scientific reports

OPEN

A contrastive learning approach for ICU false arrhythmia alarm reduction

Yuerong Zhou[1], Guoshuai Zhao[1]✉, Jun Li[2], Gan Sun[3], Xueming Qian[1], Benjamin Moody[4], Roger G. Mark[4] & Li-wei H. Lehman[4]✉

The high rate of false arrhythmia alarms in Intensive Care Units (ICUs) can lead to disruption of care, negatively impacting patients' health through noise disturbances, and slow staff response time due to alarm fatigue. Prior false-alarm reduction approaches are often rule-based and require hand-crafted features from physiological waveforms as inputs to machine learning classifiers. Despite considerable prior efforts to address the problem, false alarms are a continuing problem in the ICUs. In this work, we present a deep learning framework to automatically learn feature representations of physiological waveforms using convolutional neural networks (CNNs) to discriminate between true vs. false arrhythmia alarms. We use Contrastive Learning to simultaneously minimize a binary cross entropy classification loss and a proposed similarity loss from pair-wise comparisons of waveform segments over time as a discriminative constraint. Furthermore, we augment our deep models with learned embeddings from a rule-based method to leverage prior domain knowledge for each alarm type. We evaluate our method using the dataset from the 2015 PhysioNet Computing in Cardiology Challenge. Ablation analysis demonstrates that Contrastive Learning significantly improves the performance of a combined deep learning and rule-based-embedding approach. Our results indicate that the final proposed deep learning framework achieves superior performance in comparison to the winning entries of the Challenge.

ICUs are designed to provide acute care for patients with severe and life-threatening injuries or illnesses using sophisticated bedside monitors such as pulse oximeter (PPG), electrocardiogram (ECG), arterial blood pressure (ABP) catheter, central venous pressure catheter and ventilators. Ideally, these monitors with a built-in alarm system can send an alert to the healthcare providers when a patient's physiological signals are out of pre-defined ranges. On the one hand, arrhythmia alarms in the ICU based monitors are deliberately designed to be highly sensitive in order not to miss any life-threatening events. However, high sensitivity compromises the specificity of these alarms[1]. According to Drew et al.[2] the false alarm ratio in ICUs can be as high as 88.8%.

Alarms are falsely triggered by many factors, including noise and artifacts from patient movement, power line interference, electrode contact noise, and data collecting device noise. Falsely triggered alarms become an unseen threat in ICUs for they not only lead to sleep deprivation[3], inferior sleep structure[4], stress for both patients and staff[5] and depressed immune systems[6], but also put patients at risk for the desensitization to warnings and slowing of response times. By contrast, only 2–9% of all ICU alarms are correctly triggered and these alarms do require an urgent and professional response[7]. Therefore, false alarms present an important problem in ICUs today[8].

Methods proposed to reduce the rate of false arrhythmia alarms in the PhysioNet 2015 Challenge[9] can be roughly divided into two categories: rule-based methods and machine learning methods. The best-performing methods from the Challenge 2015 are mostly rule-based and require hand-engineered feature crafting. Rule-based methods mainly use expert-defined rule-based logic analysis to analyze patients' physiological signals[10]. Specifically, methods in this field mainly consist of signal quality evaluating and QRS-complex detection in order to analyze heart rate. Machine learning algorithms have been widely used in the medical field[11,12]. Peng et al. introduce machine learning techniques to transform the nuclear magnetic resonance (NMR) correlational map into user-friendly information for point-of-care disease diagnostic and monitoring[11]. Lau et al. utilize polymerization studies and two dimensional-nuclear magnetic resonance spectrometry (2D-NMR) to investigate the hypothesis that HOCl oxidation alters fibrinogen conformation and T2 relaxation time of water protons in the

[1]Xi'an Jiaotong University, Xi'an, China. [2]Nanjing University of Science and Technology, Nanjing, China. [3]Chinese Academy of Sciences, Shenyang, China. [4]Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ✉email: guoshuai.zhao@xjtu.edu.cn; lilehman@mit.edu

fibrin gels[12]. In the false alarm reduction problem, conventional machine learning based methods train a classification model using hand-crafted features as input to classify the alarms[13]. The performance of these prior methods depends highly on the quality of these hand-crafted features or on the design of rules which cannot automatically model the complex patterns in waveform data.

Although deep learning provides powerful representation learning techniques to automatically capture complex patterns in the data, conventional deep learning approaches for physiological waveform analysis in false alarm reduction have had limited success in out-performing the rule-based techniques in the PhysioNet 2015 Challenge[9]. False arrhythmia alarm reduction in ICUs is a challenging problem for deep learning approaches due to the high-dimensional data from long sequence length of the multi-channel physiological waveform signals, imbalanced classes of true vs. false alarms, and most importantly, a limited number of records with ground-truth labels due to the fact that expert-annotation of arrhythmia alarms is laborious and costly to obtain.

To address the above challenges, we design a novel deep contrastive learning framework to detect true arrhythmia alarms based on a CNN[14] architecture. In the proposed model, we use CNN as the signal encoder to automatically extract the features of the input signals relevant for our classification task. We propose to use the idea of Contrastive Learning with Siamese network[15] and discriminative constraints to learn an effective lower-dimensional representation of high-dimensional waveform signals to improve the signal encoder, prevent over-fitting and overcome the problem of insufficient training data. In addition, to leverage all available training records in the dataset, we train our deep learning models using records from all alarm types simultaneously, and use learned embeddings from records' alarm types as input to our deep learning models to enable classification across multiple alarm types at the same time. Finally, we augment our model by using a rule-based approach to learn an embedding as input to our deep models. This enables our technique to achieve label-efficient learning in a small labeled data setting by leveraging rule-based techniques that utilize known physiological structure of the signals for the classification task. Results on the unseen test set show that our method outperforms all submitted methods on the real-time event in the PhysioNet Challenge 2015[9].

