Impact of Label Granularity on Photoplethysmography-Based Atrial Fibrillation Detection

Yangyang Zhao[1], Matti Kaisti[1], Olli Lahdenoja[1], Tero Koivisto[1]
[1]Department of Computing, University of Turku, Turku, Finland

Abstract

Photoplethysmography (PPG)-based atrial fibrillation (AF) detection has shown promising performance in wearable applications. However, most studies still use a binary AF versus non-AF setting, where sinus rhythm and different non-AF arrhythmias are grouped into a single non-AF class. This simplified setup may obscure important sources of misclassification, especially in elderly hospitalized patients with cardiovascular comorbidities. This study investigates how label granularity affects PPG-based AF detection performance and which clinically relevant non-AF rhythm groups are more likely to be misclassified as AF. We analyzed long-term clinical wrist-PPG recordings from 49 elderly hospitalized patients, segmented into non-overlapping 30-s windows and organized into hierarchical 2-, 4-, and 6-class rhythm-label settings. The same data split and training framework were used across label settings. On the full test set, AF AUC decreased only modestly from 0.821 to 0.807 and 0.799 as label granularity increased, whereas Macro-F1 fell from 0.700 to 0.350 and 0.240. AF false positives were concentrated in specific non-AF subgroups, with ectopic-like rhythms and the Couplet-Geminy family showing the highest within-group false-positive rates. These findings indicate substantial variation in AF false-positive susceptibility across non-AF rhythm groups, which is not evident from binary AF versus non-AF evaluation alone.

1. Introduction

Wearable photoplethysmography (PPG) is widely used for long-term cardiovascular monitoring and atrial fibrillation (AF) screening. Because PPG can be continuously measured using wrist-worn devices, many studies have investigated machine learning and deep learning approaches for PPG-based AF detection [1–3]. However, most AF-screening approaches use binary AF versus non-AF classification, grouping all non-AF rhythms into a single negative class.

This simplification may be particularly limiting in clinically complex populations, where non-AF rhythms such as ectopic beats, pauses, conduction abnormalities, ventricular rhythms, and supraventricular tachyarrhythmias can produce distinct pulse patterns in wrist PPG [1, 4]. Because pulse irregularity is not specific to AF, these rhythms can increase false-positive AF detections [5–8]. In wearable screening, irregular pulse alerts may therefore reflect rhythms other than AF and generally require ECG-based confirmation [4, 9].

Recent benchmark work has begun to extend PPG-based evaluation beyond AF and atrial flutter toward multiclass arrhythmia classification, highlighting the value of evaluating a broader spectrum of rhythm abnormalities [10]. However, it remains unclear how progressively refining the non-AF label space affects AF-specific performance and the interpretation of false-positive errors.

This study evaluates how progressive rhythm-label refinement affects performance estimates and error interpretation in wrist-PPG AF detection. We compare hierarchical 2-, 4-, and 6-class label settings using the same patient-level split, preprocessing, model architecture, and training protocol. We investigate two questions: (1) how label refinement affects AF-specific and macro-averaged performance, including robustness across motion-percentile subsets, and (2) which non-AF rhythm groups are most often misclassified as AF.

2. Methods

2.1. Dataset and Preprocessing

This study used the Moore4Medical wrist-PPG dataset, whose data collection protocol and signal acquisition setup have been described previously [11]. The dataset comprises long-term recordings from 49 elderly hospitalized patients with cardiovascular comorbidities. Wrist PPG and synchronized tri-axial accelerometer (ACC) signals were recorded at 32 Hz, together with reference Holter ECG for rhythm annotation. PPG recordings were divided into non-overlapping 30-s windows (960 samples) and preprocessed using a third-order Butterworth band-pass filter at 0.5–8 Hz, followed by zero-mean, unit-variance standardization. For motion-stratified evaluation, ACC signals were segmented into the same 30-s windows. The ACC magnitude was calculated as $\sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$, and

the variance of the magnitude within each window was used as the motion score, following our previous work [3]. Test windows were ranked from lowest to highest motion and evaluated using cumulative subsets containing the lowest-motion 10%, 20%, . . . , 100% of windows.

2.2. Progressive Label Refinement

Detailed ECG annotations were cleaned and mapped according to the hierarchical definitions in Table 1. AF-related event labels were merged into AF, while standalone interval or heart-rate labels without independent rhythm meaning were excluded. The 2-, 4-, and 6-class settings were selected to represent progressively finer levels of non-AF rhythm heterogeneity while retaining AF as the target class. The 2-class setting reflects conventional AF screening, whereas the 4-class setting introduces the first clinically meaningful separation of ectopic-like rhythms from other abnormal non-AF rhythms. The 6-class setting further distinguishes PAC/SVE from repetitive ectopic patterns and separates pause/conduction abnormalities from ventricular- or rate-related rhythms, providing finer rhythm-specific analysis without excessively fragmenting rare classes. After sample-level mapping, each non-overlapping 30-s window was assigned the most frequent mapped rhythm class within that window. Because majority voting was applied after task-specific label grouping, mixed-rhythm windows could receive different final labels as the label space was refined, resulting in small differences in class counts across the 2-, 4-, and 6-class settings. The same patient-level train-test split was used for all tasks.

