33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Scalable Customization of Atrial Fibrillation Detection in Cardiac Monitoring Devices: Increasing Detection Accuracy through Personalized Monitoring in Large Patient Populations

Kuk Jin Jang, Guha Balakrishnan, Zeeshan Syed, Naveen Verma

Abstract— **To make it viable for remote monitoring to scale to large patient populations, the accuracy of detectors used to identify patient states of interests must improve. Patient-specific detectors hold the promise of higher accuracy than generic detectors, but the need to train these detectors individually for each patient using expert labeled data limits their scalability. We explore a solution to this challenge in the context of atrial fibrillation (AF) detection. Using patient recordings from the MIT-BIH AF database, we demonstrate the importance of patient specificity and present a scalable method of constructing a personalized detector based on active learning. Using a generic detector having a sensitivity of 76% and a specificity of 57% as its seed, our active learning approach constructs a detector with a sensitivity of 90% and specificity of 85%. This performance approaches that of a patient-specific detector, which has a sensitivity of 94% and specificity of 85%. By selectively choosing examples for training, the active learning approach reduces the amount of expert labeling needed by almost eight fold (compared to the patient-specific detector) while achieving accuracy within 99%.**

I. INTRODUCTION

In recent years, the growing demand for continuous care services has placed an increased focus on making healthcare scalable and cost-effective [1]. Remote monitoring systems are of particular interest, allowing healthcare professionals to network and share resources at a distance to efficiently administer care. Preliminary realizations of these systems [2] have begun to demonstrate the viability with which electronic devices and networking technologies can facilitate such methods. However, for these systems to have substantial impact, they must not only address issues of technological scalability, but also issues limiting scalability of clinical responsiveness. This implies the need for robust performance to minimize alarm fatigue, which today is a bottleneck even in hospital monitoring; the ratio of false alarms in the ICU to true critical alarms, for instance, can be as high as 100 to 1 with the most current detectors [3]. These inaccuracies compromise patient care (due to unheeded or turned-off alarms [4]), and the problem will be exacerbated in the out-patient scenario where the intent is for monitoring to scale to much larger populations.

Manuscript received April 14, 2011. This work was supported in part by the Gigascale Systems Research Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity.

K. Jang and N. Verma are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

G. Balakrishnan and Z. Syed are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

Recently, data-driven approaches for monitoring patients have demonstrated the ability to improve the accuracy of clinical detection systems. These approaches exploit the increasing large-scale availability of patient data to construct high-order models of physiological signals, thereby achieving high detection sensitivity and specificity. Through the use of machine-learning techniques, the ability to construct and apply such models has become increasingly efficient even when the datasets are large and the models must capture complex correlations. It is thus possible to construct models on even a patient-by-patient basis, which can substantially improve the specificity of detection tasks [5]. The critical challenge, however, is that with existing machine-learning approaches, this requires large volumes of patient-specific training data that must first be labeled by an expert. This motivates a need to make the customization process more scalable.

In this paper, we demonstrate a scalable approach for constructing a patient-specific atrial fibrillation (AF) detector. Our approach is based on active learning, a machine learning technique aimed at reducing the costs of training supervised learners. Active learning has been explored in a wide range of applications and has been applied in earlier medical work, most notably to epileptic seizure detection [6]. Our focus on AF is motivated by multiple considerations. First, AF is the most common form of atrial sustained arrhythmia and accounts for more hospitalizations than any other cardiac arrhythmia [7]. Second, while AF may not be lethal itself, it is associated with other cardiac conditions and increases the risk of death from cardiac disease [8]. Third, continuous, accurate, and long-term detection of AF is needed in order to correctly diagnose and treat AF [9]. These factors make it one of the key states of interest for remote-monitoring applications targeting pre-emptive response.

In this paper we:

- Develop an active learning framework for patient-specific AF detection.
- Evaluate the benefits of patient specificity and our active learning algorithm on real electrocardiogram (ECG) data containing atrial fibrillations.

