Honest vs. Naive Accuracy in HRV-Based Arrhythmia Classification

A Subject-Level Validation Study Using the MIT-BIH Arrhythmia Database

Akif Khan[1] and **Antonios E. Marsellos[2]**

[1]Sanford H. Calhoun High School, 1786 State St, Merrick, NY 11566
[2]Dept. Geology, Environment and Sustainability, Hofstra University, Hempstead, NY, USA

ABSTRACT

Premature ventricular contractions can be detected from heart rate variability, the variation in timing between successive heartbeats. Five recordings from the MIT-BIH Arrhythmia Database were analyzed in R, using expert cardiologist beat annotations to build RR intervals directly rather than detecting the beats again. HRV features over sliding 30-beat windows trained a Random Forest classifier. The central question: does the train/test split change reported accuracy? It does, by 17.7 points.

HONEST ACCURACY

79.6%

Subject-level split. Records 103 and 208 held out entirely, so the model never saw those patients.

NAIVE ACCURACY

97.3%

Random-window split. Windows from the same patients land on both sides of the split.

POINTS OVERSTATED

17.7

Accuracy added by data leakage alone, with no change to the model or the data.

Background

Silent onset. Cardiac arrhythmias, including PVCs, can occur without obvious external symptoms, making automated detection valuable for patient monitoring.

HRV as a signal. The beat-to-beat variation in timing between heartbeats changes measurably during arrhythmic events.

Leakage. Many published arrhythmia classifiers report very high accuracy, but some of that accuracy comes from allowing beats from the same patient into both the training and test sets.

Honest design. A subject-level (inter-patient) split, where no patient's beats appear on both sides, gives a realistic estimate of performance on a new, unseen patient.

Leakage is the gap between paper and bedside.

Method and Materials

Data. Five real MIT-BIH recordings (100, 103, 106, 119, 208) at 360 Hz, each with expert cardiologist beat annotations.

Beat extraction. Expert-labeled beat positions taken straight from the annotation files rather than a custom R-peak detector, preserving ground-truth accuracy.

HRV features. RR intervals between consecutive labeled beats; meanRR, meanHR, SDNN, RMSSD, pNN50 and CVrr over sliding 30-beat windows.

Classifier. Random Forest, 400 trees, separating normal-rhythm windows from arrhythmia windows (any window with at least one PVC).

Validation. Trained on 100, 106 and 119; records 103 and 208 held out entirely. A naive random-window split was run for comparison.

Software. RStudio, the randomForest package and base R.

Only the split changed, not the model.

Results

Honest split. 79.6% accuracy, 100% specificity and 64.6% sensitivity on 167 held-out test windows.

Naive split. 97.3% accuracy on the same data, a 17.7-point overstatement.

Waveform. PVCs show a wider, deeper averaged waveform and arrive earlier than normal beats (Fig. 2).

Tachogram. The PVC-heavy record shows short-long RR jumps that are absent from the normal-heavy record (Fig. 3).

Feature ranking. RMSSD (33.8) and SDNN (29.7) ranked most important; meanRR (2.4) and meanHR (1.6) contributed least (Fig. 5).

Same data, same model, two different numbers.

Discussion and Conclusion

Performance. HRV features from expert-annotated RR intervals, with a Random Forest, separate normal rhythm from PVC-containing rhythm at 79.6% accuracy under an honest split.

Validation matters. The 17.7-point gap between honest and naive accuracy shows the split matters as much as the model choice.