Our main contributions are summarized as follows:

- We propose a deep learning model in false alarm reduction, using CNN as the signal encoder to reduce the length of input signals and detect temporal and spatial patterns in multi-channel waveform data.
- We develop a Contrastive Learning framework by using Siamese Network and calculating a discriminative constraint to prevent over-fitting and address the challenge of limited training data.
- We augment our deep learning model with embeddings generated by a rule-based method to leverage domain-knowledge specific to each alarm type for label-efficient representation learning.

Related work

Rule-based methods.

Plesinger et al. test each channel in the record for regular heart activity using the QRS distribution and derived R-R information[10]. Daluwatte et al. develop an algorithm based on global heartbeat annotations generated by fusing individual heartbeat detection from multiple physiological signals and then apply an arrhythmia criterion to the global heartbeat detection to classify the alarm[16]. Ansari et al. use a multi-modal peak detection algorithm and combines the results from several peak detection algorithms to create a robust peak detection algorithm[17]. Tsimenidis et al. propose a method that includes high-pass filtering to remove baseline instability, scaling to normalize waveform amplitudes, detection of noisy and flat waveforms, differentiation to accentuate sharp waveform edges, beat detection, timing between beats preceding alarm onset, and detection of alarm conditions[18]. He et al. use a derived signal quality index (SQI) to reveal the degree of signal quality[19]. The SQI-weighted residual error of Kalman filters (KF) is used to complete the data fusion for evaluating the heart rate (HR). Finally, the algorithm of arrhythmia false alarm reduction is developed based upon the method of combining SQIs and HR estimations derived from ECG and ABP waveforms. Fallet et al. estimate heart rate from pulsatile waveforms using an adaptive frequency tracking algorithm or computed from ECGs using an adaptive mathematical morphology approach based on the quality of available signals[20]. Furthermore, they introduce a supplementary measure based on the spectral purity of the ECGs to determine if a ventricular tachycardia or flutter/fibrillation arrhythmia has taken place. Finally, alarm veracity is determined based on a set of decision rules on heart rate and spectral purity values. Couto et al. use simultaneous ECG and pulsatile waveforms[21]. QRS detectors are used to produce for each signal a set of QRS detections which are to be used for detecting false alarms. In case some of the signals may be noise-contaminated, the signal quality of each waveform is evaluated to determine whether the QRS detection obtained on that waveform is reliable. A set of rules is then used for each alarm type. Although rule-based methods are effective and commonly used in medical fields, extensive expert knowledge is needed to design the rules and evaluate.

Traditional machine learning methods.

Antink et al. present an approach that analyzes multi-modal cardiac signals in terms of their beat-to-beat intervals as well as their average rhythmicity[22]. Based on this analysis, several features in time and frequency domain are extracted and used for several machine learning approaches. Eerikainen et al. train Random Forest classifiers for every type of arrhythmia with arrhythmia-specific features computed from signal quality information and physiological features[23]. Kalidas et al. use a combination of logical analysis and SVM-based machine learning techniques[13]. Information from original signals is used for logical analysis and to form the features set. Caballero et al. develop a decision tree for each arrhythmia category, which is combined with domain knowledge to produce a set of if/else statements[24]. Using the ABP and PPG signals, separate decision trees are trained. Afghah et al. propose a model based on coalition game theory that considers the inter-features dependencies in determining the salient predictors with respect to false alarms[25]. Antink et al. present an approach that analyzes multi-modal cardiac signals in terms of their beat-to-beat intervals as well as their average rhythmicity[22]. Based on this analysis, several features in time and frequency domains are extracted and used for subsequent machine learning tasks. Zaeri-Amirani et al. propose a low-computational complexity game-theoretic feature selection method which is based on a genetic algorithm that identifies the most informative biomarkers across the signals collected from various monitoring devices[26]. Au-Yeung et al. applies a Random Forest and meanwhile performed feature selection in order to reduce the complexity of the models and improve the efficiency of the algorithm[27].

Deep learning methods. Lehman et al. present a supervised generative model to classify ventricular tachycardia alarms using non-linear embeddings of ECG dynamics[28]. The model is a variant of a Denoising Autoencoder, learned using a combination of discriminative and generative loss. Furthermore, feature transformations are explored by utilizing known physiological structure within ECG signals to enable learning under the constraints of limited labeled data. To this end, a multi-stage approach is proposed to utilize the FFT-transform of consecutive heart beats. Hooman et al. present a method for training neural networks based on neuroevolution by utilizing the Dispersive Flies Optimisation algorithm in a gradient-free population-based scheme[29]. Mousavi et al. propose a deep learning-based network composed of the CNN layers, attention mechanism, and LSTM units to reduce false alarm arrhythmia in ICUs[30]. Yu et al. design a multi-channel deep group convolutional neural network for false alarm reduction[31]. Their model takes multi-channel raw signals as input and different kernels are used for convolution operation according to the type of alarm. Zihlmann et al.[32] propose two deep learning models for classification of arbitrary-length ECG recordings using the Physionet Challenge 2017[33] dataset. The first model is a deep CNN architecture with averaging-based feature aggregation across time. The second model combines convolutional layers for feature extraction with long-short term memory (LSTM) layers for temporal aggregation of features. They use two data augmentation techniques, dropout bursts and random resampling, for ECG data during training procedure. Hong et al. propose an ensemble classifier to combine expert features and deep learning models together for ECG classification[34] on Physionet Challenge 2017 dataset. Hyvarinen et al. propose a learning principle for unsupervised representation learning on time series[35], which is based on analyzing nonstationarity in temporal data by discriminating between time segments. Kiyasseh et al. propose a family of self-supervised pretraining mechanisms based on contrastive learning for physiological signals[36]. Pei et al. propose a model for time series analysis that learns a similarity measure over pairs of time series in a supervised manner[37]. In the Siamese Network, two time series are inputted to the same recurrent network for feature extraction. Wu et al. propose an end-to-end deep learning model to learn local representations of time series[38]. A local embedding loss is applied to optimize a Siamese Network and a feature space that preserves the temporal location-wise distances between time series can be learned in their framework.