II. METHODS

A. AF Detection Architecture

We used an AF detection architecture based on a two-step process where features are first extracted from electrocardiographic (ECG) data and then classified using a support

vector machine (SVM). The feature extraction process is based on the approach described in [10] (see Fig. 1). Each patient's ECG data is first segmented into 2 minute intervals. The R-peaks of the ECG waveform are then identified, and the R-peak to R-peak (RR) intervals over this period are calculated. Subsequently, the first difference of the RR time-series is computed corresponding to the new time-series δ_{RR} where $\delta_{RR}(i) = RR(i) - RR(i-1)$. Characteristics of AF are expressed by the statistics exhibited by successive δ_{RR} values. In particular, a 2-D histogram is derived using successive values to form a coordinate pair, i.e., $(\delta_{RR}(i), \delta_{RR}(i-1))$, and the histogram bins are divided into the groups arranged as shown in Fig.1. Based on the counts in groups of bins, six features are calculated (as detailed in [10]) and concatenated to form a feature vector.

Each training feature vector is assigned a label of -1 or 1 based on expert annotations of when AF occurs. These vectors are then used to train an SVM classifier using a Gaussian radial basis function (RBF) kernel. The RBF kernel is used for high flexibility in modeling the data distributions in the feature space. The model can then be used for real-time detection of AF based on test feature vectors.

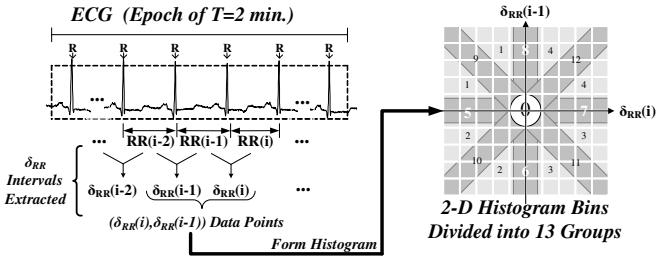

Fig. 1. Feature extraction using the process described in [10]

Using this architecture, we develop three types of AF detectors differing in the manner in which they acquire training data: a generic detector (D_G) is trained on historical records (Fig. 2a); an ideal patient-specific detector (D_{PS}) is trained on a given patient's own data (Fig. 2b); and an active learning detector (D_A) is trained using an informative subset of the given patient's data. Intuitively, D_G incurs the lowest possible annotation cost since no new labeling is required, but it may also be fairly inaccurate since it is not capable of acutely modeling each patient's unique physiology. Conversely, we expect D_{PS} to exhibit the best sensitivity and specificity, but also to require extensive expert labeling. D_A, as described in the following section, attempts to achieve D_{PS}'s performance with far less dependence on human labeling.

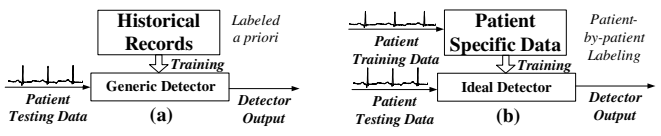

Fig. 2. Block diagram of (a) generic detector (D_G) and (b) ideal, patient-specific detector (D_{PS}).

B. Active Learning Detector (D_A)

Active learning involves the closed-loop phenomenon of a learner selecting actions or making queries that influence the choice of data to be added to its training set. An active learner attempts to select data points that are the most informative to train on, and these points are then labeled by an 'oracle' (e.g., a human expert) with some cost associated with each query. These labeled instances are added to the training set of the classifier. This cycle repeats until a stopping criterion is met (see Fig. 3). The promise of active learning is that when the instances are selected properly, the data and computation costs can be reduced dramatically [11],[12].

