



Interpretable active learning in support of clinical data annotation (#797)

D. Wang¹, M. Fiterau¹, A. Dubrawski¹, M. Hravnak², G. Clermont³, M.R. Pinsky³

Auton Lab, Carnegie-Mellon University¹
Schools of Nursing² and Medicine³, University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Funding: NIH NINR R01NR013912; NSF IIS-0911032



INTRODUCTION

Vital Sign Monitoring Systems issue alerts whenever one of the signals is abnormal. However, a large percentage of these alerts are in fact due to equipment malfunction or misapplication of sensors. It is vital to distinguish true alerts from artifacts in order to reduce alarm fatigue. We propose to automatically adjudicate artifacts, however, training a classifier usually requires a supply of adjudicated data to work. Annotation of large amounts of clinical data consumes valuable time of expert clinicians. We propose to determine whether active learning (AL) can be used to reduce expert effort and, to this end, we prototyped a protocol to collect reliable training data and a framework to build adjudication models.

DESIGN

Prospective study recruited 308 admissions to a 24-bed surgical stepdown unit over 8 weeks.

INCLUSION CRITERIA

- Adults age>21
- Continuous noninvasive monitoring

EXCLUSION CRITERIA

- None—the full census of patients in the 8-week timeframe were included.

METHODS

A. Monitoring

Noninvasive VS monitoring data recorded at a frequency of 1/20Hz consisted of heart rate (HR), respiratory rate (RR; bioimpedance), noninvasive (oscillometric) systolic (SBP) and diastolic (DBP) blood pressure, and peripheral oximetry (SpO₂).

B. Event Detection

VS events were detected as any VS violation of stability thresholds (HR< 40 or >140 bpm, RR< 8 or >36 bpm, SBP < 80 or >200 mmHg, DBP>110 mmHg, SpO₂< 85%). The first VS signal to exceed the stability threshold determines the type of the alert event.

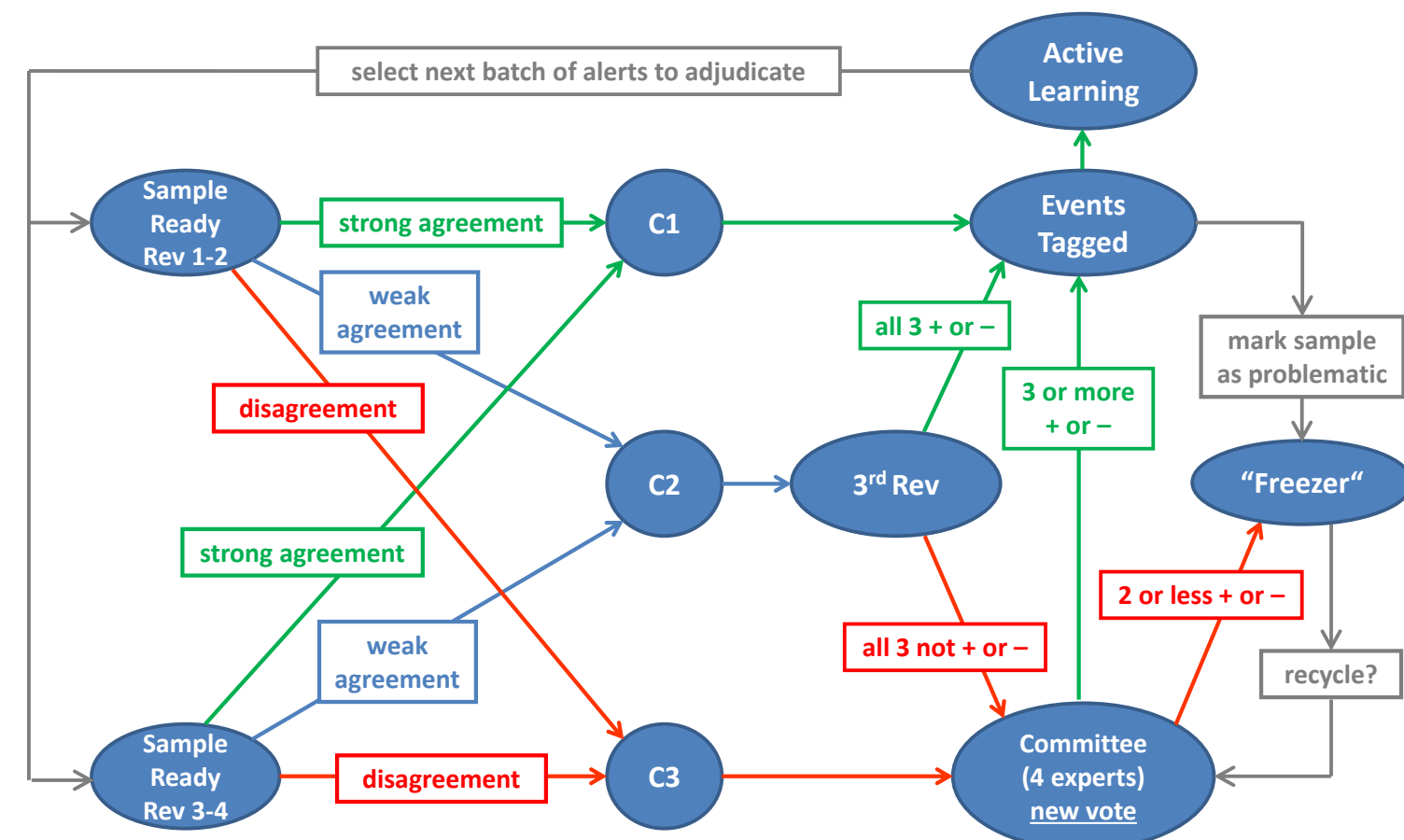
C. Feature Extraction from Vital Sign Time Series

