Computers in Biology and Medicine 109 (2019) 14–21

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ECG anomaly class identification using LSTM and error profile modeling

Sucheta Chauhan[a,**], Lovekesh Vig[b], Shandar Ahmad[a,*]

[a] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India
[b] Tata Consultancy Services - Research and Innovation, New Delhi, India

ARTICLE INFO

Keywords:
ECG signal
Deep learning
Long short term memory (LSTM)
Multi layer perceptron
Logistic regression

ABSTRACT

Automatic diagnosis of cardiac events is a current problem of interest in which deep learning has shown promising success. We have earlier reported the use of Long Short Term Memory (LSTM) networks-trained on *normal* ECG patterns-to the detection of anomalies from the prediction errors for real-time diagnostic applications. In this work, we extend our anomaly detection algorithm by introducing a second stage predictor that can identify the *actual anomaly class* from the error outputs of the first stage model. Results from seven types of anomalies have been presented including Atrial Premature Contraction (APC), Paced Beat (PB), Premature Ventricular Contraction (PVC), Right Bundle Branch Block (RBBB), Ventricular Bigeminy (VB), Ventricular Couplets (VCs) and Ventricular Tachycardia (VT). To optimize anomaly class prediction performance, multiple choices of second stage models such as multilayer perceptron (MLP), support vector machine (SVM) and logistic regression have been employed. A featurization scheme for LSTM prediction errors in the form of overall summaries has been proposed and a successful predictor for the same was developed with good performance. Our results indicate that the error vectors represented by their summary features carry useful predictive information about actual ECG anomaly type. We discuss how the accuracy scores without attention to inherent class imbalances and paucity of data instances may produce misleading performance estimates and hence accurate background models are needed to estimate true predictive performance of multi-class predictors such as those presented in this work. The training data sets and related resources for this study are provided at http://ecg.sciwhylab.org.

1. Introduction

Heart disease is one of the major health problems worldwide. Early detection of heart diseases and proper medical treatment can prevent sudden deaths [17]. A large number of cardiac and seriously ill patients are under constant monitoring of their cardiac conditions with the help of ECG or EKG (Electrocardiogram) recording devices as they are at a risk of undergoing cardiac events needing critical care. Intensive real-time monitoring of ECG in the form of human attention is not practical in most such cases and hence there is a need to automatically detect the anomalous cardiac events directly from machine-readable, recorded ECG signals in an exact or approximate real time scenario. Successful real-time diagnosis or classification of ECG signals is achieved by finding characteristic shapes of the ECG that discriminate effectively between the previously defined diagnostic categories (See Fig. 1 for typical ECG patterns of normal and various types of abnormal heart beats). An arrhythmia is an abnormal heart rhythm. Various machine learning and data mining methods have already been developed to improve the detection and classification of different types of cardiac arrhythmias based on ECG signal data but the levels of predictive performances are still low [16]. Reported approaches include Self-Organizing Maps (SOM), Support Vector Machines (SVM) [19] Multilayer Perceptron (MLP) [4]; Markov Models and Fuzzy or Neuro-fuzzy Systems [2,6,10]. Historically, Khadra et al. way back in 1997 used time-frequency wavelet theory to differentiate among three types of arrhythmias i.e. ventricular fibrillation, atrial fibrillation, and ventricular tachycardia [12]. Patra et al. [18] proposed an integration methods using fuzzy c-means (FCM) clustering, Principal Component Analysis (PCA) and Neural networks for classifying five types of ECG beats [18]. Chazal et al. [9] used the wavelet coefficients to describe the ECG shape, wherein classification is performed directly on the wavelet coefficients. In another work by Srinivasan et al. an autoregressive modeling was applied on the ECG signals and AR coefficients were used for their classification into different types of arrhythmias [20]. More recently deep learning techniques have found their way into this problem [3,7]. We reported an LSTM based model, whereas Acharya et al.

* Corresponding author.
** Corresponding author.
E-mail addresses: sucheta@sciwhylab.org (S. Chauhan), shandar@jnu.ac.in (S. Ahmad).

https://doi.org/10.1016/j.compbiomed.2019.04.009
Received 29 January 2019; Received in revised form 11 April 2019; Accepted 13 April 2019