Methodology

In deep learning, CNN has been successfully applied in many different domains such as image classification[39] as well as different natural language processing tasks[40,41]. Motivated by the success of CNN and its variants in these various domains, researchers have started using CNN for time series classification[42]. Commonly used CNN models such as FCN[43] and ResNet[44] are good at extracting local spatial features and ResNet can support a very deep architecture. However, in order to take advantage of the power of FCN and ResNet, these models use as many CNN layers as possible. It is hard for us to train since we have a limited number of training records. Therefore, we train a deep learning model to classify the arrhythmia alarms using a 1-dimensional CNN which can extract local features of 1-dimensional data such as time series data. It is small and easy to train. In the proposed model, four CNN blocks are used to extract the features of the raw input signals. Each block has a different kernel size. In order to improve the performance of the feature extractor and to avoid the overfitting problem, which is crucial in this problem, we propose a pair-wise loss function which utilizes contrastive learning. While other approaches use contrastive learning between different records, we utilize contrastive learning method and compute pair-wise loss between two different segments inside the same alarm record. Specifically, the proposed Siamese Network architecture learns latent representations of the signals through contrastive learning from two segments of the same patient waveform record, namely, the 'alarm-trigger signal' (or 'alarm signal' for short), which is the waveform segment that triggered the alarm, and a pre-alarm 'baseline signal' which is a randomly sampled waveform segment of the same length representing the baseline of the same patient prior to the alarm-triggering event. In addition, in order to leverage the domain-knowledge encoded in the rule-based methods, we augment our model with output from the rule-based method proposed by Plesinger et al.[10]. For each record, we feed the output of the rule-based method into our networks to generate an embedding. After converting raw signals into a representation vector, it is then concatenated with the alarm-type embedding and the rule-based embedding as input to a classification layer to generate the probability of a false alarm for each record. At last, the study was performed in accordance with the relevant guidelines and regulations and in accordance with the Declaration of Helsinki.

Model architecture. Figure 1 illustrates the proposed model architecture and deep learning framework. The architecture includes a signal encoder, two fully connected layers and a classification layer. The signal encoder is based on a CNN architecture, and hence, does not rely on heavy hand-crafting of feature engineering. We denote the number of time steps as T_l, the number of variables as M_l, the kernel size of CNN layer as D_l, the input as $X^{(l)} \in \mathbb{R}^{M_l \times T_l}$, the full sizes of filters as $W^{(l)} \in \mathbb{R}^{M_{l-1} \times D_l \times M_l}$, and the bias as $B^{(l)} \in \mathbb{R}^{M_l \times T_l}$ in the lth CNN layer. Let $X_{m,t}^{(l)}$ be the value of the mth ($0 < m \leq M$) variables with the tth ($0 < t \leq T$) time step input series. By the activation function $f(\cdot)$ Rectified Linear Unit(ReLU), we can get the value of each position for $\forall m \in \{1, 2, 3, \ldots, M_l\}, \forall t \in \{1, 2, 3, \ldots, T_l\}, \forall l \in \{1, 2, 3, \ldots, L\}$.