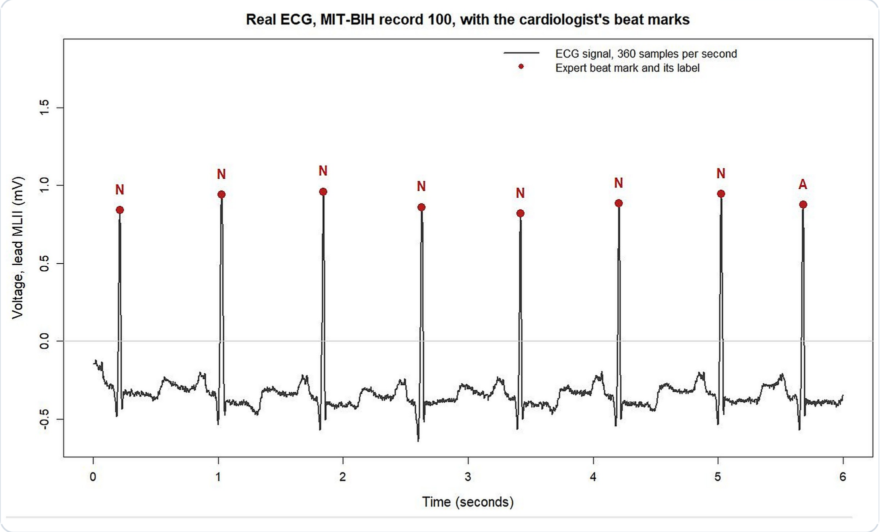
Conservative model. Perfect specificity (100%) with moderate sensitivity (64.6%): it rarely raises false alarms but misses some true arrhythmia windows.

What it keys on. Irregularity features outrank rate features, so the classifier reads beat-timing irregularity rather than heart rate itself.

Future work. Scale to more of the 48 MIT-BIH records, compare against an artificial neural network, and build an R Shiny app for real-time classification.

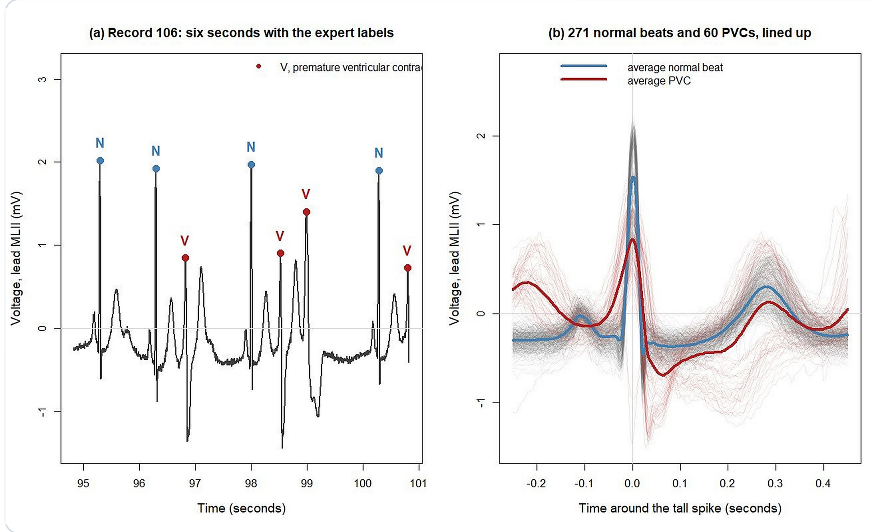
Validation design is a result, not a footnote.

FIG. 1



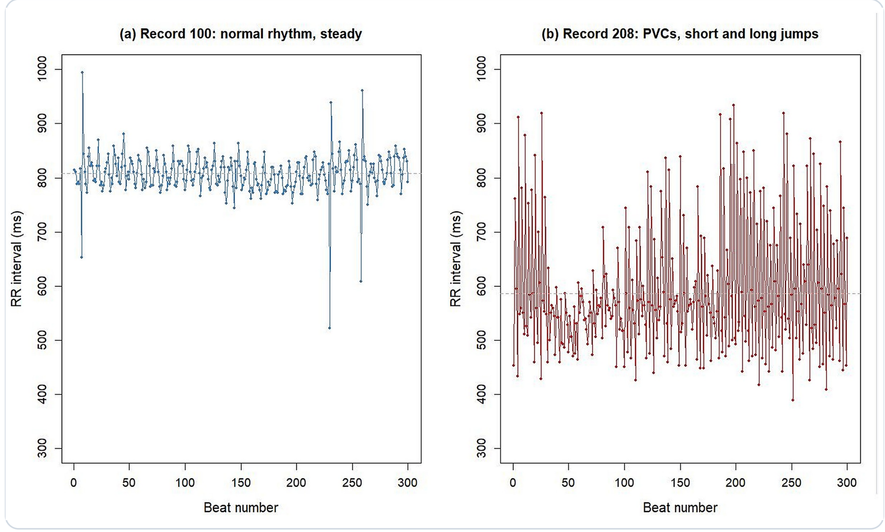
Raw lead-MLII ECG from record 100 with the cardiologist's beat marks overlaid. These expert beat positions are the ground truth used to build every RR interval in this study.

