Window-level AF detector: feasibility report

Window-level AF detector: feasibility report

Date: 14 September 2026, draft; section 7 will be completed after the engine-native AFDB test. All scripts, features, models and reports are on elon in `/home/sabir/ecg_engine/feasibility/`; the full experiment log is `RESULTS.md`. Published summary: https://ecg-algo.sabir.pro/af-window-detector/

1. Question

Can an AF detector that passes a standard validation (ANSI/AAMI EC57) on data never used in training be built in a short time, keeping the existing beat detector and beat measurements and replacing only the decision stage?

2. What was built

A classifier that decides on a window of consecutive beats rather than on single beats as the previous model did.

- **Input:** RR intervals from the engine's beats (from PhysioNet reference annotations for the public databases), plus, where the engine has run, its beat measurements: P-wave presence and amplitude, QRS width, deformation index, quality index, R-S amplitude, ventricular fraction from its beat classifier, and the fraction of the window marked out of signal.
- **RR features (10 per window):** heart rate, normalised RMSSD, coefficient of variation, pNN50, Shannon entropy of the RR histogram, sample entropy, turning point ratio, Poincare SD1/SD2, irregularity index, normalised median absolute deviation.
- **Model:** histogram gradient boosting (scikit-learn), depth 4, 300 iterations, with monotonic constraints so that more RR irregularity can only raise the AF probability. About 0.5 MB, against 189 MB for the previous Random Forest.
- **Physiological guard:** a window with irregularity index below 0.1 and RR CV below 0.05, practically regular rhythm, is never called AF.
- **Windows:** 32 beats stepping 8 for long recordings; 16 beats stepping 4 for 30-second strips, because a 32-beat window needs 33 beats and a strip below about 66 beats per minute does not have them.
- **Strip decision:** AF if the highest window probability is at least 0.8, not AF if at most 0.5, otherwise inconclusive. Strips whose beats are implausible (RR outside 0.2-3.0 s or more than 200 beats per minute) yield no window and are inconclusive.
- **Episodes (long recordings):** hysteresis over windows, on 0.6 / off 0.3, minimum one window; parameters chosen on separate tuning records.

3. Data and protocol

Set	Role	Size	Beats	Labels
		84 records, ~24 h each	reference	

Set	Role	Size	Beats	Labels
Long-Term AF Database (PhysioNet)	training 60 records, tuning 24 records			interval, cardiologist-reviewed
MIT-BIH Arrhythmia Database	training	48 records	engine	interval, PhysioNet
Clinical AF set (Kyiv clinic)	training	218 segments, 52 patients	engine	whole segment, physician
Sensor strips, 97 patients	training	6,156 windows	engine	whole strip, original developer
MIT-BIH AF Database (AFDB)	**test, never used in training**	23 records, 10 h each	PhysioNet detections	interval, cardiologist
Sensor strips, 47 patients	**test, patients disjoint from training**	755 strips	engine	whole strip, original developer

Sensor patients are split by a deterministic hash of the patient identifier; zero overlap was verified. Windows are labelled AF at 80 % or more coverage by an annotated AF interval, not AF at 20 % or less, and discarded in between. Expert-flagged artefact segments are excluded from training. EC57 scoring (episode and duration sensitivity and positive predictivity) is a Python implementation of the epicmp definitions; a cross-check with the WFDB tools is pending.

4. Training-data audit

Out-of-fold predictions by patient over 111,629 training windows plus physiological checks found and fixed:

1. **A loader bug.** About two hours at the start of one LTAFDB record precede its first rhythm annotation and had been treated as non-AF although the rhythm is irregular. 11,601 windows removed; the loader now drops beats before the first rhythm marker.
2. **Unflagged artefacts.** 754 training windows with engine beat rates above 200 per minute, mostly in one sensor set. Dropped by rule.
3. **Impossible labels.** 205 windows labelled AF with practically regular RR intervals, from whole-segment labels covering non-AF stretches. Dropped by rule.

Two LTAFDB records with atrial bigeminy and supraventricular tachycardia that resemble AF on RR intervals are genuine hard negatives and were kept. Model-based disagreements were not used for cleaning, to avoid circularity. Retraining on the corrected data raised cross-validated specificity from 93.5 % to 95.5 %.