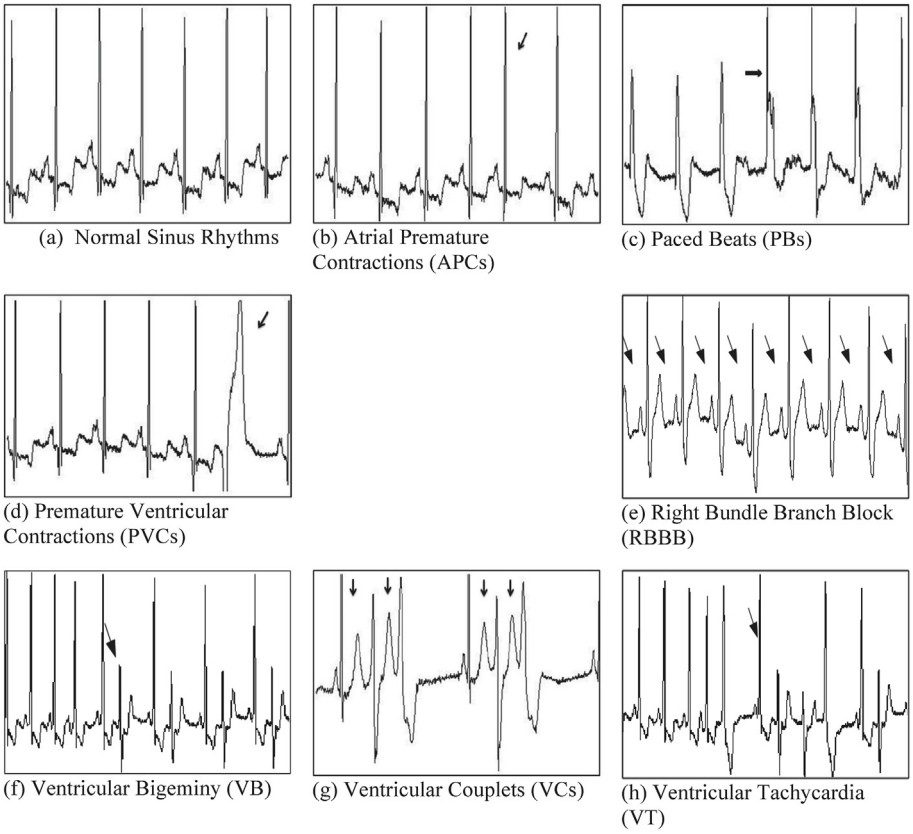

Fig. 1. Eight different types of beats in ECG recordings (seven anomaly types and a normal).

have presented a convolutional neural network (CNN) technique to automatically diagnose different ECG segments with potential anomalies [3]. The latter group designed eleven-layer deep CNN network with the output layer of four neurons, each representing the normal (NSR), A-Fib, AFL, and V-Fib abnormal ECG class resulting in encouraging performance levels.

In our previous work [7], we utilized the probability distribution of prediction errors from LSTM based recurrent models, trained on normal ECG signals to label normal and abnormal patterns on previously unseen recordings. The main advantage of using LSTM network was that the ECG signal could be directly fed into the network without any elaborate preprocessing as required by other techniques. Also, the networks needed no prior information about abnormal signals as they were trained only on normal data. The application of LSTM model-derived prediction errors as the indicator of anomaly proved to be successful but the study was limited to the detection of anomaly *in general* and employed a basic error annotation technique. The model did not evaluate if the same error patterns could be retrained to predict the actual anomaly class instead of making a simple binary class prediction. Such a detailed prediction would be helpful in better diagnostic systems and naturally provide for a better patient care in clinical conditions.

In this work, we propose to extend the LSTM anomaly detection algorithm by introducing a second stage predictor that can identify the actual anomaly class from the outputs of the first stage model. The anomaly classes that we have investigated in this work are seven types of arrhythmias: Atrial Premature Contraction (APC), Paced Beat (PB), Premature Ventricular Contraction (PVC), Right Bundle Branch Block (RBBB), Ventricular Bigeminy (VB), Ventricular Couplets (VCs) and Ventricular Tachycardia (VT). Using the prediction errors of LSTM model as the inputs, we have evaluated the performances of multi layer perceptron (MLP), Logistic regression (LR), and support vector machine (SVM) in estimating the true class of ECG anomaly out of these seven. Our results indicate that the ECG data sets trained on LSTM may be used to compute error vectors and subsequently their summary features over a fixed time interval. These newly introduced features have useful information to be further trained to classify original signals with the help of a second stage predictor such as MLP, SVM or LR. We also discuss how the accuracy scores without attention to inherent class imbalances may be misleading and hence an accurate background model is needed to estimate true predictive performance of multi-class models.

2. Methods

2.1. Source data sets and annotations

All the ECG dataset used in this work and their annotations have been taken from MIT-BIH Arrhythmia Database of which anomaly without a class reference was used in our previous paper [7,15]. In brief, we have used 1 min-long ECG recordings for each annotation. The seven types of pathological ECG beats, being considered here and as shown in Fig. 1, are as follows:

1) Normal Sinus Rhythm: It is a normal heart rhythm shown in Fig. 1(a).
2) Atrial Premature Contraction (APC): It is a premature beat that commences in atria and occurs earlier than normal beats, shown in Fig. 1(b).
3) Paced Beat (PB): Refers to an increase in heart rate. Paced beats start from 4th beat in Fig. 1 (c).
4) Premature Ventricular Contraction (PVC): It is also a premature beat but commences in ventricles. One single PVC is shown in Fig. 1(d)
5) Right Bundle Branch Block (RBBB): These rhythms originate above the ventricles. Here right ventricles are not directly activated by impulses travelling through the right bundle branch. In Fig. 1(e), all rhythms belong to RBBB.

6) Ventricular Bigeminy (VB): Each normal sinus impulse is followed by a ventricular premature beat. VB starts from fifth beat in the Fig. 1(f).
7) Ventricular Couplets (VCs): These are sequence of two consecutive PVCs. In Fig. 1(g), first, second, fourth and fifth beats are PVCs.
8) Ventricular Tachycardia (VT): It is fast abnormal heart rate commences in ventricles. Three or more PVCs in a row make VT. In Fig. 1(h), sixth, seventh and eighth beats are VT.