2.3. Model Training

The classifier followed the residual temporal-attention architecture used in our previous work [3], using only the PPG waveform as input. ACC was used exclusively for motion-stratified evaluation. The same architecture and training hyperparameters were used across label settings, with the output layer adapted to the number of classes. Models were trained for 60 epochs using Adam with a learning rate of 0.0001, batch size 64, and weight decay 10^{-5}. Weighted cross-entropy (WCE) was used to address class imbalance, with class weights computed from the training set for each task. Each experiment was repeated with three random seeds. Models were implemented in PyTorch 2.3.0 and trained on an NVIDIA TITAN RTX GPU.

2.4. Evaluation

Performance was assessed using AF-specific metrics (AUC, sensitivity, precision, and F1-score) and multiclass metrics (Macro-F1, weighted-F1, accuracy, and balanced accuracy). AF AUC was computed one-vs-rest against all non-AF classes.

AF false positives were further analyzed by their true non-AF rhythm group. For each group, we report the within-group AF false-positive rate and the corresponding number of false-positive windows. For this analysis, predicted probabilities were averaged across the three random seeds before assigning the final class. For motion-stratified evaluation, metrics were computed over cumulative p10–p100 motion subsets and averaged across the three random seeds. The p100 value represents performance on the full test set, while medians and interquartile ranges across p10–p100 summarize performance variation with increasing motion.

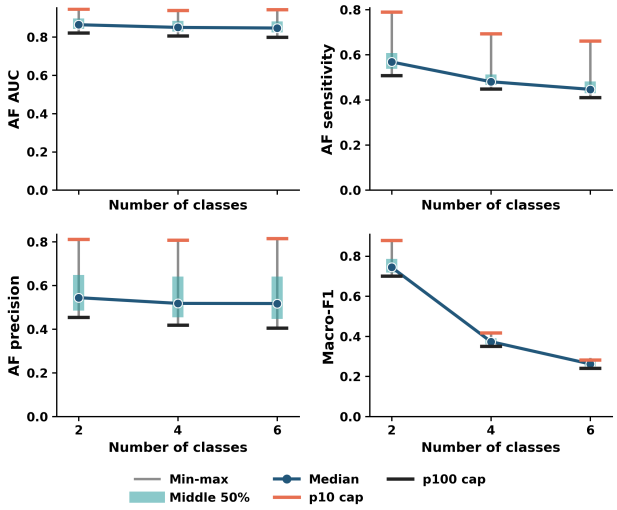

Figure 1. Performance across hierarchical 2-, 4-, and 6-class rhythm-label settings under increasing motion. Test windows were ranked from lowest to highest motion and evaluated cumulatively from p10 (the lowest-motion 10% of windows) to p100 (the full test set). For each class setting, the gray vertical line shows the minimum–maximum range across p10–p100, the cyan box the interquartile range, and the blue circle the median; blue lines connect the medians across label settings. Orange and black caps indicate the p10 and p100 values, respectively.

3. Results

3.1. Performance Across Label Granularities and Motion Levels

Increasing label granularity affected multiclass performance more strongly than AF discrimination (Figure 1). At p100, AF AUC decreased modestly from 0.821 in the 2-class task to 0.807 and 0.799 in the 4- and 6-class tasks, whereas Macro-F1 decreased from 0.700 to 0.350 and

Table 1. Hierarchical 2-, 4-, and 6-class rhythm-label definitions and segment-level class distributions used in the main analysis. Colors indicate the correspondence of rhythm labels across progressively refined classification settings.