Features computed, for each VS signal, during span of each alert, and a short window of 4 minutes preceding alert onset. Features include common statistics of each VS: mean, std. dev., minimum, maximum, gradients, min and max of first order differences, duty cycle etc.

METHODS

D. Expert Annotation Workflow

Time plots of all VS parameters during events were visually adjudicated by a group of 4 reviewers as real alerts or artifact according to the protocol:



Each alert is initially reviewed by two experts which classify it as genuine or artifact with some level of confidence.

Strong Expert Agreement: if the two initial reviewers agree on the alert with high confidence, this label is assigned, added to the repository, and treated as ground truth by the machine learning system.

Weak Expert Agreement: if the confidence of the reviews is not high, a 3rd reviewer annotates the alert independently; in the case of an agreement, the label is added to the repository, whereas if the arbiter disagrees, the decision regarding the alert is made by a committee of 4 experts.

Expert Disagreement: if the initial reviewers disagree, the alert is given directly to the committee.

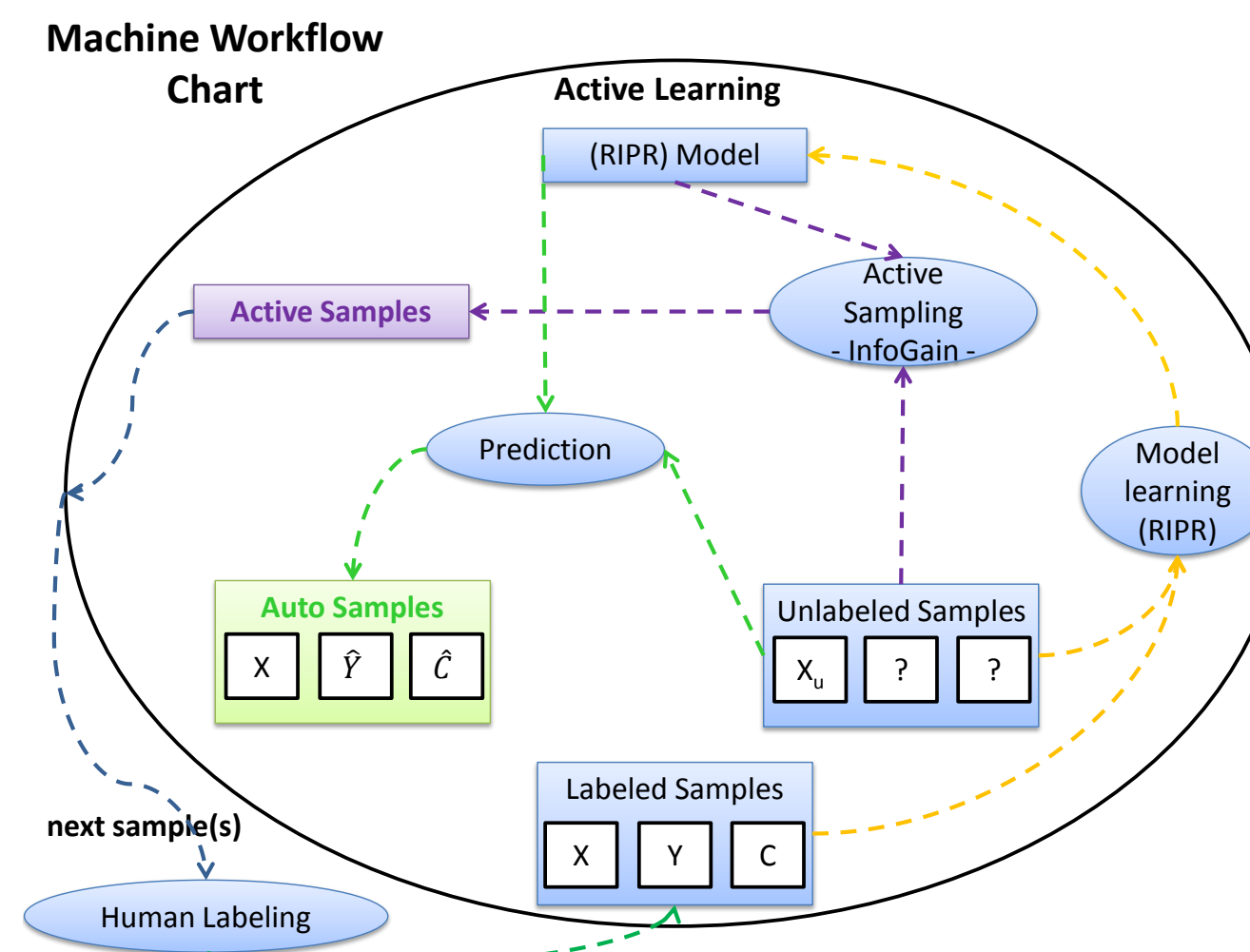
Committee Vote: the committee of 4 experts discusses the alerts on which a label could not be established in the previous stages. Each expert then adjudicates the alert. The label is determined by a majority vote. In the case that opinions are divided, the sample is marked as problematic, not used in training, and placed in a "freezer", to be disambiguated at a future time.