It may be clarified that prediction models could have also been developed by first splitting the one minute ECG recordings into short pulses and attempting to predict abnormal *pulse* from a pathological one. Indeed some of the previous methods seem to use this partitioning scheme of the data. However, in this work, we have used an entire one-minute signal as pathological or normal one, which is explained below.

(1) In a whole one-minute recording, the number of pathological *pulses* are very small compared to normal pulses and the prediction models are likely to be biased towards predicting trivial majority class. By annotating the whole one-minute recordings as pathological or normal, class balance is improved.
(2) Models trained in this way are now class-balanced but the challenge in accurate predictions still remain large because now the pathological signal may easily get lost in the overall normal pattern. The model performances, therefore can be better assessed by introducing a background model. The background model mimics the behavior of a trivial predictor. Results for such background models have been included in this work.
(3) One-minute ECG recordings may contain more than one loci of abnormal pulse. A predictive model using an entire recording can utilize this conveniently.
(4) Once a pathological annotation is assigned to the whole recording, our previously published LSTM-based model can still identify the exact site of anomaly for which performances have been benchmarked and reported earlier.

2.2. Overall prediction model

The details of the proposed models evaluated in this work, together with the previously reported LSTM workflow, are illustrated in Fig. 2. Each component of this scheme is briefly described below:

2.3. Long Short-Term Memory (LSTM) networks

Long Short-Term Memory (LSTM) networks are an extension of Recurrent Neural Networks (RNNs) and basically extend the memory of the latter. RNN was known to suffer from the vanishing gradient problem, and to overcome this, Hochreiter et al. in 1997 [11] used multiplicative units of a new model called Long Short Term Memory (LSTM). Multiplicative units are called Input gate, forget gate and output gate in order to preserve and select information useful for training and prediction [11]. Detailed architecture of LSTM network employed in this work is the same as we reported earlier [7]. The essential details of this LSTM network are included below for quick understanding of the current model (see Fig. 2). In our previous work, we trained stacked LSTM based Recurrent Neural Network on non-anomalous training set (s_N) using (val_N) as the validation set for early stopping. Subsequently, the trained LSTM network was used to predict signal strengths on $test_{N+A}$ i.e. the combination of anomalous and non-anomalous sequences [14]. For the input $x^{(t)}$, the predictions are made for every $l < t < n - l$ values l times i.e. next 10 predictions, where n is the length of the sequence.

2.4. Error calculation for deriving features for second stage model inputs

For each pattern in our anomaly detection data, starting with an initial window as the LSTM inputs, we predict an entire ECG waveform from scratch and compare it with the actual ECG signal. The difference between the predicted and actual ECG signals is an error vector that has the same dimensionality as the original ECG signal but from which the normal components of the signal have been effectively subtracted. Formally, the network prediction at time t is $p_i^{(t)} = [\,\bar{x}_i^{(t+1)}, \ldots\ldots, \bar{x}_i^{(t+l)}]$ where $\bar{x}_i^{(t+1)}$ is the next predicted value in the sequence corresponding to $x_{i+1}^{(t+1)}$ for that i^{th} time point. We calculate an error vector $e^{(t)}$ for point $x^{(t)}$ as $e^{(t)} = [e_1^{(t)}, \ldots., e_l^{(t)}]$, for simplification we can write: $e_i^{(t)} = [(\bar{x}_{i-1}^{(t+1)} - x_i^{(t+1)}), \ldots.,(\bar{x}_{i-l}^{(t+l)} - x_i^{(t+1)})]$ where subscript gives the information of the next i^{th} time point ($i \le l$) in the sequence and superscript gives the information of $1 \le l \le 10$ next predictions for that time point. So, $e_1^{(t)} = (\bar{x}_{i-1}^{(t+1)} - x_i^{(t+1)})$, $e_2^{(t)} = (\bar{x}_{i-2}^{(t+2)} - x_i^{(t+1)})$ and $e_l^{(t)} = (\bar{x}_{i-l}^{(t+l)} - x_i^{(t+1)})$ likewise we computed ten error vectors corresponding to each value in the sequence. As described above, even though, the anomaly may be in any region within this ECG error vector, the entire error vector is used for computing the summary features

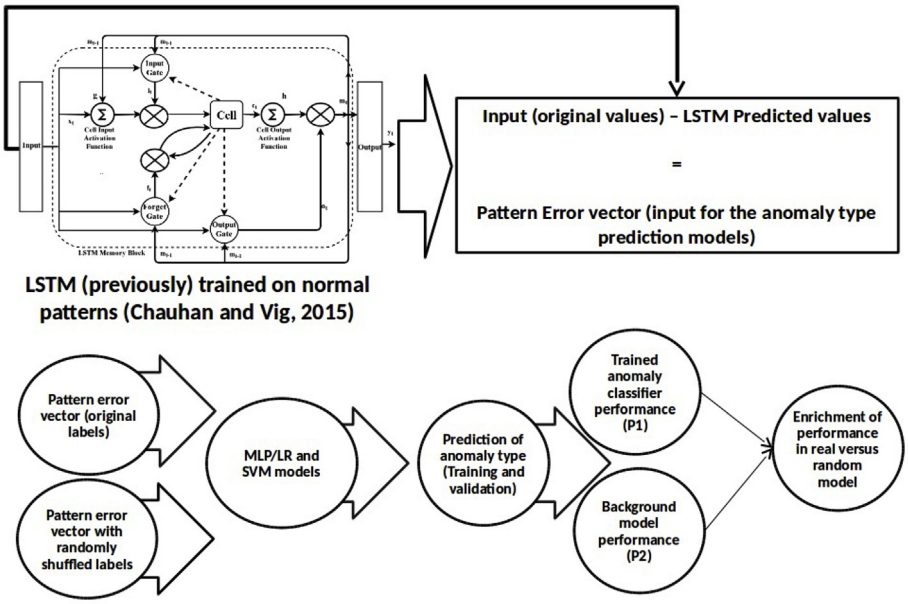

Fig. 2. Workflow of ECG arrhythmia classification.

