Deep Learning Based Classification of True/False Arrhythmia Alarms in the Intensive Care Unit

Event 2: Retroactive Classification Performance

Jack Boynton[1] **and Byung Suk Lee**[2]

[1]Department of Electrical and Biomedical Engineering
[2]Department of Computer Science; University of Vermont
[1]jwboynto@uvm.edu, [2]bslee@uvm.edu

THE UNIVERSITY OF VERMONT
COLLEGE OF ENGINEERING & MATHEMATICAL SCIENCES

Abstract

Once a cardiac alarm is triggered in the intensive care unit (ICU), accurately classifying whether the alarm is true of false is of critical importance. Incorrect classification may lead to patient's death if the alarm is true or to disruption in patient care if false. There has been a body of research, as signified by the 2015 PhysioNet/CinC Challenge; due accomplishments have been made in the relevant computational technology, and yet the highest accuracy known thus far is in the mid-80% range (85%). Our work achieved much higher accuracy and, additionally, very early classification almost at the onset of an arrhythmia alarm, by utilizing state of the art deep learning methods. The machine learning model used is a Residual Network (ResNet) and a Bi-directional Long Short Term Memory (BiLSTM) connected in tandem. Using the PhysioNet dataset of 750 recorded ECG segments published with the Challenge, our method performed the classification with 96% accuracy in 0.52 seconds from the onset of an alarm on average over all test ECG segments.

Introduction

The research presented in this paper stems from the 2015 PhysioNet/CinC Challenge [1], particularly the "retroactive" classification test to determine within 10 seconds whether an arrhythmia alarm is true or false in the ICU. The accuracy of classification is undoubtedly important; misclassifying a true alarm as false may result in a patient's death and misclassifying a false alarm as true may result in wasteful disruption and disturbance. Thus, the goal of the Challenge was to achieve as high true/false alarm classification accuracy as possible. The result of the Challenge [2] was limited to 85% accuracy for the top method, and there has been no further advancement since then. Our work adds novel advancements to this state of the art.

There are two advancements. First, our work enhanced the accuracy to a high 90% range (96%) by using a *deep learning* model as the computational method. None of the published papers on this problem (e.g., reported by Clifford et al. [2]) used deep learning, which has proven to produce powerful models given adequate training data. Second, our work addresses *early* classification of true or false alarms, which is attributed to the excellent feature-extraction ability of a deep learning model. Early classification is as important as accurate classification. A delayed classification of true alarm is potentially dangerous to the patient, and a delayed classification of false alarm deprives the opportunity to suppress it in time. Thus, this paper presents the computational methods and the results of using a deep learning model to determine whether an alarm is true or false accurately and early in the ICU.

Methods

Deep learning model A model comprising ResNet and BiLSTM in tandem.ResNet is used to extract complex features from the ECG time series, and BiLSTM is used to build a prediction model based on the temporal order of features. The model output is the probability of an alarm being true. The ResNet architecture we used consists of three residual convolutional blocks with filter sizes 64, 128, and 128, respectively, effectively extracting 128 features from a given ECG segment. The BiLSTM is a standard TensorFlow model 'Bidirectional.'

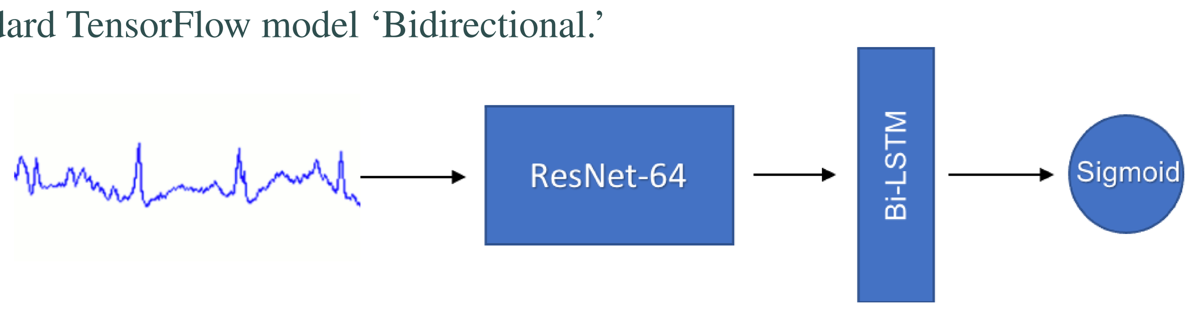


Figure 1: Deep learning model structure.

Prequential evaluation Time series data are arriving as chronologically ordered samples. Therefore, we used prequential evaluation (as opposed to the conventional cross-validation) to consider the effect of such temporal ordering. Basically, each new batch of data is first used as test data and then appended to the existing training data. Thus, the training data size keeps increasing, and so does the training time as the total training size increases (see Figure 2).

Data sets The training dataset provided in the 2015 PhysioNet/CinC Challenge was split to training and test data sets at the ratio of 80% to 20% for the purpose of our evaluation, as the test data set used in the Challenge was not available.