Task	Class	Mapped rhythm labels	Segments, n (%)
2-class	non-AF	Sinus, Brady, ST-, AV1, ST+, Tachy, ST(ref), AV2/I; PAC/SVE; COUP, COUP(mf), GEM, GEM(mf), SALV, SALV(mf); Arrest, Block, AV2/II; IVR, VT, VT(mf), SVT	95,782 (86.0%)
	AF	AF and AF-related event markers	15,650 (14.0%)
4-class	SR	Sinus, Brady, ST-, AV1, ST+, Tachy, ST(ref), AV2/I	86,576 (77.7%)
	AF	AF and AF-related event markers	15,709 (14.1%)
	Ectopic-like	PAC/SVE; COUP, COUP(mf), GEM, GEM(mf), SALV, SALV(mf)	7,882 (7.1%)
	Non-ectopic Other	Arrest, Block, AV2/II; IVR, VT, VT(mf), SVT	1,265 (1.1%)
6-class	SR	Sinus, Brady, ST-, AV1, ST+, Tachy, ST(ref), AV2/I	86,588 (77.7%)
	AF	AF and AF-related event markers	15,712 (14.1%)
	PAC-SVE family	PAC/SVE	1,682 (1.5%)
	Couplet-Geminy family	COUP, COUP(mf), GEM, GEM(mf), SALV, SALV(mf)	6,192 (5.6%)
	Pause-Conduction	Arrest, Block, AV2/II	728 (0.7%)
	Ventricular-Rate-related	IVR, VT, VT(mf), SVT	530 (0.5%)

0.240. AF sensitivity decreased from 0.508 to 0.449 and 0.411, while AF precision decreased from 0.455 to 0.418 and 0.406. Across the cumulative p10–p100 motion subsets, AF AUC remained relatively stable, whereas AF sensitivity, AF precision, and Macro-F1 varied more noticeably with increasing motion.

3.2. Non-AF Rhythm Groups Misclassified as AF

Figure 2 shows substantial differences in AF false-positive rates across non-AF rhythm groups. In the 4-class task, Ectopic-like rhythms had the highest within-group false-positive rate (26.1%). In the 6-class task, the Couplet-Geminy family showed the highest rate (35.8%), whereas PAC-SVE and Ventricular-Rate-related rhythms showed much lower rates. SR showed a lower within-group AF false-positive rate (6.9% in the 6-class task) but produced the largest absolute number of false-positive windows because of its high prevalence.

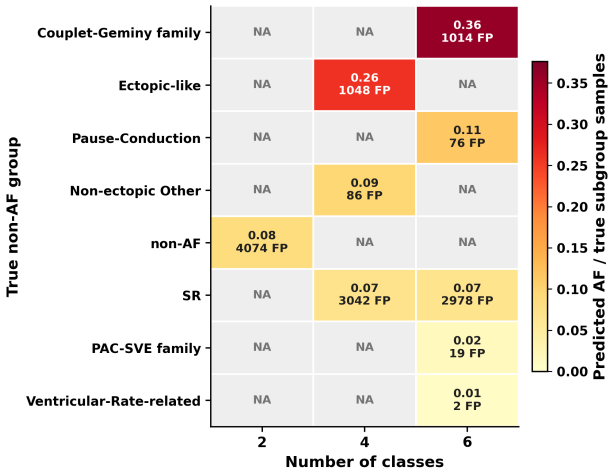

Figure 2. AF false-positive patterns across hierarchical 2-, 4-, and 6-class label settings. Rows denote the true non-AF classes or rhythm groups and columns the label settings. Each colored cell shows the within-group AF false-positive rate, defined as the proportion of windows in that rhythm group predicted as AF; the corresponding number of false-positive windows is shown below. "NA" indicates that the rhythm group is not defined at that label granularity.

4. Discussion

Progressive rhythm-label refinement affected multiclass performance more strongly than AF discrimination. AF AUC remained relatively stable across the 2-, 4-, and 6-class settings, whereas Macro-F1 decreased substantially as the non-AF label space was refined. This suggests that finer labels expose difficulties in distinguishing heterogeneous non-AF rhythms that are not apparent in binary AF-versus-non-AF evaluation. Recent PPG benchmark work has similarly shown that broader multiclass arrhythmia classification is more challenging than AF detection alone [10]. Similar concerns have been raised in mobile ECG, where Tuboly et al. [8] noted that evaluations focused mainly on AF and sinus rhythm may provide limited information about false positives caused by other arrhythmias.

AF false positives were not uniformly distributed across non-AF rhythm groups. SR produced the largest absolute number of false-positive windows because of its high prevalence, whereas Ectopic-like rhythms showed a much higher within-group false-positive rate. This agrees with previous PPG studies reporting reduced AF-detection

specificity in the presence of frequent premature contractions [6, 7]. Further refinement showed that the Couplet-Geminy family had the highest within-group false-positive rate (35.8%), whereas the PAC-SVE family showed much lower confusion. Bacevicius et al. [6] reported no reduction in accuracy for bigeminy/trigeminy in a detector incorporating dedicated bigeminy suppression, suggesting that rhythm-specific false-positive patterns may also depend on the detection algorithm. The difference between the two ectopic subgroups illustrates information obscured when they are combined into a broader category.

Motion-stratified evaluation showed that AF AUC remained relatively stable across cumulative motion subsets, whereas sensitivity, precision, and Macro-F1 varied more noticeably.