representing the entire waveform (see below). This summary feature vector is used for the second stage models for anomaly class prediction.

2.5. Second stage model inputs

As stated above, second stage models -employing prediction errors derived from the first stage (LSTM network)- are computed as derivative summary features described below without further preprocessing. We found these summary features to be more effective than raw error vectors and therefore discuss only the models based on them in this work. Summary feature construction protocol is described below:

Step 1: using initial seed from an ECG signal, predict an entire one-minute recording using trained LSTM networks. 10 predictions are made for each point for every sliding window position in the initial seed, which will be merged in Step 3.
Step 2: compute error vector of size $e_{(2140\times10)}$ by subtracting original ECG signal from the LSTM-predicted one.
Step 3: average the ten values of each time point to generate the final error vector of length 2140 i.e. $e_{(2140\times1)}$
Step 4: compute summary features:

a) Select ten maximum and minimum error values from $e_{(2140\times1)}$ (top and last 10 values in the sorted vector).
b) Compute three statistical measures viz.
   i) Mean of all 2140 values in $e_{(2140\times1)}$
   ii) Median of all 2140 values in $e_{(2140\times1)}$
   iii) Standard deviation of all 2140 values in $e_{(2140\times1)}$
c) Take ten previous and ten next time points from the position with highest error (to enrich the most likely anomaly position in the summary features). For $e^{(t)}$where $e^{(t-10)} \le e^{(t)} \le e^{(t+10)}$
d) Build summary feature *sf* vector from (a, b, and c) of length $sf_{(43\times1)}$

From the above-explained steps, we have $sf_{(43\times1)}$ i.e. 43 summary features for each ECG recording.

2.6. Target vector representation

We have used one hot sparse encoding for creating the target vector. We maintained target vector of size (1×8) for each sample. Only one bit is labeled as bit "1" at a time, leaving the remaining bits as "0". The exact position of non-zero bit in a target vector shows the presence of the particular category (anomalous or normal) of that waveform.

2.7. Multilayer perceptron

In this work, we have used a deep MLP architecture with 43 input units at the input layer, two hidden layers with 300 units in each and 8 output units at the output layer. In the *forward pass*, the signal flows from the input layer through the hidden layers to the output layer, and the decision of the output layer is measured against the ground truth labels (see target vector representation, defined above). In the *backward pass (during training only)*, we used backpropagation and the chain rule of differential calculus taking partial derivatives of the error function with respect to the various weights and biases and the same are back-propagated through the MLP. We have used the *ReLU* activation function at the hidden layers and sigmoid activation function at the output layer respectively. We used *softmax* activation function at the output layer:

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{k} e^{y_j}}$$

2.8. Logistic regression (LR)

Logistic Regression is a classification algorithm that predicts the probability of occurrence of an event by fitting data to a *logit* function. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In multiclass logistic regression, instead of $y = 0,1$ we can expand to $y = 1,\ 2,\ 3,\ ....,\ n$, where $n = 8$ as we are working with 7 anomalous and one normal ECG classes. Basically, LR model runs binary classification multiple times, once for each class. So, effectively one LR model is developed for each dimension in the target vector defined above and corresponds to labeling a single anomaly type as positive class while all others anomaly classes are negative for that LR model. This will produce multiple outcomes, one from each model, of which the final prediction from the class, which has the maximum probability (predicted score) is selected as a consensus prediction of LR models.

2.9. Support vector machine

SVMs (Support Vector Machines) are a useful technique for data classification. Given a training set of instance-label pairs: $(x_i, y_i),\ i = 1, ...,l\ \ and\ x_i \in \Re^n and\ y \in \{1, 0\}^l$, where l is the number of data points i.e. 140, the support vector machines (SVM) [5,8] require the solution of the following optimization problem:

For each data point i, we need to introduce a variable $\xi_i = \max(0,1 - y_i(w.\ x_i - b)$. where ξ_i is the smallest non negative number which satisfies $y_i(w.\ x_i - b) \ge 1 - \xi_i$. Thus the optimization-$min\ \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i$. Subject to $y_i(w^T\varphi(x_i) - b) \ge 1 - \xi_i,\ where \xi_i \ge 0$ for all i.

Here training vectors x_i are mapped into a higher dimensional space by the function φ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $k(x_i, x_j) \equiv \varphi(x_i)^T\varphi(x_j)$ is called the kernel function [21]. We have used radial basis function (RBF):

$$k(x_i, x_j) = exp(-\gamma\ ||x_i - x_j\ ||^2),\ \ \gamma > 0$$

here, γ is a kernel parameter.