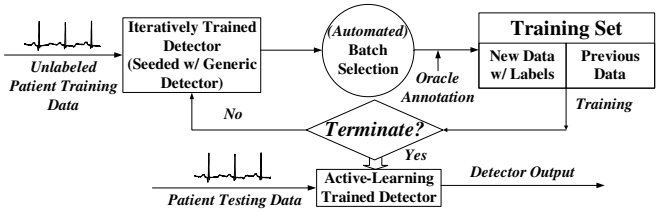

Fig. 3. Diagram of active learning process

The active learning process can be separated into three distinct steps: *initialization*, *selection*, and *termination*. During initialization, the active learner is seeded with an initial detector. In this work, the initial detector is the generic detector D_G. This initial detector is then subsequently modified through the selection phase as unlabeled data instances (i.e., feature vectors) acquired from the patient are labeled and added to the training set. To reduce the number of training computation cycles, we adopted a process in which data instances are added in batches of 100. While several heuristics have been proposed in the literature to select these batches intelligently, we use a simple approach that selects instances nearest to the SVM decision hyperplane [13]. More formally, for a given data instance $x_i \notin I_\tau$, where I_τ corresponds to the labeled training set prior to the selection of the τ-th batch, new data instances are chosen to minimize:

$$\left| \sum_{x_\tau \in I_\tau} y_\tau \alpha_\tau K(x_\tau, x_i) \right| \tag{1}$$

where $K(x_\tau, x_i)$ corresponds to the RBF SVM kernel, and α_τ denotes the Lagrangian multiplier for x_i.

In general, as training instances are added through selection, more support vectors are produced (i.e., the SVM model complexity increases). Fig. 4 shows a typical scenario of how the number of support vectors evolves in this application during the selection process; a distinct 'knee' occurs in the curve, after which the rate of support vectors produced decreases. We consider halting the process at this knee since learning then begins to noticeably stagnate. In order to approximate this point, we developed a novel termination criterion based on the rate of change of the number of support vectors. Formally, we terminate the learner at batch τ when:

$$\frac{(\theta(\tau-10) - \theta(\tau))/10}{(\theta(\tau-20) - \theta(\tau))/20} > \beta \tag{2}$$

where $\theta(\tau)$ is the number of support vectors in the trained model at batch τ and β is a threshold. A value of $\beta = 2$ is used. Intuitively, this criterion terminates the learner when the slope of the θ curve starts rapidly decreasing.

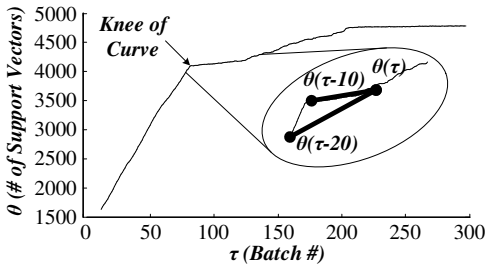

Fig. 4. Example of active learning process (*Results of Patient 20*). A batch corresponds to 100 selected feature vectors.

III. Testing Methodology and Results

A. Experimental Setup

We used the MIT-BIH AF database [14] to evaluate our work. This dataset consists of ECG data (sampled at 250 Hz with 12-bit resolution) for 21 patients (10 hours of data per patient). When training D_G, we used 2-fold cross-validation for training and evaluation. This involved randomly dividing the patients in the MIT-BIH AF database into two sets and alternately training D_G on records from one set and then testing the detector on records from the other set. We set the SVM parameter $\gamma = 20$ and used equal class-specific penalties. When evaluating D_{PS} and D_A, 2-fold cross-validation was performed on each patient's own data record. For D_G and D_{PS}, optimal SVM parameters were chosen using 5-fold cross-validation on the training set. This option was not feasible for D_A since no labels are available at the onset of the active learning process; we thus used the same SVM parameters as those for D_G. Finally, we also developed a detector (D_R) that was initialized with the generic model and then trained by selecting patient-specific data batches randomly. This was used as a control in our experiments to evaluate the performance of D_A. We used the LIBSVM [15] software package for our work.