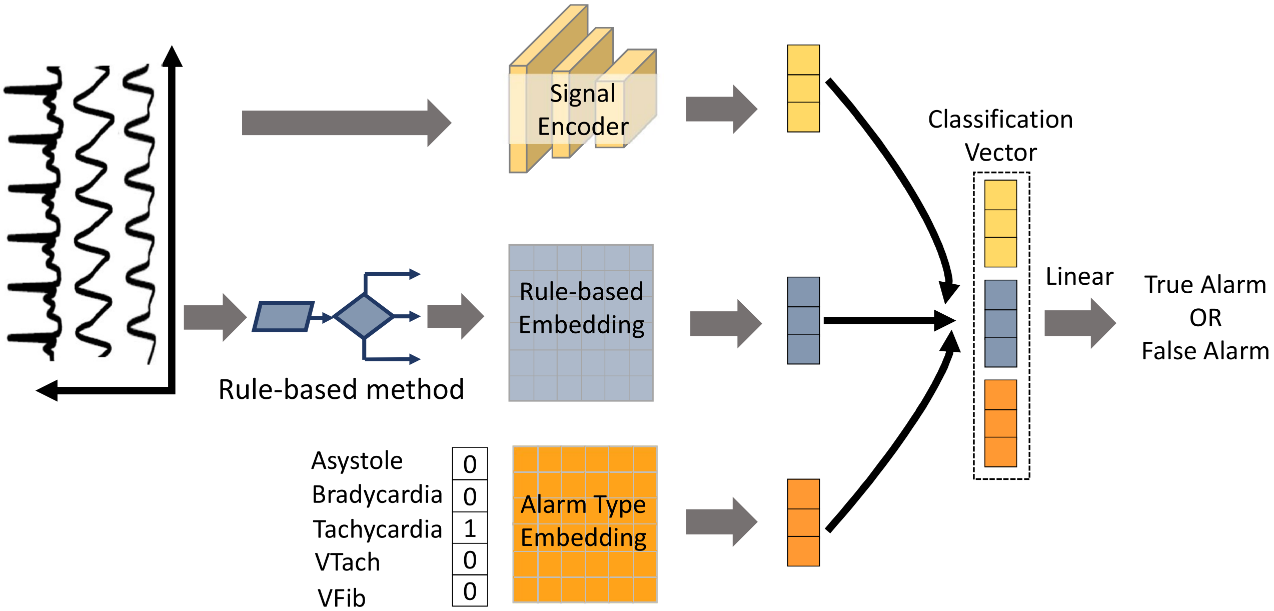

Figure 1. Illustration of the proposed model architecture and deep learning framework. Figure generated in PowerPoint version 1808, https://www.microsoft.com.

$$X_{m,t}^{(l)} = f\left(B_{m,t}^{(l)} + \sum_{j=0}^{D_l-1}\sum_{i=0}^{M_l-1} W_{i,j,m}^{(l)} X_{i,t+j}^{(l-1)}\right). \tag{1}$$

Two fully connected layers are the alarm embedding layer and rule-based embedding layer respectively. With different alarm types, signals may have different features and characteristics unique to each alarm type in distinguishing between a true vs. a false alarm. Since we already have the information about the triggered alarm type of each record in the training datasets, alarm type of each sample can provide useful information for the model when classifying. The alarm embedding layer converts the alarm type of a sample to an embedding with a fixed size. Given a one-hot vector $A \in \mathbb{R}^{1\times N}$ as an alarm type of a given record, the alarm embedding layer transforms it into:

$$E_a = AW_a, \tag{2}$$

where $W_a \in \mathbb{R}^{N\times S_a}$ is a trainable parameter of the alarm embedding layer, N indicates the number of arrhythmia alarm types and S_a indicates the size of alarm embedding.

The rule-based embedding layer converts the output by a rule-based method into an embedding. We denote the result output by the rule-based method as R. The rule-based embedding layer transforms it into:

$$E_r = RW_r, \tag{3}$$

where E_r is the embedding of the rule-based output, $W_r \in \mathbb{R}^{S_r}$ is a trainable parameter of the rule-based embedding layer and S_r indicates the size of the rule-based embedding.

We combine the strong learning ability of deep learning models and clinical rules and experiences from rule-based method by concatenating the latent encoding of raw signals E_e, the alarm type embedding E_a and the rule-based embedding E_r as follows:

$$E = E_e \oplus E_a \oplus E_r. \tag{4}$$

Finally the concatenated vector $E \in \mathbb{R}^{1\times(S_e+S_a+S_r)}$ is fed into the classification layer, which is comprised of a fully-connected layer followed by a sigmoidal output layer to get the output probability O of the triggered alarm being true:

$$O = \sigma(EW_c), \tag{5}$$

where $W_c \in \mathbb{R}^{(S_e+S_a+S_r)\times 1}$ is the trainable parameter of the classification layer and σ is the sigmoid function.

Loss function. In the false alarm reduction problem, the label of each record is just TRUE and FALSE, while the length of each waveform channel of each record is 75,000 samples (5-min segment sampled at 250 Hz), which means the supervised information is too small to train a deep learning model. Therefore, instead of only using the label of each record, we design a pair-wise similarity-based loss function calculated by using different segments in the waveforms of the same record as additional information to train our model. The choice of similarity-based loss is motivated by the fact that, clinically, the detection of a true VT event often involves compar-