2.10. Implementation and performance evaluation

All models trained in this work were implemented in Python programming language using Adam optimizer for training the deep MLP and LSTM model [13] in Tensorflow framework [1]. C-SVM i.e. Support vector classification was used which is based on libSVM framework implemented in python using *scikit-learn* module. All machine learning techniques used in this work like *deep MLP*, *Logistic Regression*, and Support *vector machine* have been trained and validated using Jackknife or leave-one-out cross validation method. In jackknife method, in each fold only one sample is taken out from the complete dataset to test, rest n− 1 samples are used in training. The performance of the prediction system is evaluated using precision, recall, F-score, and Matthews correlation coefficient (MCC). These measurements are expressed in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP). The measurements are defined as follows:

$$Accuracy\ (\%)\ =\ \frac{TP + TN}{(TP + FP + TN + FN)} \times 100$$

$$Precision\ (\%)\ =\ \frac{TP}{(TP + FP)} \times 100$$

$$Recall\ (\%)\ =\ \frac{TP}{(TP + FN)} \times 100$$

$$F_{\beta=1}(\%) = 2\frac{(Precision * Recall)}{(Precision + Recall)} \times 100$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}}$$

2.11. Computation of a background model

In this work, we are dealing with a highly imbalanced data with multiclass predictions. Performance scores can be heavily indicative of the class balance for any given dimension instead of representing true predictive value of a model. To overcome this problem, we computed a background model for each anomaly class to estimate how much performance score is expected for a randomly assigned set of class labels, so that the gain in performance of every trained model in its reference may be computed. Two background models are defined for our work as follows:

1) Background Model 1 (BG Model 1): We obtained this model after shuffling the actual target labels 20 times for each class without retraining the model and measured the performance on various metrics. It is a common background model.
2) Background Model 2 (BG Model 2): We shuffled the actual target values 20 times for each class and retrained the model for those values and measured the average performance on each model.

Final unbiased performance estimates for any of the given metrics, defined above (e.g. F-score or AUC or ROC) can be easily computed from the assessment of how much that score has been enriched with respect to these background models.

3. Results

As stated in the Methods, we have first predicted ECG signals using our previously trained LSTM and then taken a (signed) difference between the predicted and actual signal intensities at all available time points in a given recoding. These error vectors for each recording are then converted to 43 summary features as described above. Using these 43 input features for the second stage predictor, we have used multiple models such as SVM, MLR and MLP in a leave-one-out manner. For comparison full length error vectors without converting them to summary features are also developed to assess predictive performance of different feature sets. Two background models by reassigning class labels are also trained likewise as stated in Methods. Results of these training experiments are presented below.

3.1. The best model selection choice to predict each anomaly type in three different conditions

Table 1 shows the ability of different classifiers to solve the anomaly classification problem being addressed in this work in a nutshell. More detailed results are presented in Fig. 3. Further, Fig. 4 is added to highlight key points from these results and the same are also summarized in the following:

1. Table 2 shows the model-wise classification performance obtained using jackknife cross-validation for the classifier processing for each category. As observed from this table, the classification accuracy of deep MLP, Logistic regression, and SVM model varies substantially across anomaly type, featurization and selected model (e.g. MLP, LR and SVM model resulted in 42.86%, 51.43%, and 50.0% accuracy overall respectively).
2. SVM is the best model to predict PB anomaly when trained on summary features. However, SVM is not the best model for all types of anomalies or when raw ECG signals are used for training them.

Table 1
On the basis of F-score, the best and worst predictive models are selected for each anomaly type in three different conditions. Symbol notations are defined below.

When network is trained on 3 conditioned data set:	Best Prediction	No prediction
1) Original ECG signals	✓	X
2) complete error vectors	✓	✗
3) Summary Features	☑	☒

Blank cells and missing symbols in each cell represent that these anomalies are somehow captured by the different models with low F-score.

Category (#anomalies)	LSTM	DeepMLP	LR	SVM
Normal (34)			✓ ☑	✓
PVC (26)			☑	✓ ✓
VC (20)		✓	☑	✓
RBBB (16)	X ✗	✓ ✓	X ✗ ☑	X ✗
VT (16)	✓	✓ ☑	X	✗
APC (14)	X ✗	X	X ✓ ☑	✗
VB (8)	X ✓	X	X ✗ ☑	X ✗ ☒
PB (6)	X	X ✗	✓ ✓	X ✗ ☑

Superiority of summary features on SVM models could be because summary based featurization creates smaller input vectors which can be trained to higher level of accuracies on training data without causing over-fitting.

