-BIH Noise Stress Test Database (NSTDB) [6] records for MA, EM, and BW were used to simulate real-world noise. Clean AFDB windows were contaminated using a segment-adaptive protocol. Each clean ECG window x_i was corrupted by an NSTDB noise segment n_i of equal length according to (1):

$$\tilde{x}_i = x_i + \alpha_i n_i, \tag{1}$$

where the amplitude scaling factor α_i was defined as (2):

$$\alpha_i = \frac{\rho_i}{R_n/(R_x+\epsilon)+\epsilon}, \rho_i \sim U(0.2,2.0) \tag{2}$$

where R_x and R_n denote the dynamic ranges of the ECG and noise segments respectively, ρi ~ U(0.2, 2.0) controls contamination intensity, while $\epsilon = 10^{-12}$ prevents numerical issues. This created different levels of noise across windows, which is closer to real-world noise than using a fixed noise level.

3 Deep learning models

Three DL classification models were selected to represent different approaches to AFib detection. CTRhythm (Convolutional Neural Network–Transformer Rhythm Classifier) [3] is a hybrid CNN–Transformer model that combines local CNN feature extraction with the Transformer's ability to learn global patterns. The CTRhythm architecture consists of a six-block residual CNN that first extracts local morphological features from the raw ECG signal.

The first three blocks apply stride 2 for progressive downsampling, which reduces sequence length, increases feature dimensionality, and improves extraction of informative local patterns. Six Transformer encoder layers with eight-head self-attention are used to learn patterns across the full signal window in parallel. This gives it an advantage over the CNN models with limited receptive field and recurrent networks that have sequential processing constraints.

MFEGNet (multiscale feature-enhanced gating network) [8] was proposed to address challenges in AFib detection under noisy conditions. The architecture includes an initial convolutional layer, a multiscale convolution module, two Feature Enhancement blocks, three residual blocks that perform downsampling for deep feature extraction, and a GRU layer that models rhythmic patterns. The multiscale convolution module applies four parallel convolutional branches that allow the model to extract ECG morphological features at multiple temporal scales simultaneously. To suppress noisy and redundant features, each feature enhancement block combines a soft-threshold residual shrinkage unit, a dilated convolution to widen the receptive field, and a Squeeze-and-Excitation module that assigns a learned weight to each feature channel. Therefore, channels that have AFib-relevant information are amplified and others are suppressed.

MGCNet (multimodal gated contrastive network) [2] is a dual-branch multimodal network proposed to improve AFib detection robustness and cross-dataset generalization. The architecture includes two parallel branches. The time-domain branch processes the raw ECG segment through a 1D convolutional encoder, and the frequency-domain branch converts the same segment into a Short-Time Fourier Transform spectrogram and processes it through a 2D convolutional encoder. The reasoning behind this is that AFib manifests as irregular rhythm intervals in the time domain and chaotic atrial activity in the frequency domain; therefore, processing them simultaneously provides a more complete representation than either domain alone.

At three hierarchical levels, the branches are connected through a bidirectional gating module, which computes per-channel weight vectors and applies them to suppress or amplify channels in the opposite branch. The time-domain branch uses a bidirectional GRU to model temporal patterns bidirectionally, while the frequency branch uses global average pooling to summarise spectral content. In addition to binary cross-entropy, the model is trained with a cross-modal supervised contrastive loss, which is designed to improve cross-dataset generalisation by pulling together

same-class temporal and spectral representations and separating cross-class pairs.

Although the three models are different in architecture, all are designed for AFib detection tasks. The aim of this study is to determine which architectural approach generalises most reliably across datasets and under realistic noise conditions and how noise affects their performance.

4 Results

All three models were trained on the SHDB-AF training split for each of the 10s and 30s window lengths using 5-fold patient-stratified cross-validation and evaluated on AFDB. The evaluation metrics were selected to enable meaningful comparison of model classification performance. The primary metric is macro F1 score, calculated as the unweighted average of per-class F1 scores. It was chosen because it is appropriate for imbalanced datasets as it treats both classes equally regardless of their size. Also, the AFib class precision and recall are reported to indicate if the model tends to over-predict AFib or to miss true positive cases. All metric values are reported as mean and standard deviation across the five cross-validation folds. Results are presented in Tables 1 – 2, with the best values per condition highlighted in bold.