In order to evaluate the performance of the detectors, the following metrics were used: sensitivity (*Sn*), specificity (*Sp*), and overall accuracy (*OA*). Sensitivity was defined as $Sn = TP/(TP+FN)$, specificity was defined as $Sp = TN/(TN+FP)$, and overall accuracy was defined as $OA = (TP+TN)/Total$ (where TP is the true positive count, FP is the false positive count, TN is the true negative count, FN is the false negative count, and $Total$ is the sum of these).

B. Results

Table I presents the performances of the three detectors and Fig. 5 compares their overall accuracies. D_G achieved a mean performance of 64.7%, 76.0%, and 57.4% for *OA*, *Sn*, and *Sp*. As expected, D_{PS} achieved much better performance, with mean values of 92.0%, 94.4%, and 84.6%. For some patients (e.g., patients 5 and 6), D_{PS} exhibited a slight decrease in sensitivity, though this reduction was minimal compared to the improvement in specificity, amounting to considerable improvement in *OA*. D_A had a mean performance of 92.2%, 90.3%, and 84.9%, which was comparable to D_{PS}.

TABLE I
Performance Summary of AF Detectors

Patient	Generic (D_G)			Patient-specific (D_{PS})			Active Learning (D_A)		
	OA(%)[a]	*Sn*(%)	*Sp*(%)	*OA*(%)	*Sn*(%)	*Sp*(%)	*OA*(%)	*Sn*(%)	*Sp*(%)
1	87.51	94.25	75.26	99.86	99.79	99.99	99.85	99.78	99.99
2	61.82	100.0	61.39	99.99	100.0	99.99	99.90	100.0	99.90
3	85.90	100.0	85.72	99.97	98.00	99.99	99.96	97.60	99.99
4	47.14	94.41	43.53	99.90	99.41	99.94	99.94	99.87	99.94
5	39.48	93.33	16.82	80.83	78.66	81.74	82.24	40.44	99.82
6	68.38	99.86	0.70	82.30	93.63	60.13	83.04	93.14	61.34
7	58.39	75.85	17.33	90.59	95.33	79.43	90.32	95.44	78.28
8	71.79	90.55	29.51	93.60	99.85	79.50	93.73	99.72	80.23
9	73.85	69.07	76.91	84.83	83.20	85.81	88.43	71.00	99.62
10	70.85	74.39	69.82	83.88	90.60	81.93	87.98	63.00	95.23
11	91.57	90.21	95.17	99.15	99.34	98.66	99.49	99.63	99.10
12	33.50	06.85	95.98	84.49	82.83	88.38	84.04	88.97	72.49
13	86.35	83.77	94.04	97.91	97.70	98.52	99.02	99.05	98.93
14	36.90	07.71	87.49	87.48	97.65	69.85	87.17	97.03	70.09
15	88.21	97.19	16.46	93.19	99.41	43.48	93.08	99.37	42.83
16	31.95	11.68	86.94	88.41	90.54	82.65	86.60	95.19	63.28
17	59.33	76.95	23.50	79.63	93.19	52.08	79.55	92.57	53.09
18	89.62	91.87	81.63	97.42	97.03	98.79	97.55	96.95	99.68
19	79.82	97.20	73.74	98.91	99.58	98.68	99.34	99.67	99.22
20	48.42	50.28	38.97	95.92	98.20	84.30	94.26	98.49	72.69
21	47.59	91.37	34.86	92.11	87.92	93.32	90.84	69.99	96.91
Mean	64.68	76.04	57.42	91.96	94.38	84.63	92.21	90.33	84.88
Std.	20.18	29.65	60.70	7.08	6.44	16.53	6.81	15.79	18.45

[a] *OA*: Overall Accuracy, S_n: Sensitivity, S_p: Specificity

Fig. 6 shows the progression in accuracy over 21 patients as data batches are added for training (the curves are normalized to the final accuracy achieved by using the entire patient training dataset). As shown, convergence is achieved with very few batches; the termination criterion is met after adding approximately 20% of the patient-specific data. In fact, the detector achieves 99% of its final *OA* value with only 13% of the data, suggesting that a more aggressive termination criterion may be chosen. Table II shows the effectiveness of the hyperplane-distance selection criterion by comparing the amount of data required by both D_A and D_R to achieve 99% of D_{PS} accuracy. On average, D_A required less than half the data compared to D_R. Fig. 7 shows how the accuracy of D_A and D_R progresses for two representative cases (Patients 2 and 14), illustrating the rapid convergence achieved by D_A.