3. SVM models completely fail to detect the anomaly type VB. Even though other classifiers are slightly better, they also show quite weak prediction scores. Exact cause of difference between performance for PB and VB on the same model and featurization is difficult. Intutively, paced beat (PB) anomaly spreads over a longer region in the entire ECG recording and therefore summary features will receive stronger signals to enable their identification. It may also be that PB anomalous patterns are more conserved across patients and are easier to learn from most models. On the contrary VB patterns are more complex and due to a small number of patients in the category, achieving generalization without overfitting is more challenging. It may also be due to the fact that VB is more of a sequential property of a signal, which might be lost upon summarization.
4. RBBB is only captured by deep MLP on complete error vector, while Logistic regression is more successful when trained on summary features.
5. Although the overall accuracy of SVM classifier is higher than other classifiers in case when network is trained and tested on complete error vector and on original ECG signal, these good scores mainly come from the majority class data (i.e. normal, PVC and VC class sequences).
6. The other classes (minority classes i.e. RBBB, VT, APC, PB, VB) have very low F-score. RBBB and PB are two anomalies, which are strongly predicted by Logistic regression and SVM model when trained on summary features.
7. On summary features Logistic regression (LR) model performs the best in terms of accuracy and F-score as well because most of the anomalies are better predicted than other models by them. RBBB and APC are never captured by LSTM.
8. Logistic regression has poor predictability for RBBB and VB on complete error vector while it strongly predicts both anomalies on

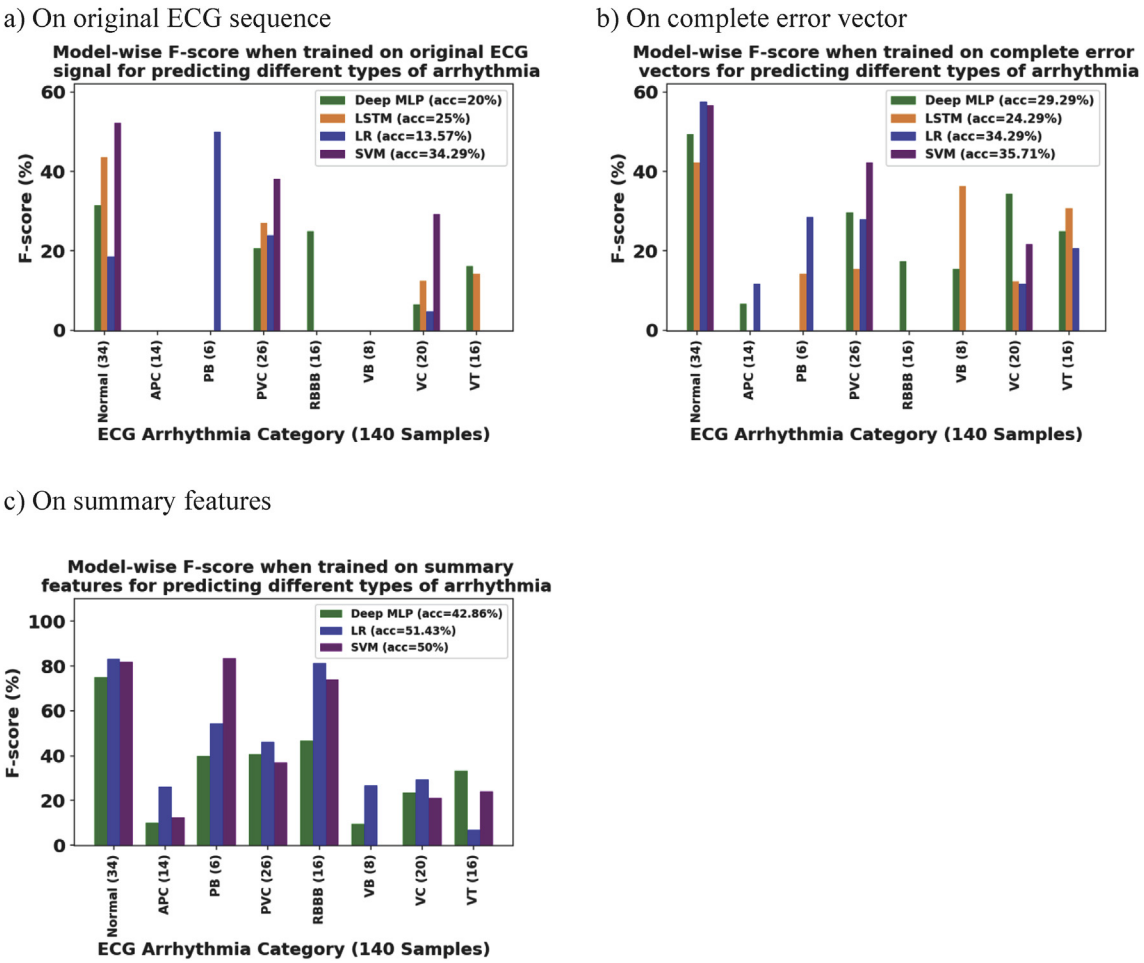

Fig. 3. F-score performance, measured when four different models have been trained and tested on three different data conditions.

summary features with highest F-score.