On clean signals, CTRhythm and MFEGNet both performed better on 10s segments. CTRhythm had the highest performance overall in accuracy (0.877 ± 0.044), macro F1 score (0.865 ± 0.050) and AFib precision (0.956 ± 0.018), and MFEGNet had the highest AFib recall (0.810 ± 0.146) when using the same window length (Table 1).

Table 1. Cross-dataset classification results for 10s and 30s window lengths on clean signals. The highest mean values within each window length are highlighted in bold.

Model	Window	Accuracy	Macro F1	AFib Recall	AFib precision
CTRhythm	10 s	**0.877 ± 0.044**	**0.865 ± 0.050**	0.728 ± 0.105	**0.956 ± 0.018**
	30s	0.841 ± 0.034	0.823 ± 0.046	0.697 ± 0.109	**0.885 ± 0.067**
MFEGNet	10 s	0.869 ± 0.050	0.860 ± 0.056	**0.810 ± 0.146**	0.869 ± 0.089
	30s	0.800 ± 0.071	0.774 ± 0.111	0.708 ± 0.277	0.801 ± 0.091
MGCNet	10 s	0.814 ± 0.076	0.787 ± 0.107	0.616 ± 0.181	0.867 ± 0.082
	30s	**0.852 ± 0.053**	**0.841 ± 0.058**	**0.712 ± 0.167**	0.856 ± 0.064

For noise robustness, the models were re-evaluated on test sets contaminated with three different noise types (BW, MA, EM). This caused performance degradation across all models and all noise types (Table 2). CTRhythm showed the sharpest decline in performance, macro F1 score dropped from 0.865 to 0.531 on ECG signals with MA, which was caused by a recall collapse while precision remained high (0.914 ± 0.050). However, MGCNet overpredicted AFib on ECG with MA and EM noise as it had higher AFib recall than the other models, but it also showed substantially reduced precision (0.649 ± 0.1 and 0.663 ± 0.138 respectively).

Table 2. Results of classification models on 10s windows with different noise type contaminations. The highest mean values within each condition are highlighted in bold.

Model	Noise type	Accuracy	Macro F1	AFib Recall	AFib precision
CTRhythm	clean	**0.877 ± 0.044**	**0.865 ± 0.050**	0.728 ± 0.105	**0.956 ± 0.018**
	EM	0.687 ± 0.024	0.585 ± 0.052	0.251 ± 0.085	**0.888 ± 0.051**
	MA	0.663 ± 0.035	0.531 ± 0.065	0.169 ± 0.083	**0.914 ± 0.050**
	BW	0.692 ± 0.032	0.594 ± 0.070	0.270 ± 0.122	**0.911 ± 0.086**
MFEGNet	clean	0.869 ± 0.050	0.860 ± 0.056	**0.810 ± 0.146**	0.869 ± 0.089
	EM	**0.711 ± 0.037**	**0.643 ± 0.084**	0.399 ± 0.219	0.815 ± 0.091
	MA	**0.735 ± 0.045**	**0.673 ± 0.091**	0.434 ± 0.225	0.868 ± 0.098
	BW	**0.723 ± 0.041**	**0.658 ± 0.089**	**0.430 ± 0.248**	0.851 ± 0.116
MGCNet	clean	0.814 ± 0.076	0.787 ± 0.107	0.616 ± 0.181	0.867 ± 0.082
	EM	0.671 ± 0.034	0.634 ± 0.020	**0.468 ± 0.118**	0.649 ± 0.100
	MA	0.669 ± 0.031	0.627 ± 0.026	**0.437 ± 0.115**	0.663 ± 0.138
	BW	0.670 ± 0.021	0.603 ± 0.050	0.399 ± 0.235	0.755 ± 0.171

5 Conclusions

In this paper, we evaluated three DL architectures on cross-dataset testing that included noise consistent with wearable device recordings. Results show a trade-off between clean signal accuracy and noise robustness. CTRhythm achieved the highest performance on clean signals, which reflects Transformer encoder's ability to learn global rhythm patterns. However, it also showed the most severe degradation under noise. MFEGNet correctly identified the most AFib cases on clean signals (AFib recall = 0.81 ± 0.146) and its noise suppression mechanisms enabled the model to have the highest

accuracy and F1 score under all noise types. Furthermore, MGCNet retained the highest AFib recall under noisy conditions, likely because its cross-modal supervised contrastive loss results in better class discrimination even under signal distortion. Despite this, all three models degraded substantially under noise. Therefore, in cases where signal quality cannot be controlled, MFEGNet and MGCNet architectures would be more suitable. CTRhythm may be the best choice if a more robust denoising pipeline was integrated.