5. Results

5.1 Sensor strips, end to end, all 755 held-out strips, expert flags not used

Metric	Value
Inconclusive	31.8 % (28 % of this set is artefact by the expert’s flags; 199 strips have no plausible window, 41 fall in the probability band)

Metric	Value
Classified strips	515
Sensitivity	98.7 %
Specificity	98.6 %
PPV / NPV	96.9 % / 99.4 %
Accuracy	98.6 %
Errors	5 false positives, 2 misses

This is the level reported by cleared consumer ECG apps (about 98 % / 99.6 % with roughly 12 % unclassifiable) on a set with far more artefact. A separately trained artefact detector was evaluated and rejected: it reproduced the expert flags only at 86 % / 85 % and added refusals without improving accuracy. Noise handling in the product is the plausibility gate, the engine's out-of-signal and quality measures as features, and the probability band.

5.2 Public benchmark, AFDB, never used in training, RR features only

Metric	Value
Window sensitivity / specificity / PPV	95.6 % / 96.7 % / 95.8 %
EC57 duration sensitivity / positive predictivity	96.5 % / 92.0 % (94.0 % with the two-scale cascade)
EC57 episode sensitivity / positive predictivity	93.9 % / 46.8 % (96.2 % / 45.4 % with 16-beat windows; 93.5 % / 55.7 % with the cascade)
Episodes of 30 s or longer detected	100 %
Episodes 10-30 s detected	89 % (32-beat), 100 % (16-beat)
Episodes under 10 s detected	50-61 %

False episodes number 410 with a median length of 33 s; 63 exceed one minute and two exceed five minutes. Restricted to detected episodes of at least 60 s, positive predictivity is 83.9 %. The false windows are statistically indistinguishable from AF on RR intervals (irregularity index 0.65 vs 0.68, RR CV 0.205 vs 0.208) and sit in eight records with frequent ectopy. Episode counting weights a 30-second false alarm the same as a 6-hour true episode, which is why this metric is low while duration precision is 92-94 %. It does not apply to a 30-second strip claim.

5.3 Sensitivity by heart-rate band (window level)

Set	< 60 bpm	60-100	100-130	> 130
Sensor held-out, Se	n/a	98.1 %	98.4 %	97.4 %
AFDB, Se	99.3 %	95.4 %	98.1 %	95.7 %
AFDB, Sp	99.9 %	97.6 %	92.6 %	(65 windows)

5.4 Comparison with the previous model

Same 424 clean held-out sensor strips, per strip: previous model's delivered output Se 99.4 %, Sp 96.1 %, PPV 94.3 %, with all 47 patients in its training set and labels written with its output at hand; ours (32-beat model, patients unseen) Se 94.5 %, Sp 99.2 %, PPV 98.7 %; the 16-beat product configuration on the broader 515-strip set Se 98.7 %, Sp 98.6 %. On

MIT-BIH by EC57, the previous model on its own training data: episode 91.6 / 58.6, duration 92.9 / 86.5; ours on AFDB, never seen: duration 96.5 / 92.0.

Beat detection, unchanged engine, MIT-BIH, bxb rule: Se 98.1 %, +P 98.7 %.

6. Conclusion on feasibility

Feasible. On sensor patients never seen in training the detector reaches 98.7 % sensitivity and 98.6 % specificity per strip, and on a public benchmark it never saw, 96.5 % duration sensitivity and 92 % duration precision by EC57 with every AF episode of clinical length detected. The model is small, explainable, deterministic, constrained by physiology, and its data pipeline has been audited.

Targets proposed for the 30-second strip claim and status:

Metric	Acceptance	Current	Status
Strip sensitivity, held-out patients	>= 95 %, lower 95 % CI >= 90 %	98.7 %	met
Strip specificity	>= 97 %, lower 95 % CI >= 95 %	98.6 %	met
Inconclusive rate	<= 15 % on a clinical-grade set	31.8 % on an artefact-heavy set	needs a clean study set
AFDB duration Se / +P	>= 95 % / >= 90 %	96.5 % / 92-94 %	met
AFDB episode Se, episodes >= 30 s	>= 95 %	100 %	met
Beat detection Se / +P	>= 98 % / 98 %	98.1 % / 98.7 %	met

What remains before a submission, in order: cardiologist re-annotation of the existing strips; a clinical validation study on the device with a simultaneous 12-lead reference; freezing and versioning the algorithm with tests and a design description; a written validation protocol followed by a single pre-specified run scored with the WFDB tools; software, machine-learning and labeling documentation.

Caveats: the sensor labels are the original developer’s until re-annotated; 165 AF-positive strips give wide confidence intervals; AFDB was explored before the protocol was frozen.

7. Engine-native features on AFDB

Pending: the engine is processing all 14,041 one-minute segments of AFDB so that P-wave, QRS and quality features can be tested on this benchmark. Expected effect: higher episode precision in the ectopy-rich records. This affects a future continuous-monitoring claim only.