9. APC, RBBB, PB and VB are not captured by LSTM with either the original ECG signals or their error-based summaries. However, original ECG signals work better for this model (see below).
10. LSTM performs poorly on whole waveform data but does not do well on summary data. It is intuitively caused because LSTM are more suitable for time series signals, and use of summary features actually disrupts the time-series nature of data (see point [3] above).
11. Detailed analysis of confusion matrices from predicted scores confirms that the performance levels are generally dependent on data availability in each class but some exceptions do occur (see

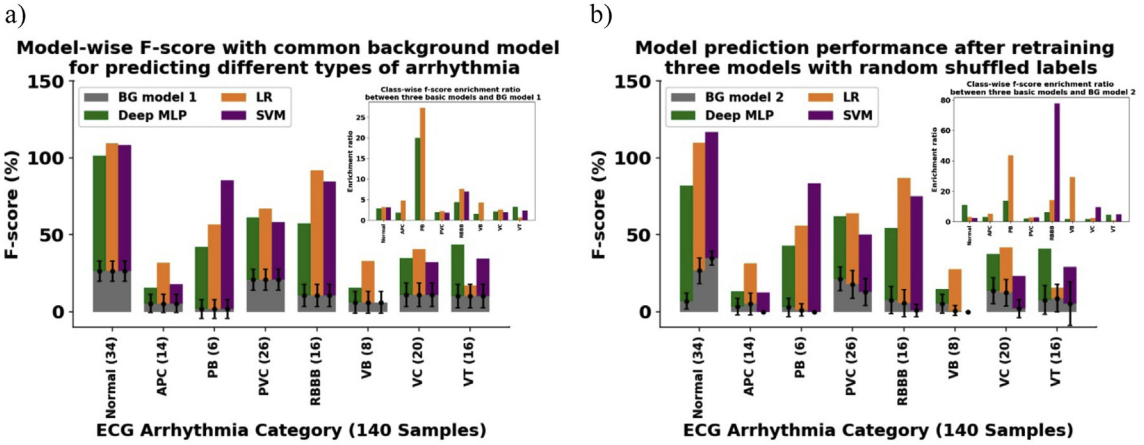

Fig. 4. Model-wise F-score performance measure for predicting class labels for different types of arrhythmia present in ECG recordings. Here, (a) one common background model is obtained after shuffling the actual class labels randomly (without retraining) and (b) is obtained after retraining the network with randomly shuffled class labels of arrhythmias. Inset in (a) shows the ratio difference among three basic models (i.e. deep MLP, LR, and SVM) and background model 1 (model obtained without retraining) and inset (b) shows the ratio differences among three basic models and background model 2 (obtained after retraining the model with randomly shuffled labels).

Table 2
Model-wise performance measures for predicting each anomaly class present in ECG recordings. Pr: Precision, Re: Recall, F: F-score, and MCC: Matthews correlation coefficient. Additional details of confusion matrix in predictions have been included in Supplementary data.

Anomaly type (#of patterns)	Deep MLP				LR				SVM			
	Pr (%)	Re (%)	F (%)	MCC	Pr (%)	Re (%)	F (%)	MCC	Pr (%)	Re (%)	F (%)	MCC
Normal (34)	71.05	79.41	75	0.67	74.42	94.12	83.12	0.78	69.39	100	81.93	0.77
APC (14)	16.67	7.14	10	0.05	33.33	21.43	26.09	0.20	50.00	7.14	12.50	0.16
PB (6)	50	33.33	40	0.39	60.00	50.00	54.55	0.53	83.33	83.33	83.33	0.83
PVC (26)	32.56	53.85	40.58	0.24	46.15	46.15	46.15	0.34	29.55	50	37.14	0.19
RBBB (16)	50	43.75	46.67	0.40	81.25	81.25	81.25	0.79	90.91	62.50	74.07	0.73
VB (8)	7.69	12.50	9.52	0.03	28.57	25.00	26.67	0.23	0	0	0	−0.02
VC (20)	28.57	20	23.53	0.14	28.57	30.00	29.27	0.17	22.22	20.00	21.05	0.09
VT (16)	50	25	33.33	0.30	7.69	6.25	6.90	−0.04	33.33	18.75	24	0.18

Table 3
F-score prediction enrichment score for each anomaly class from Background Model 1 and 2. Enrichment score (EnR) is the ratio between actual model prediction F-score and background model prediction F-score, shown in Fig. 4.

Anomaly type (#of patterns)	Background Model 1			Background Model 2		
	EnR-MLP	EnR-LR	EnR-SVM	EnR-MLP	EnR-LR	EnR-SVM
Normal (34)	2.84	3.15	3.10	10.83	3.12	2.35
APC (14)	1.82	4.74	0	2.97	5.05	0
PB (6)	20	27.27	0	13.66	43.64	0
PVC (26)	1.94	2.21	1.78	1.88	2.59	2.84
RBBB (16)	4.38	7.62	6.94	6.16	14.19	77.78
VB (8)	1.54	4.31	0	1.77	29.33	0
VC (20)	2.11	2.62	1.88	1.70	2.35	9.44
VT (16)	3.20	0.66	2.30	4.30	0.78	4.61

Supplementary Table ST1 and Supplementary Figure SF1).

3.2. Background model performances and real model enrichments

While training multiclass models, standalone accuracy scores such as precision, recall, AUC and F-score may be misleading as seen from the Tables 2 and 3. Some of the scores are less than 50% giving the impression that prediction scores are poorer than a random model. However, in the given data class imbalance scenario, random model scores are not 50%. To accurately estimate these scores, we created two background models as described in Methods, by shuffling class labels in two ways i.e. by randomly assigning class labels for all 8 categories or by shuffling within a binary class label column. Table 3 summarizes the enrichment scores for each of these performance estimates together with performances on background models. We observed that in all cases, RBBB anomalies are best predicted among all the anomaly classes considered here and reaches as much as more than 70% the background models. Some of the models are found to show poorly enriched scores, suggesting corresponding model is not suitable for such anomaly detection.