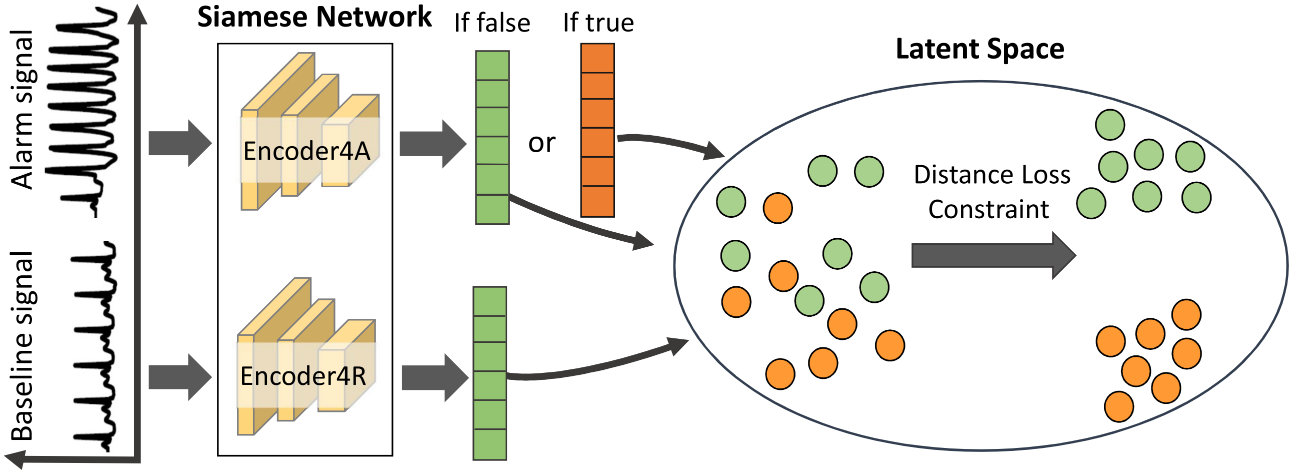

Figure 2. We use the idea of Siamese Network to calculate our discriminative constraint. Here, Encoder4A and Encoder4R take an 'alarm signal' and a 'baseline signal' sampled from the same waveform record as inputs and output their feature vectors respectively. The distance loss is then used to find the similarity of the inputs by comparing their feature vectors. The 'alarm signal' refers to the multi-channel waveform segment that triggered the alarm (e.g., 10-s segment prior to the alarm onset), and the 'baseline signal' segment is randomly sampled from a prior time interval of the same record. Figure generated in PowerPoint version 1808, https://www.microsoft.com.

ing the signals immediately prior to the alarm onset with signals from the same patient at an earlier time point to determine whether there has been a change in the patient's ECG from his or her baseline. Additionally, sampling from the same patient's baseline signals functions as a data augmentation scheme to increase our effective sample size to improve performance in a small labeled data setting. This similarity-based loss works as a discriminative constraint and is combined with the binary entropy loss during training. Our model takes the 10 s sequence of the multi-channel waveforms immediately prior to the alarm onset as the 'alarm signals'. We use alarm signals instead of the whole signals for classification since the exact time of the event that triggered the alarm is within 10 s of the alarm[45]. Using the alarm signals can effectively reduce the computational complexity and improve the accuracy of classification due to the difficulty of very long time series classification. Meanwhile, we also randomly sample a sequence with the same length as random baseline signals prior to time $t - 10$ s from the same patient as the pre-alarm 'baseline signal', where t is the alarm onset time. The signal encoder is a Siamese Network, which means it can be seen as two identical encoders, Encoder4A and Encoder4R. They have the same configuration with the same parameters and weights. Alarm signals and baseline signals are fed into Encoder4A and Encoder4R respectively to get their corresponding feature vector E_e^R and E_e^A. These two feature vectors are then used for calculating the discriminative constraints. Figure 2 illustrates the Siamese Network architecture in calculating the discriminative constraint. The discriminative constraint of a record depends on the ground-truth of its alarm. If the triggered alarm is false, then the feature vector of its baseline signals E_e^R should be close to the feature of the alarm signals E_e^A since the monitoring system misjudged the vital signs of the last 10 seconds by mistakenly triggering an alarm and these two feature vectors should be considered as a similar representation. Their constraint can be defined as follows:

$$C_{false}^{(i)} = -\log(\sigma(f(E^A)^T f(E^R))), \tag{6}$$

where $C_{false}^{(i)}$ is the discriminative constraint of a record with a false alarm in a mini-batch. σ is the sigmoid function. $f(\cdot)$ is the signal encoder. If the alarm is true, E_e^R should be distant from E_e^A for the alarm signals to represent real abnormal vital signs while baseline signals do not. Then the constraint should be defined as:

$$C_{true}^{(j)} = -\log(\sigma(-(f(E^A)^T f(E^R)))), \tag{7}$$

where $C_{true}^{(j)}$ is the discriminative constraint of a record with a true alarm in a mini-batch. The discriminative constraint in a mini-batch can be calculated as:

$$C = \frac{1}{N_1}\sum C_{false}^{(i)} + \frac{1}{N_2}\sum C_{true}^{(j)}. \tag{8}$$

In the above equation, C is the calculated discriminative constraint in a mini-batch. N_1 is the number of false alarm records in a mini-batch and N_2 is the number of true alarm samples.

In the training procedure, we try to minimize the discriminative constraint for each mini-batch. Meanwhile, we also use binary cross entropy loss(BCE) to ensure that the classification part of the model can correctly classify the classification vector of input signals. The binary cross entropy loss is calculated as:

Arrhythmia	Definition	Count/ratio	True alarms	False alarms
ASY	0 beats in 4s	122/17%	22	100
EBR	> 5 beats, HR < 40 bpm	89/11%	43	46
ETC	> 17 beats, HR > 140 bpm	140/17%	131	9
VTA	> 5 ventricular beats, HR > 100 bpm	341/47%	89	252
VFB	Fibrillation waves in 4 s	58/7%	6	52

Table 1. Statistics of the PhysioNet Challenge 2015 training set.

$$L_{BCE} = \frac{1}{N}\sum_{i}^{N} -y_i \log \widehat{y}_i - \left(1 - y_i\right) \log\left(1 - \widehat{y}_i\right). \quad (9)$$

In the above equation, L_{BCE} is the calculated binary cross entropy loss. N is the number of alarm records in a mini-batch. y_i is the label of the triggered alarm of a sample. $\widehat{y}_i$ is the probability of the alarm of a sample being true.

The loss function we minimize during training is the combination of discriminative constraint and the binary cross entropy loss. We formulate the loss function as the weighted sum of a binary cross entropy loss and the discriminative constraint as:

$$L = L_{BCE} + w * C, \quad (10)$$

where w is the weight.