E. Batch Adjudication

A committee of experts adjudicated the initial batch (201) of these alerts labeling them as **true instabilities (133)**, **artifacts (39)** or unclear (29). These were used to train a Random Forest (RF) classifier for artifact adjudication. At the conclusion of each of the 3 cycles of expert adjudication, yielding 37, 43 and 33 annotated cases, respectively. 149 alerts could not be adjudicated due to expert disagreement. We measured the number of still unlabeled data that cannot be confidently adjudicated by the respective models

E. Training Procedure using Active Learning

We used a method derived from (Fiterau, Dubrawski: Projection Retrieval for Classification, NIPS 2012) to select data that maximizes the expected information gain and presents it in a human-interpretable fashion, and compared it against a RF classifier that selects the most uncertain data.

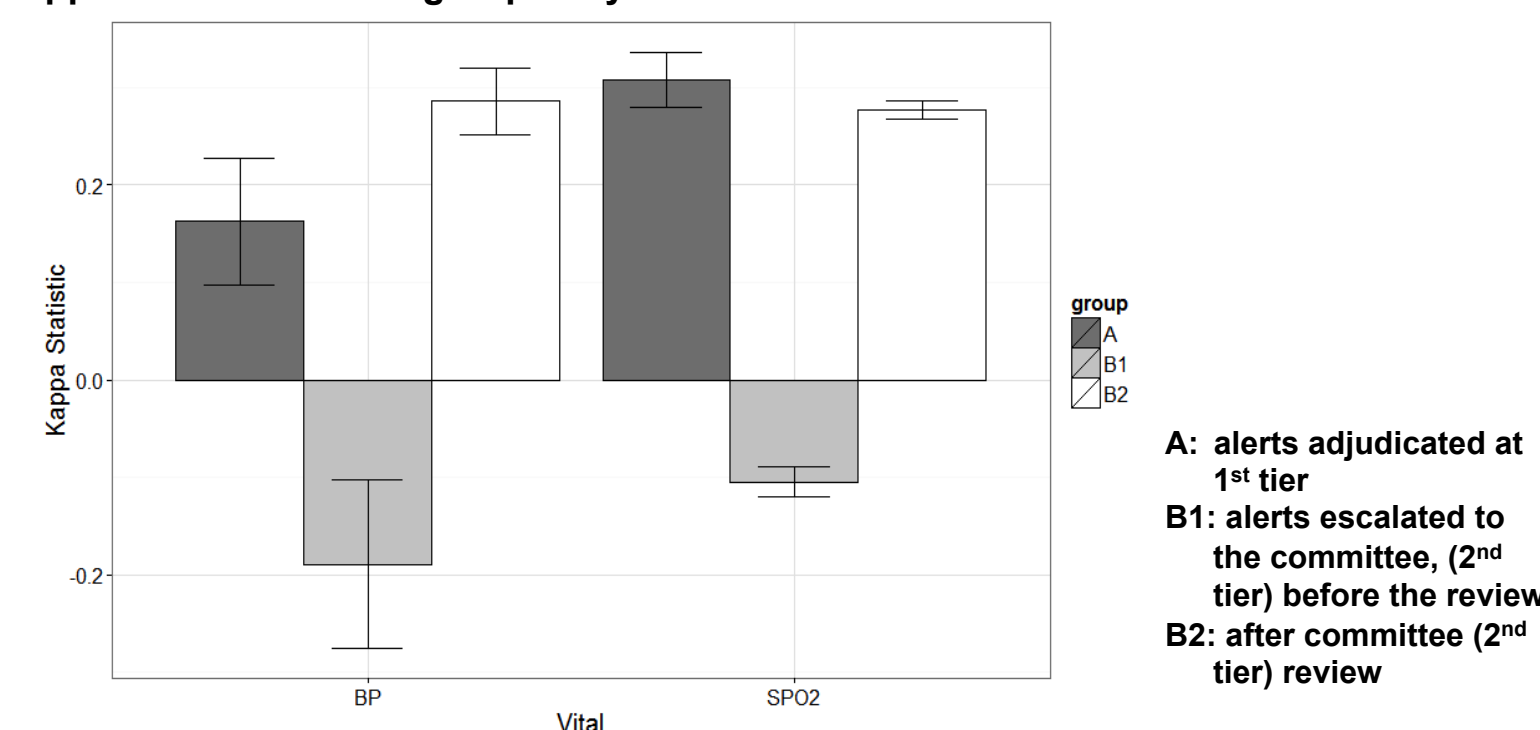


RESULTS

The proportion of alerts **escalated to the 2nd tier review** was 26 (32.5%) BP, 115 (50%) SpO₂. Almost all of HR and RR alerts could be adjudicated in the first tier.

The results show that the **consensus for alerts initially conflicted improved significantly** as a result of the 2nd tier committee review. Weighted pairwise Kappa statistic increases from -0.19 to 0.29 for BP, and from -0.10 to 0.28 for SpO₂ alerts.