FIG. 2



Record 106: six seconds with expert labels (left), and 271 normal beats against 60 PVCs lined up on their tall spike (right). The average PVC is wider and dips deeper than the average normal beat.

FIG. 3



RR interval series (tachograms) for normal-heavy record 100 (left) and PVC-heavy record 208 (right). Record 208 shows the short-long jumps caused by an early PVC followed by a compensatory pause.

FIG. 4

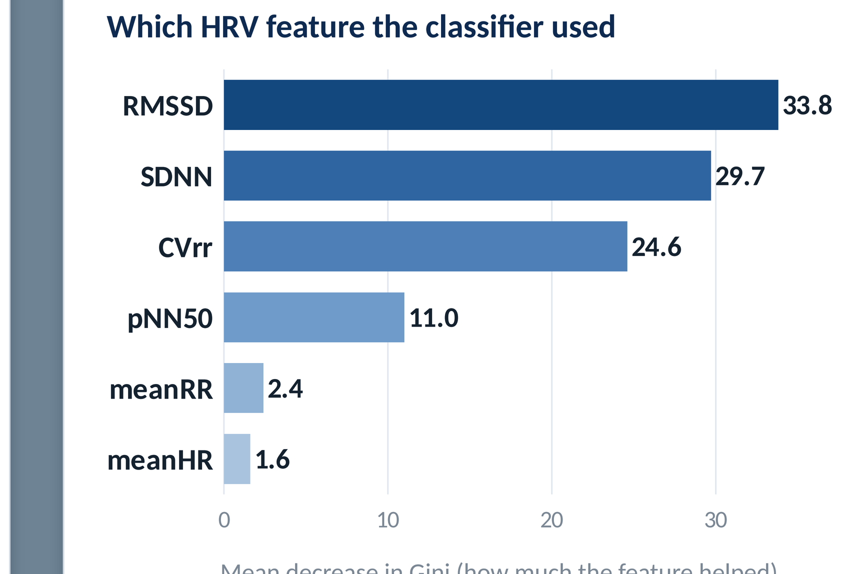
ACTUAL LABEL FROM THE CARDIOLOGIST

	NORMAL	ARRHYTHMIA
Predicted NORMAL	71 correct (42.5% of test windows)	34 missed (20.4% of test windows)
Predicted ARRHYTHMIA	0 false alarm (0.0% of test windows)	62 correct (37.1% of test windows)

ACCURACY	SENSITIVITY	SPECIFICITY
79.6%	64.6%	100.0%

Confusion matrix under the subject-level split: trained on records 100, 106 and 119, tested on held-out records 103 and 208. A naive random-window split scores 97.3% on this same data.

FIG. 5



Importance of each HRV feature to the classifier. RMSSD and SDNN, both measures of beat-to-beat irregularity, rank highest, well ahead of mean RR interval and mean heart rate.

REFERENCES

1. Moody GB, Mark RG (2001). The impact of the MIT-BIH Arrhythmia Database. IEEE Eng. Med. Biol. Mag., 20(3), 45-50.
2. Goldberger AL, Amaral LAN, Glass L, et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation, 101(23), e215-e220.
3. de Chazal P, O'Dwyer M, Reilly RB (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng., 51(7), 1196-1206.
4. Task Force of the ESC and NASPE (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation, 93(5), 1043-1065.
5. Breiman L (2001). Random Forests. Machine Learning, 45(1), 5-32.
6. Liaw A, Wiener M (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.

ACKNOWLEDGEMENTS

Thank you to Dr. Antonios E. Marsellos and the Department of Geology, Environment and Sustainability at Hofstra University for providing the tools to conduct this research.