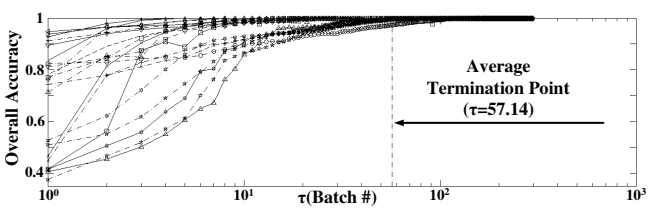

Fig. 6. Progression in accuracy (over 21 patients) for active-learning trained detector. On average, the termination criterion is met with 20% of the data instances.

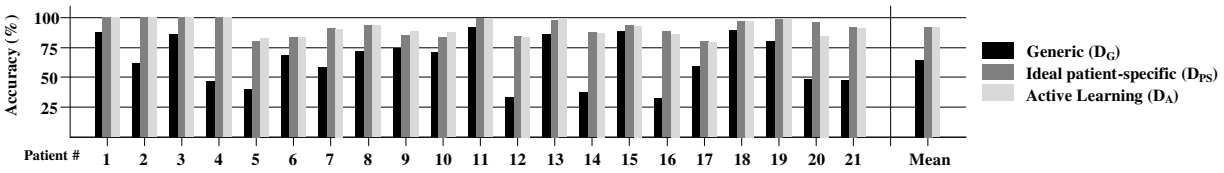

Fig. 5. Overall accuracy of generic, ideal patient-specific, and active-learning trained detectors

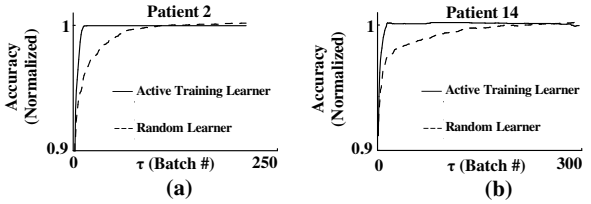

Fig. 7. Normalized progression in accuracy of D_A compared to D_R for (a) Patient 2 and (b) Patient 14.

TABLE II

NUMBER OF DATA POINTS FOR 99% OF D_{PS} ACCURACY

Patient	Active (D_A)		Random (D_R)		Patient	Active (D_A)		Random (D_R)	
	Amount	(%)[a]	Amount	(%)		Amount	(%)	Amount	(%)
1	200	1	700	3	12	4900	27	10400	57
2	900	4	4900	22	13	1300	5	6300	25
3	200	1	400	2	14	3400	15	15200	67
4	700	3	3100	14	15	800	3	7800	28
5	9200	30	5700	18	16	4500	26	13100	75
6	9000	30	21800	72	17	5200	26	12700	65
7	4700	26	10200	56	18	700	2	7800	28
8	700	3	3100	14	19	500	2	2000	7
9	2100	9	1400	6	20	3800	14	13400	49
10	3500	18	4700	24	21	4900	16	16600	56
11	400	1	1900	7	Mean	2933.3	13	7771.4	33

[a] Percent of data points used out of total training instances for each patient.