Experiments

Datasets. The study uses publicly-available, de-identified dataset from PhysioNet[9]. The PhysioNet/Computing in Cardiology Challenge 2015 provides a dataset with 750 records for algorithm development and 500 unrevealed records. These records consist of patients' physiological signals that have been collected from four hospitals in the United States and Europe, sourced from the devices made by three major manufacturing companies of intensive care monitor devices[9]. Each record contained an alarm for one arrhythmia event and the triggered alarm was reviewed and labeled by a team of expert annotators to either true or false. Asystole (ASY), Extreme Bradycardia (EBR), Extreme Tachycardia (ETC), Ventricular Fibrillation or Flutter (VFB), or Ventricular Tachycardia (VTA) are the five alarm types in the datasets. Each record contained two leads of ECG, and at least one pulsatile waveform of either PPG or ABP. In some records, both pulsatile waveforms or a respiratory signal were present. All signals have been re-sampled to a resolution of 12 bits and had a sampling frequency of 250 Hz, therefore each record is 5 min or 5 min and 30 s long. The alarm onsets are 5 min from the beginning of each record. The exact time of the event that triggered the alarm varies somewhat from one record to another, but in order to meet the AAMI standards, the commencement of the event must be within 10 seconds of the alarm[45]. For Event 1, which is a real-time alarm suppression problem, each record is exactly 5 min long. For Event 2, each record contains an additional 30 s of signals following the time of the alarm. We focus on Event 1 in this paper. Some statistics are shown in Table 1, and more detailed statistics about this dataset could be found in Supplementary Appendix.

Pre-processing. In this paper, we focus on the real-time event and only use the first 300 s of signals for each record, which means only information prior to the alarm onset is used. Therefore, for each record we use, the event that triggered the alarm is during the last 10 s immediately prior to the alarm onset. Before we feed the signals into the model, raw signals are subjected to imputation and standardization. In the imputation part, some patients do not have the record of certain signals. Therefore, these missing signals are imputed with 0. In the standardization part, each signal is normalized to a range of 0 to 1.

Setup. In the experiments, we use fivefold cross validation. For each cross validation, one fold that is used for evaluating the model has 150 records and the remaining fourfold that are used to train the model has 600 records. In the end, all evaluation results are averaged.

We use 4 parallel CNN blocks in the signal encoder. These 4 CNN blocks have different filter sizes, which are 50, 100, 200, and 400 respectively. Each CNN block has two convolutional layers with the same filter size. The first convolutional layer is composed of 64 filters with a stride of 5, a Batch Normalization layer[46] and a Rectified Linear Unit(ReLU) layer[47]. The second convolutional layer has the same hyperparameters as the first layer except for the number of input channels. Convolutional layers are followed by a Global Max Pooling layer to aggregate high-level discriminative features and flatten features across channels. The output sizes of the rule-based embedding layer and the alarm type embedding layer are both set as 64.

Our model was trained with a maximum of 1000 epochs and a mini-batch size of 256. We use Adam optimizer[48] to minimize the BCE loss and the discriminative constraint with learning rate of 0.001. To prevent the over-fitting problem, we use L2 regularization with a value of 0.0005. The dropout rate before each CNN block is set as 0.8. To overcome the problem of imbalanced classes, a weight of 1.5 was added to the positive samples in the BCE loss function as the number of negative alarm records is roughly 1.5 times than the number of positive alarm records. The weight of the discriminative constraint is set to 1.5 during training. The rule-based method

	MLP			FCN			ResNet			RB2			ML1			RB1			EDGCN			Ours		
Arrhythmia	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**	**TPR (%)**	**TNR (%)**	**Score**
ASY	17	82	53.95	61	78	64.48	83	64	61.68	75	94	82.46	75	90	78.95	100	97	97.06	78	82	73.68	100	97	97.42
EBR	92	9	37.61	67	64	42.28	87	41	49.57	96	63	72.06	92	84	77.78	100	74	84.38	100	71	82.47	100	72	83.51
ETC	100	0	87.80	100	0	95.50	100	0	95.50	100	80	98.63	100	80	98.63	97	100	87.65	100	60	98.20	97	100	87.80
VTA	21	87	44.40	15	85	41.18	56	69	48.55	71	95	73.26	89	90	75.10	82	84	72.73	90	80	75.91	91	83	78.75
VFB	0	96	50.00	56	98	71.62	78	90	77.27	83	94	84.09	100	71	75.00	83	91	81.82	100	92	93.10	78	94	80.30
Real-time	66	77	51.54	66	80	52.79	83	66	59.11	89	**91**	79.02	94	82	79.44	93	87	81.62	96	80	80.68	**96**	86	**84.47**

Table 2. Performance comparison on the test set of PhysioNet Challenge 2015. Best performing values in each performance metric are given in bold.

Components								
Basic	**Rule**	**Constraint**	**ASY**	**EBR**	**ETA**	**VTA**	**VFB**	**Real-time**
✓			74.85	40.43	95.50	51.69	52.13	60.35
✓		✓	62.30	46.62	**96.40**	56.63	55.56	60.69
✓	✓		96.13	80.20	95.50	69.93	**80.30**	78.90
✓	✓	✓	**97.42**	**83.51**	87.80	**78.75**	**80.30**	**84.47**

Table 3. Quantitative results of ablation study on test set. Best performing challenge scores in each column are given in bold.

with which we choose to combine our model is proposed by Ref.[10]. Our model was implemented using Python programming language and PyTorch deep learning library[49].