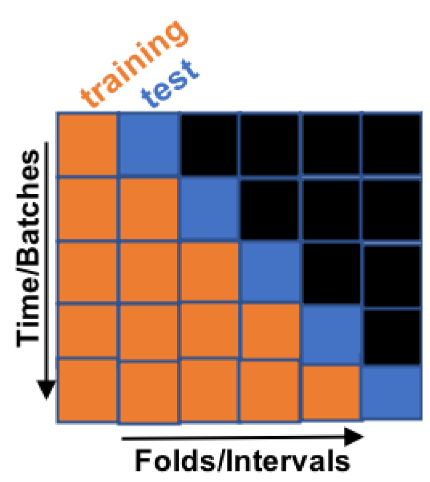


Figure 2: Prequential evaluation

Results and Discussion

True/false alarm classification performance

The performance has two factors: classification accuracy (i.e., how accurate the classification output is) and classification time (i.e., how quickly the classification can be made). The classification accuracy was measured as in Equation 1 following the 2015 PhysioNet/CinC Challenge, and the classification time was measured as wall-clock time (or, equivalently, the number of ECG samples).

$$\text{Accuracy score} = \frac{TP + TN}{TP + TN + FP + 5 \times FN} \quad (1)$$

We first tried a threshold-based approach, that is, wait until the probability reaches a certain threshold value before outputting the classification result. Specifically, the model training and testing were done at different batch intervals of ECG samples, progressively decreased from 10 seconds down to 2, 1, 0.5 seconds and then further down to one sample interval (4 milliseconds). The result showed an increase in the classification time without any gain in the classification accuracy when the interval was reduced (see Figure 4(left)).

Careful inspection of this phenomenon revealed that, for all test ECG segments, the probability changed rapidly and monotonously, either positive (true alarm case) or negative (false alarm case)—see Figure 3. Based on this observation, the notion of threshold was dismissed, and the earliest possible sample point of detecting the polarity of the probability change was examined. We would call this a "polarity" approach. The resulting classification time was about 125 samples, amounting to 0.52 seconds, on average, which was far shorter than 1.88 seconds on average for the threshold-based approach (see Figure 4(right)). Notably, the classification time in the polarity approach was also consistent across different test ECG segments; moreover, earlier classification did not compromise the accuracy at all, and actually raised it a bit to 96.23% compared with 95.00% for the threshold approach.

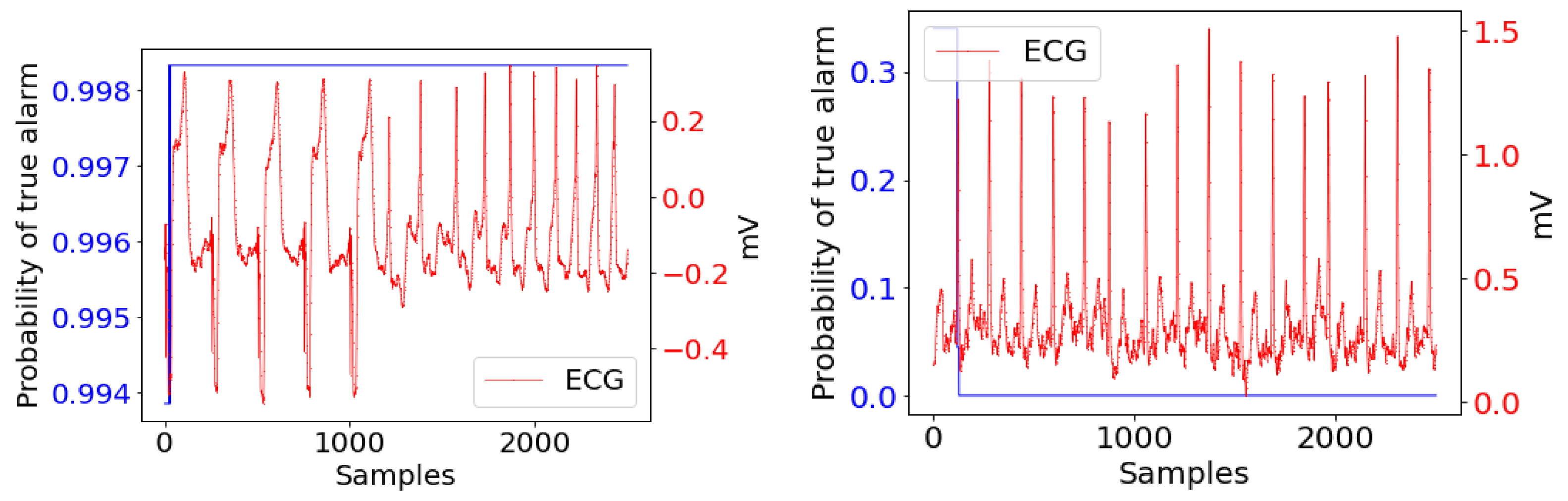
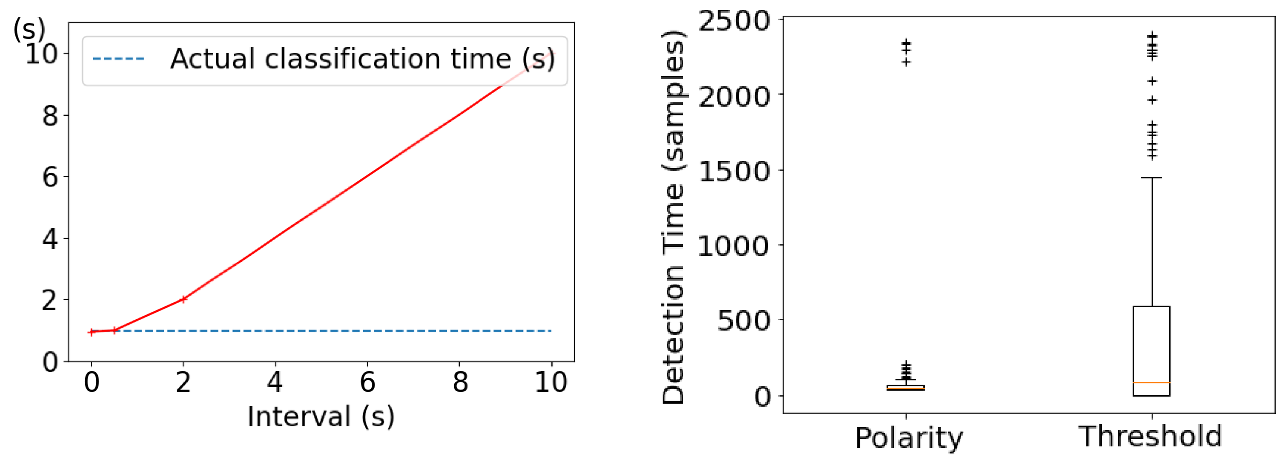