IV. DISCUSSION & CONCLUSION

Our results show that patient specificity improves AF-detection, a finding consistent with outcomes from studies in other clinical domains. In particular, our generic detector exhibited poor specificity. This characteristic is common to many detection systems in use today that are trained on a population-level and thus suffer from an inability to precisely differentiate between true events and artifacts due to their need to detect events across a broad range of individuals. Despite the improved performance provided by patient specificity, however, the excessive annotation costs associated with training a detector for each individual has led to generic systems still being widely employed.

To address this issue, we developed a novel architecture based on active learning that may make patient-specific AF detection more scalable. Our SVM-based detector makes use of population-level prior knowledge for an initial model, and refines this knowledge by selectively interacting with human experts to query examples from a new patient until a termination criterion is met. Our novel criterion assesses the rate at which support vectors are being added to the detector's model and terminates the process when this rate begins to stagnate. Through this approach, our detector achieved very similar accuracy to a patient-specific detector while requiring 80% fewer examples.

Our results are promising, but it is also useful to test our system on other annotated datasets as well as real clinical settings. Such studies can also help refine the initialization, batch selection, and termination heuristics used in this work.

REFERENCES

[1] S. Keehan, A. Sisko, C. Truffer, S. Smith, C. Cowan, J. Poisal, M. K. Clemens, and the National Health Expenditure Accounts Projections Team, "Health spending projections through 2017: The baby-boom generation is coming to medicare," *Health Affairs*, vol. 27, no. 2, pp. 145–155, 2008.

[2] A. D. Jurik and A. C. Weaver, "Remote medical monitoring," *Computer*, vol. 41, pp. 96–99, 2008.

[3] T.-Y. Leong, D. Aronsky, and M. M. Shabot, "Guest editorial: Computer-based decision support for critical and emergency care," *J. of Biomedical Informatics*, vol. 41, pp. 409–412, June 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1379917.1380259

[4] L. Wallis, "Alarm fatigue linked to patient's death," *AJN The American Journal of Nursing*, vol. 110, no. 7, p. 16, 2010.

[5] A. Shoeb and J. Guttag, "Application of machine learning to epileptic seizure detection," in *Proc. Int. Conf. on Machine Learning*, Nagoya, Japan, June 2010.

[6] G. Balakrishnan and Z. Syed, "Scalable personalized medicine with active learning: detecting seizures with minimum labeled data," in *Proceedings of the 1st ACM International Health Informatics Symposium*, ser. IHI '10. ACM, 2010, pp. 83–90.

[7] A. S. Go, E. M. Hylek, K. A. Phillips, Y. Chang, L. E. Henault, J. V. Selby, and D. E. Singer, "Prevalence of diagnosed atrial fibrillation in adults," *JAMA: The Journal of the American Medical Association*, vol. 285, no. 18, pp. 2370–2375, 2001.

[8] E. J.Benjamin, P. A. Wolf, R. B. D'Agostino, H. Silbershatz, W. B. Kannel, and D. Levy, "Impact of atrial fibrillation on the risk of death : The framingham heart study," *Circulation*, vol. 98, no. 10, pp. 946–952, 1998.

[9] C. W. Israel, G. Gronefeld, J. R. Ehrlich, Y.-G. Li, and S. H. Hohnloser, "Long-term risk of recurrent atrial fibrillation as documented by an implantable monitoring device: Implications for optimal patient care," *J Am Coll Cardiol*, vol. 43, no. 1, pp. 47–52, 2004.

[10] S. Sarkar, D. Ritscher, and R. Mehra, "A detector for a chronic implantable atrial tachyarrhythmia monitor," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1219 –1224, Mar. 2008.

[11] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, pp. 319–342, 1988, 10.1023/A:1022821128753. [Online]. Available: http://dx.doi.org/10.1023/A:1022821128753

[12] E. Baum, "Neural net algorithms that learn in polynomial time from examples and queries," *IEEE Trans. Neural Networks*, vol. 2, no. 1, pp. 5 –19, Jan. 1991.

[13] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. Morgan Kaufmann Publishers Inc., 2000, pp. 839–846.

[14] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, June 2000.

[15] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.