This study has several limitations. The analysis was based on a single clinical wrist-PPG dataset with 49 patients, and some refined rhythm groups contained relatively few samples. Validation on larger and more diverse datasets is therefore needed, and the observed subgroup-specific patterns may depend on the model and training strategy. In addition, task-specific label grouping before window-level majority voting caused a small number of mixed-rhythm windows to receive different final labels across settings. Overall, finer rhythm labels provided additional information about subgroup-specific AF false positives that was not visible in the binary setting.

Acknowledgements

This study was part of the CARE-DETECT clinical trial (ClinicalTrials.gov ID: NCT05351775), approved by the Ethics Committee of the Hospital District of Southwest Finland and conducted in accordance with the Declaration of Helsinki, with written informed consent obtained from all participants. This work was funded by the Moore4Medical project, supported by the ECSEL JU and Business Finland under grant agreements H2020-ECSEL-2019-IA-876190 and 7215/31/2019, and by the ITEA project RM4HEALTH, supported by Business Finland under Grant 8139/31/2022. Yangyang Zhao acknowledges support from the Nokia Foundation and the Finnish Foundation for Cardiovascular Research through the Maria Tapanainen Fund.

References

[1] Pereira T, Tran N, Gadhoumi K, Pelter MM, Do DH, Lee RJ, Colorado R, Meisel K, Hu X. Photoplethysmography based atrial fibrillation detection: a review. NPJ digital medicine 2020;3(1):3.

[2] Zhao Y, Lahdenoja O, Elnaggar I, Vasankari T, Jaakkola S, Kiviniemi T, Airaksinen J, Kaisti M, Koivisto T. Validation of an mems-based pressure sensor system for atrial fibrillation detection from wrist and finger. IEEE Sensors Journal 2025;25(13):25615–25624.

[3] Zhao Y, Kaisti M, Lahdenoja O, Koivisto T. Motion-robust multimodal fusion of ppg and accelerometer signals for three-class heart rhythm classification. In Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2025; 171–175.

[4] Joglar JA, Chung MK, Armbruster AL, Benjamin EJ, Chyou JY, Cronin EM, Deswal A, Eckhardt LL, Goldberger ZD, Gopinathannair R, et al. 2023 acc/aha/accp/hrs guideline for the diagnosis and management of atrial fibrillation: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. Journal of the American College of Cardiology 2024; 83(1):109–279.

[5] Han D, Bashar SK, Mohagheghian F, Ding E, Whitcomb C, McManus DD, Chon KH. Premature atrial and ventricular contraction detection using photoplethysmographic data from a smartwatch. Sensors 2020;20(19):5683.

[6] Bacevicius J, Abramikas Z, Dvinelis E, Audzijoniene D, Petrylaite M, Marinskiene J, Staigyte J, Karuzas A, Juknevicius V, Jakaite R, et al. High specificity wearable device with photoplethysmography and six-lead electrocardiography for atrial fibrillation detection challenged by frequent premature contractions: Doublecheck-af. Frontiers in cardiovascular medicine 2022;9:869730.

[7] Liao MT, Yu CC, Lin LY, Pan KH, Tsai TH, Wu YC, Liu YB. Impact of recording length and other arrhythmias on atrial fibrillation detection from wrist photoplethysmogram using smartwatches. Scientific Reports 2022;12(1):5364.

[8] Tuboly G, Kozmann G, Kiss O, Merkely B. Atrial fibrillation detection with and without atrial activity analysis using lead-i mobile ecg technology. Biomedical Signal Processing and Control 2021;66:102462.

[9] Perino AC, Gummidipundi SE, Lee J, Hedlin H, Garcia A, Ferris T, Balasubramanian V, Gardner RM, Cheung L, Hung G, et al. Arrhythmias other than atrial fibrillation in those with an irregular pulse detected with a smartwatch: findings from the apple heart study. Circulation Arrhythmia and Electrophysiology 2021;14(10):e010063.

[10] Moulaeifard M, Aston PJ, Charlton PH, Strodthoff N. Deriving health metrics from the photoplethysmogram: Benchmarks and insights from mimic-iii-ext-ppg. arXiv preprint arXiv260321832 2026;.

[11] Zhao Y, Lahdenoja O, Sandelin J, Seifizarei S, Anzanpour A, Lehto J, Nuotio J, Jaakkola J, Relander A, Vasankari T, et al. Assessing the impact of signal quality on heart rate detection from long-term clinical wrist ppg under varying cardiac rhythms. Biomedical Signal Processing and Control 2026;112:108688.

Address for correspondence:

Yangyang Zhao
Digital Health Technology Group, Department of Computing
University of Turku, Vesilinnantie 5, 20014 Turku, Finland
yazhao@utu.fi