Compared methods. We evaluate our proposed model on the hidden test set and present comparisons to the existing deep learning methods commonly used for time-series classification and the top three methods of the real-time event listed on the challenge website. The compared methods are summarized as follows. (1) *MLP* We apply the multi-layer perceptron as a feature extractor of the input waveform and then use a dense layer to classify. (2) *FCN* We use a fully-connected convolutional network as the feature extractor of the input waveform. (3) *ResNet* We use ResNet as the feature extractor of the input waveform (4) *RB1* A rule-based method based on descriptive statistics and Fourier and Hilbert transforms[10]. (5) *ML1* A machine learning model that uses a combination of logical and SVM-based techniques[13]. (6) *RB2* A rule-based method that detects QRS and analyses the signal quality and then applies a different rule to each alarm type[21]. (7) *EDGCN* A deep group convolutional neural network proposed by Yu et al.[31].

Results. The evaluation metrics for false alarm reduction are classification accuracy (ACC), true positive rate (TPR) and true negative rate (TNR). The PhysioNet Challenge 2015[9] also provides an official scoring mechanism for evaluating. It is defined as $score = (TP + TN)/(TP + TN + FP + 5 \times FN)$, where TP is true positives, FP is false positives, FN is false negatives, and TN is true negatives. The Challenge Score focuses more on the TPR value, since mistakenly classifying a true alarm as false results in significantly more severe consequences.

During training, we evaluate our model on the fivefold validation set. The average of the challenge score from the fivefold result is 87.00 with a standard deviation of 4.84. The detailed results are included in Supplementary Appendix.

Table 2 shows the detailed performance comparisons with the compared methods on the hidden test set. The performance of each alarm type is denoted as N/A since there are no such statistics in the paper. The higher the ACC, TPR, TNR and score, the better the performance. It is observed that our proposed contrastive learning model out-performs other baseline methods in the hidden test set. Note that the baseline methods in Table 2, including FCN and ResNet, are trained *without* contrastive learning. In Supplementary Appendix, we also present the performance comparison of FCN, ResNet and our CNN as different "backbone" encoders in our proposed contrastive learning framework.

Comparison with other models, especially with rule-based models, shows that it is difficult for common deep learning models to achieve high performance. There are many reasons for the poor performance of these models. First of all, many possible reasons can lead to false alarms including noise, patient manipulation or movement, mis-configuring, staff manipulation and leads falling off or mis-identification of signals. Second, the very long sequence length and imbalanced classes in the given datasets are big challenges for deep learning models. In addition, the number of labeled samples is crucial for classification using deep learning models while there are only 750 samples in the given training set which extremely limits the performance of these models. In our proposed model, we use 4 different kernel sizes from 50 to 400 in the CNN layers, which help alleviate the problem that the sequence length is too long for larger kernel sizes and can more effectively detect abnormal signals.

Ablation study. Our method has three components: signal encoder, discriminative constraint and rule-based embedding. We implement an ablation study by analyzing the quantitative results on the hidden test set as shown in Table 3. It can be observed that the more components we use the higher the performance is. Using only the signal encoder, the real-time event score can only achieve a challenge score of 60.35, which is only slightly better than the performance of ResNet, a common deep learning model with the best score in Table 2.

We observe that using discriminative constraints during training leads to improved performance in the overall real-time event score. Furthermore, it improves the performance of all alarm types except asystole. Augmenting the CNN-based signal encoder with a rule-based embedding can improve the performance significantly. The rule-based model contains information about the descriptive statistics and fuzzy logic derived based on domain-knowledge of the waveform for each alarm type, and thus enhances the signal encoder's ability to more accurately distinguish between true vs. false alarms.

The discriminative constraints utilize the idea of contrastive learning to address the problem of over-fitting and imbalanced classes. Our results indicate that adding discriminative constraints to the combined a model of signal encoder and rule-based embeddings leads to the best performance in the Challenge score, with significant performance improvement from 78.90 to 84.47. Our results demonstrate the effectiveness of the proposed loss function.

Conclusions and future work

False arrhythmia alarm reduction in ICUs is a challenging task for deep learning due to the very long sequence length of physiological signals, imbalanced classes and a limited number of labeled records. In this paper, we present a contrastive learning framework based on Siamese Network for false alarm reduction. During training, we use discriminative constraints to improve the feature extraction of signals. Furthermore, we augment our proposed model with a rule-based technique by using embeddings from the outputs of the rule-based method to regularize our deep learning model for label-efficient representation learning. Results show that the proposed method detected 86% of false alarms in the test set. The detection rate of true alarms was 96%. Using the official given scoring equation of the challenge, we achieved a score of 84.47 in the real-time event, outperforming other methods in the same task in the 2015 PhysioNet Challenge for the false arrhythmia alarm reduction.

Since the supervised information is too small to train a common deep learning model, making better use of the input data may be a promising direction. In future work, we will consider using self-supervised learning techniques to expand the scale of training data since we can set multiple pseudo-labels according to the downstream task to pretrain the model. Another potential direction is to pretrain an unsupervised learning model, such as BERT for the unlabeled time series data and then finetune the model in the downstream task.