However, all model's performance degraded below clinically acceptable performance under noisy conditions. AFib recall was the most affected metric under noise, likely because noise introduces new frequency components that obscure features, such as the absence of the P wave, that are critical to AFib detection. As a result, many true AFib segments fall below the classification threshold and are predicted as non-AFib windows. This is clinically significant, since missed AFib episodes carry greater risk than false positive detections.

These findings suggest that strong performance on clean benchmark datasets does not necessarily guarantee reliable performance in real-world settings and highlight the need for effective denoising before classification. MA and EM are especially challenging to denoise, because their frequency content overlaps with the ECG signal and traditional methods have limited ability to remove them without obscuring diagnostically relevant features. Recently, diffusion-based models have demonstrated strong potential for ECG noise removal under the noise types evaluated in this study, and have outperformed conventional methods [7]. However, their effect on AFib detection performance remains unexplored. Therefore, future work should evaluate diffusion-based denoising as a preprocessing step and quantify its effect on AFib detection performance across noise types and architectures.

Acknowledgements

The authors are thankful for the high-performance computing resources provided by the Information Technology Open Access Center of Vilnius University.

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-ITP-25-9.

References

[1] Tsutsui, K., Biton Brimer, S., Ben-Moshe, N., Sellal, J.-M., Oster, J., Mori, H., Ikeda, Y., Arai, T., Nakano, S., Kato, R. and Behar, J.A. (2025) ‚SHDB-AF: A Japanese Holter ECG database of atrial fibrillation', Scientific Data, 12, 454. https://doi.org/10.1038/s41597-025-04777-4

[2] H. Li, J. Wei, J. Xiao, Y. Lai, M. Liu, S. Lv, X. Ouyang, Robust and generalizable atrial fibrillation detection from ECG using time-frequency fusion and supervised contrastive learning, arXiv preprint arXiv:2601.10202 (2026), https://arxiv.org/abs/2601.10202.

[3] Y. Liu, Z. Lin, P. Li, H. Song, J. Lu, P. Cao, Z. Yan, H. Li, Z. Huang, X. Li, Y. Yang, Y. Liang, P. Fang, P. Li, CTRhythm: Accurate atrial fibrillation detection from single-lead ECG by convolutional neural network and transformer integration (2024), https://doi.org/10.1101/2024.10.26.24316175.

[4] R. Masiliūnas, A. Dapkutė, J. Grigaitė et al., High prevalence of atrial fibrillation in a Lithuanian stroke patient cohort, Medicina 58 (6) (2022) 800, https://doi.org/10.3390/medicina58060800.

[5] Moody GB, Mark RG. MIT-BIH Atrial Fibrillation Database. Version 1.0.0. PhysioNet; 2000. doi:10.13026/C2MW2D

[6] G.B. Moody, W.E. Muldrow, R.G. Mark, A noise stress test for arrhythmia detectors, in: Comput. Cardiol. 11 (1984) 381–384.

[7] Šimėnas, S., Klepachevskyi, D., Medvedev, V., Treigys, P., Abramikas, Ž., Bernataviičienė, J.: Inter-patient evaluation of diffusion-based ECG denoising under standardized noise types. (2025)

[8] X. Wu, M. Yan, R. Wang, L. Xie, MFEG-Net: Multiscale feature enhanced gating network for atrial fibrillation detection, Comput. Methods Programs Biomed. 261 (2025) 108606, https://doi.org/10.1016/j.cmpb.2025.108606.

[9] Van Gelder, I.C., Rienstra, M., Bunting, K.V. et al. (2024) ‚2024 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS', European Heart Journal, 45(36), pp. 3314–3414. https://doi.org/10.1093/eurheartj/ehae176