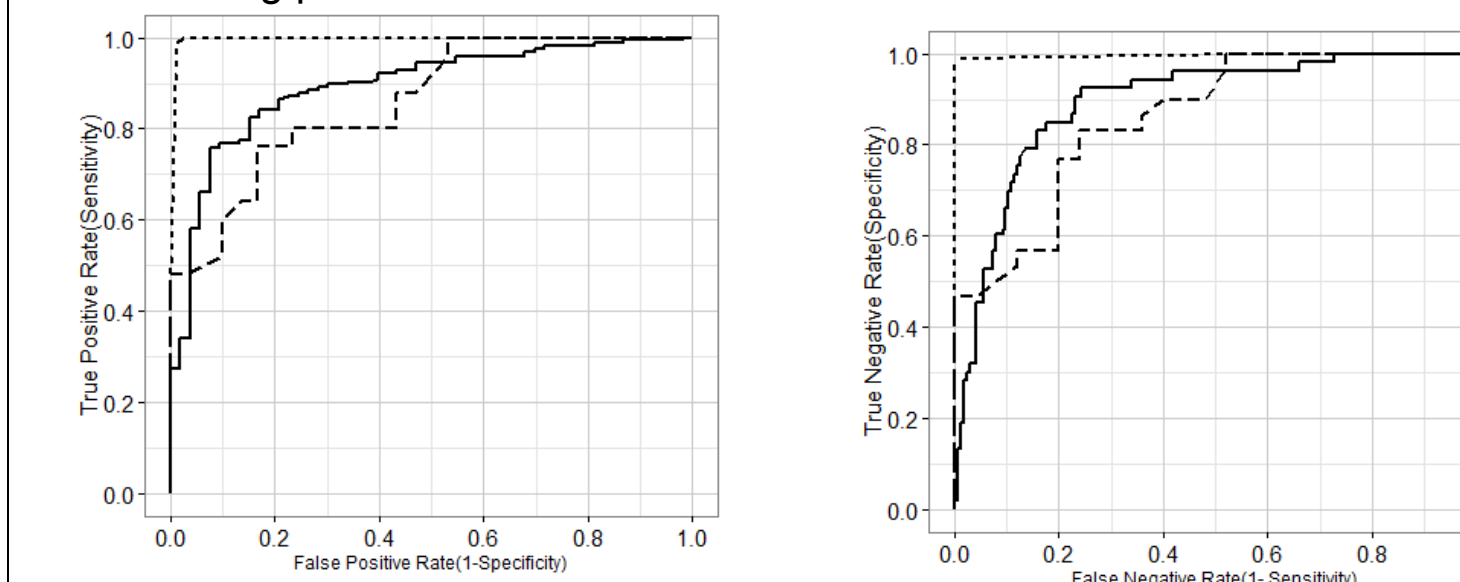
Kappa scores for alerts grouped by tier



ROC Curves for Artifact Adjudication

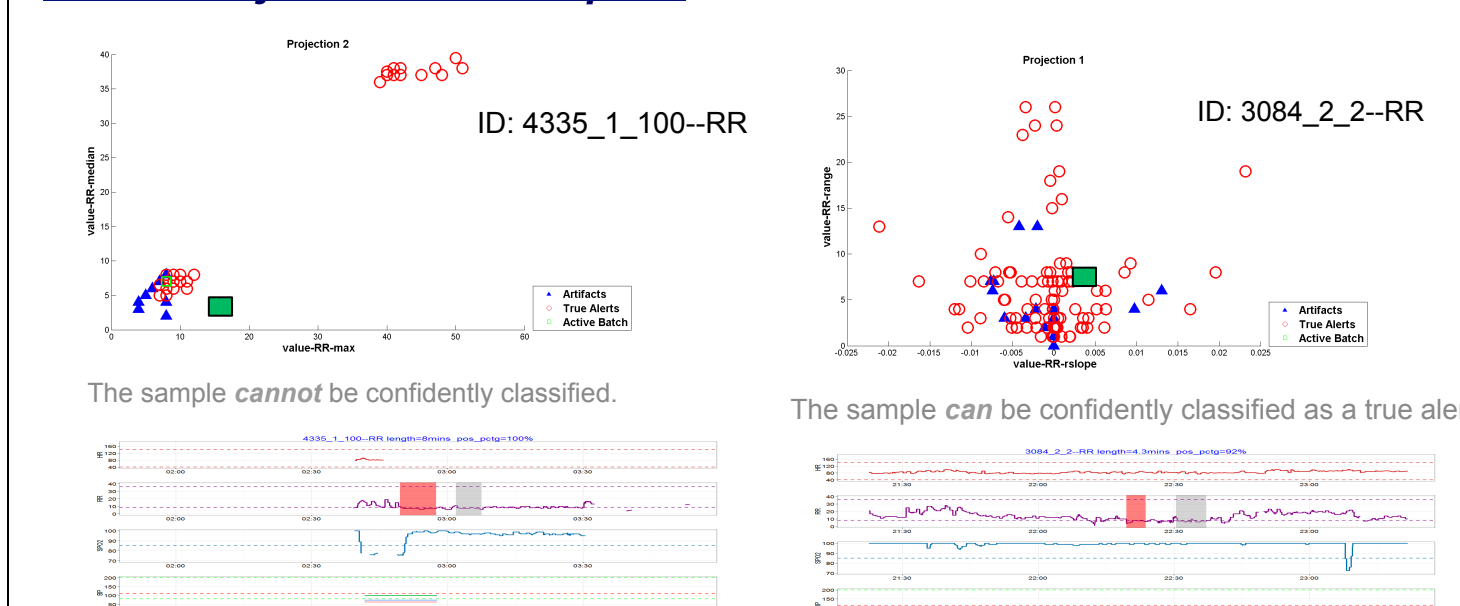
Preliminary **artifact adjudication model** was built from these annotations:

- Very strong ability to identify RR alerts and artifacts** (dense dashed line in the ROC diagrams).
- Ability to very confidently isolate** more than 47% of **true BP alerts** and more than 45% of **BP artifacts**
- Very good performance in isolating SpO₂ artifact**, equivalent to **what can be attained with 50% more annotated training data** if the Active Machine Learning protocol has not been used.



The classifier was not able to adjudicate 114 cases at the end of the 1st cycle, but at the end of 2nd it could confidently process all unlabeled data saving experts from having to label 304 episodes at that point (52% effort reduction), while RF method would allow 11% effort reduction at the end of cycle 3.

Artifact Adjudication Examples:



CONCLUSIONS

We implemented a multi-tier framework to elicit ground truth from multiple reviewers to support development of a prototype of the automated artifact adjudication system.

The initial results show that precious human expertise can be utilized efficiently and without loss of performance of the resulting models of instability.

The proposed annotation framework can yield accurate alert adjudication systems while minimizing effort of human experts required to produce ground truth evidence, even if very large libraries of reference data are available.