4. Discussion

In this work, we have implemented various computational models to predict anomaly class from LSTM model errors. The power of the method is that LSTM model is trained on normal data sets and errors that are returned from such models on a new data are scale-free as LSTM is automatically scaled to signal strengths. We observed that the second stage LSTM model was not effective and that is likely due to two reasons. First of all summary features of LSTM model errors are not a time-series data and LSTM is not a suitable model for such stationary data. Secondly the errors are derived from LSTM model itself and any data structure or patterns detectable by LSTM are already implicit in the first stage error vectors.

A significant point that we made in this work is that background models are important for reporting performance levels on highly imbalanced multi-class data. We notice that ECG data with randomized class labels are not only sensitive to class population but also whether the anomaly patterns are conserved within the category. In other words, the same anomaly class could be seen in somewhat different ECG waveforms. Performance scores of each anomaly class also reflect the inherent structure of the specific anomaly e.g. they may occur over a longer or narrow subspace. Knowledge of these diversities or lack thereof is crucial for accurate estimates of model performances and confidence to apply such models on clinically deployable systems.

Performance levels in this work are often several folds better than a random background model, which shows that the experimental design and outcomes are effective. However, much ground needs to be covered to take such results to substitute for human expertize. Major bottleneck in having a high performance diagnostic system appears to be the availability of high quality data for many patients with manually assigned class labels. We believe that more clinical data collection with diverse scenarios is needed to develop ECG anomaly class prediction methods and model performances will improve over time within the framework proposed in this work.

5. Conclusion

We successfully implemented a novel model that utilizes LSTM prediction errors as input features for a second stage anomaly class prediction problem. Results are promising although not coming close to a 100% accurate model. Data paucity and class imbalance require definition of accurate background models, two of which have been proposed in this work. We believe these findings will be helpful in real-time monitoring and cardiac anomaly class alert systems. However, more data will be needed to improve the performances further.

Funding

This work has been partially supported by grants from CSIR (37 (1695)/17/EMR-II), DBT (BT/PR24208/BID/7/801/2017), DST-SERB (EMR/2017/005485) and DST-ICPS (DST/ICPS/CLUSTER/Data Science/2018/General/T-136) projects to SA and DST-PURSE grant to JNU.

Conflicts of interest

None declared.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.compbiomed.2019.04.009.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a System for Large-Scale Machine Learning, OSDI, 2016.

[2] R. Acharya, A. Kumar, P. Bhat, C. Lim, N. Kannathal, S. Krishnan, Classification of cardiac abnormalities using heart rate signals, Med. Biol. Eng. Comput. 42 (3) (2004) 288–293.

[3] U.R. Acharya, H. Fujita, O.S. Lih, Y. Hagiwara, J.H. Tan, M. Adam, Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network, Inf. Sci. 405 (2017) 81–90.

[4] N. Belgacem, M. Chikh, F.B. Reguig, Supervised Classification of ECG Using Neural Networks, (2003).

[5] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, 1992.

[6] R. Ceylan, Y. Özbay, B. Karlik, A novel approach for classification of ECG arrhythmias: type-2 fuzzy clustering neural network, Expert Syst. Appl. 36 (3) (2009) 6721–6726.

[7] S. Chauhan, L. Vig, Anomaly detection in ECG time signals via deep long short-term memory networks, Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, IEEE, 2015.

[8] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[9] P. De Chazal, B. Celler, R. Reilly, Using wavelet coefficients for the classification of the electrocardiogram, Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE, IEEE, 2000.

[10] M. Engin, ECG beat classification using neuro-fuzzy network, Pattern Recogn. Lett. 25 (15) (2004) 1715–1722.

[11] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[12] L. Khadra, A. Al-Fahoum, H. Al-Nashash, Detection of life-threatening cardiac arrhythmias using the wavelet transformation, Med. Biol. Eng. Comput. 35 (6) (1997) 626–632.

[13] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014) arXiv preprint arXiv:1412.6980.

[14] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, Long short term memory networks for anomaly detection in time series, Proceedings, Presses universitaires de Louvain, 2015.

[15] R. Mark, P. Schluter, G. Moody, P. Devlin, D. Chernoff, An annotated ECG database for evaluating arrhythmia detectors, IEEE Transactions on Biomedical Engineering, IEEE-Inst Electrical Electronics Engineers Inc 345 E 47TH St, NEW YORK, NY, 1982.

[16] M. Mitra, R. Samanta, Cardiac arrhythmia classification using neural networks with selected features, Proc. Technol. 10 (2013) 76–84.

[17] Y. Özbay, B. Karlik, A Recognition of ECG Arrhythmias Using Artificial Neural Networks, Seluk univ konya (turkey) electrical and electronics engineering, 2001.

[18] D. Patra, M.K. Das, S. Pradhan, Integration of Fcm, Pca and neural networks for classification of Ecg arrhythmias, IAENG Int. J. Comput. Sci. 36 (3) (2009).

[19] K. Polat, S. Güneş, Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine, Appl. Math. Comput. 186 (1) (2007) 898–906.

[20] N. Srinivasan, D. Ge, S. Krishnan, Autoregressive modeling and classification of cardiac arrhythmias, Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint, IEEE, 2002.

[21] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, A Practical Guide to Support Vector Classification, (2003), pp. 1–16.