Figure 3: Change of model's output probability over time: (left) Increase (true alarm); (right) Decrease (false alarm).



(Interval = sample interval (4 ms))

Figure 4: Mean classification time for varying interval in the threshold approach (left); classification times of the two approaches (right).

The remarkably early classification with such a high accuracy drew our suspicion at first and triggered a thorough investigation. The prequential evaluation was part of the investigation effort, to prevent overfit of the model for time series. Visual inspection of a significant number of the ECG segments used suggested that the key to such early classification is the substantial regularity inherent in the ECG waveform morphology. This regularity enables the deep learning model to quickly capture the signature features of samples that predict the polarity of the output probability (i.e., increase or decrease). In our work, it was typically after "seeing" the first wavelet (e.g., P-, QRS, T-wave) in an ECG beat. A further, larger-scale investigation is warranted involving more diverse ECG segments reflecting different patient cohorts (e.g., gender, age, body mass index).

Comparison with prior art

Table 1 summarizes the top four methods that achieved higher than 80% accuracy (according to Equation 1) in the retroactive classification test among the contestants [2]; the 'voting algorithm' was not a contestant but added by the Challenge organizer to always pick the best result via majority vote of the select top 13 methods. Note that none of them used deep learning. Note as well that the aspect of "early classification" is unique to our work, so is not applicable to their work.

Publication	Method	Accuracy
This paper	ResNet+BiLSTM	96.23%
Voting algorithm (top 13) [2]	Majority vote	87.04%
Fallet et al. [3]	Rule-based	85.04%
Plesinger et al. [4]	Rule-based	84.96%
Kalidas et al. [5]	Support vector machine	81.85%

Table 1: Comparison with other work.

Conclusions

This paper presented a novel work that achieved accurate and early classification of true or false arrhythmia alarm in the ICU, surpassing the state of the art. The enabling computational method was deep learning. The deep learning model, the prequential evaluation, and the experiments were presented.

Our immediate further work is to adopt the method into the "real-time test" case of the PhysioNet/CinC 2015 Challenge [2], to predict a true alarm early and accurately before an arrhythmia alarm is triggered. The method can be also applied more broadly to other time-critical arrhythmia monitoring settings as well, like remote cardiac care via an implanted monitor.

References

[1] Clifford G, Silva I, Moody B, Mark R. Reducing false arrhythmia alarms in the ICU - the PhysioNet computing in cardiology challenge 2015. https://www.physionet.org/content/challenge-2015/1.0.0/, February 2015.

[2] Clifford G, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D, Mark R. The PhysioNet/Computing in Cardiology challenge 2015: Reducing false arrhythmia alarms in the ICU. In Proc. CinC. September 2015; 273–276.

[3] Fallet S, Yazdani S, Vesin J. A multimodal approach to reduce false arrhythmia alarms in the intensive care unit. In Proc. CinC. September 2015; 277–280.

[4] Plesinger F, Klimes P, Halamek J, Jurak P. False alarms in intensive care unit monitors: Detection of life-threatening arrhythmias using elementary algebra, descriptive statistics and fuzzy logic. In Pro. CinC. September 2015; 281–284.

[5] Kalidas V, Tamil LS. Enhancing accuracy of arrhythmia classification by combining logical and machine learning techniques. In Proc. CinC. September 2015; 733–736.

Acknowledgements

This research was funded by the University of Vermont Summer Undergraduate Research Fellowship (SURF) and College of Engineering and Mathematical Sciences Research Experience for Undergraduates (REU) award.