Received: 17 August 2021; Accepted: 7 February 2022
Published online: 18 March 2022

References

1. Drew, B. J. *et al.* Practice standards for electrocardiographic monitoring in hospital settings: An American Heart Association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young: Endorsed by the international society of computerized electrocardiology and the American Association of Critical-care Nurses. *Circulation* **110**, 2721–2746 (2004).
2. Drew, B. J. *et al.* Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PLoS ONE* **9**, e110274 (2014).
3. Parthasarathy, S. & Tobin, M. Sleep in the intensive care unit. *Intens. Care Med.* **30**, 197 (2004).
4. Johnson, A. N. Neonatal response to control of noise inside the incubator. *Pediatr. Nurs.* **27**, 600 (2001).
5. Morrison, W. E., Haas, E. C., Shaffner, D. H., Garrett, E. S. & Fackler, J. C. Noise, stress, and annoyance in a pediatric intensive care unit. *Crit. Care Med.* **31**, 113–119 (2003).
6. Berg, S. Impact of reduced reverberation time on sound-induced arousals during sleep. *Sleep* **24**, 289–292 (2001).
7. Tsien, C. L. & Fackler, J. C. Poor prognosis for existing monitors in the intensive care unit. *Crit. Care Med.* **25**, 614–619 (1997).
8. Cvach, M. Monitor alarm fatigue: An integrative review. *Biomed. Instrum. Technol.* **46**, 268–277 (2012).
9. Clifford, G. D. *et al.* The physionet/computing in cardiology challenge 2015: reducing false arrhythmia alarms in the icu. In *2015 Computing in Cardiology Conference (CinC)*, 273–276 (IEEE, 2015).
10. Plesinger, F., Klimes, P., Halamek, J. & Jurak, P. Taming of the monitors: reducing false alarms in intensive care units. *Physiological Measurements, 37:1313–1325* (2016).
11. Peng, W. K., Ng, T.-T. & Loh, T. P. Machine learning assistive rapid, label-free molecular phenotyping of blood with two-dimensional nmr correlational spectroscopy. *Commun. Biol.* **3**, 535. https://doi.org/10.1038/s42003-020-01262-z (2020).
12. Lau, W.-H., White, N. J., Yeo, T.-W., Gruen, R. L. & Pervushin, K. Tracking oxidation-induced alterations in fibrin clot formation by nmr-based methods. *Sci. Rep.* **11**, 15691. https://doi.org/10.1038/s41598-021-94401-3 (2021).
13. Kalidas, V. & Tamil, L. S. Cardiac arrhythmia classification using multi-modal signal analysis. *Physiological Measurements*, **37**(8), 1253–1272 (2016).
14. LeCun, Y. *et al.* Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, 396–404 (1990).
15. Chopra, S., Hadsell, R. & LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 539–546 (IEEE, 2005).
16. Daluwatte, C. *et al.* Heartbeat fusion algorithm to reduce false alarms for arrhythmias. In *2015 Computing in Cardiology Conference (CinC)*, 745–748 (IEEE, 2015).
17. Ansari, S., Belle, A. & Najarian, K. Multi-modal integrated approach towards reducing false arrhythmia alarms during continuous patient monitoring: The physionet challenge 2015. In *2015 Computing in Cardiology Conference (CinC)*, 1181–1184 (IEEE, 2015).
18. Tsimenidis, C. & Murray, A. Reliability of clinical alarm detection in intensive care units. In *2015 Computing in Cardiology Conference (CinC)*, 1185–1188 (IEEE, 2015).
19. He, R. *et al.* Reducing false arrhythmia alarms in the icu using novel signal quality indices assessment method. In *2015 Computing in Cardiology Conference (CinC)*, 1189–1192 (IEEE, 2015).

20. Fallet, S., Yazdani, S. & Vesin, J.-M. A multimodal approach to reduce false arrhythmia alarms in the intensive care unit. In *2015 Computing in Cardiology Conference (CinC)*, 277–280 (IEEE, 2015).
21. Couto, P., Ramalho, R. & Rodrigues, R. Suppression of false arrhythmia alarms using ecg and pulsatile waveforms. In *2015 Computing in Cardiology Conference (CinC)*, 749–752 (IEEE, 2015).
22. Antink, C. H., Leonhardt, S. & Walter, M. Reducing false alarms in the icu by quantifying self-similarity of multimodal biosignals. *Physiol. Meas.* **37**, 1233 (2016).
23. Eerikäinen, L. M., Vanschoren, J., Rooijakkers, M. J., Vullings, R. & Aarts, R. M. Decreasing the false alarm rate of arrhythmias in intensive care using a machine learning approach. In *2015 Computing in Cardiology Conference (CinC)*, 293–296 (IEEE, 2015).
24. Caballero, M. & Mirsky, G. M. Reduction of false cardiac arrhythmia alarms through the use of machine learning techniques. In *2015 Computing in Cardiology Conference (CinC)*, 1169–1172 (IEEE, 2015).
25. Afghah, F., Razi, A. & Najarian, K. A shapley value solution to game theoretic-based feature reduction in false alarm detection. Preprint at http://arxiv.org/abs/1512.01680 (2015).
26. Zaeri-Amirani, M., Afghah, F. & Mousavi, S. A feature selection method based on shapley value to false alarm reduction in icus a genetic-algorithm approach. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 319–323 (IEEE, 2018).
27. Au-Yeung, W.-T.M., Sahani, A. K., Isselbacher, E. M. & Armoundas, A. A. Reduction of false alarms in the intensive care unit using an optimized machine learning based approach. *NPJ Dig. Med.* **2**, 1–5 (2019).
28. Lehman, E. P., Krishnan, R. G., Zhao, X., Mark, R. G. & Lehman, L. H. Representation learning approaches to detect false arrhythmia alarms from ecg dynamics. In *Machine Learning for Healthcare Conference*, 571–586 (PMLR, 2018).
29. Hooman, O. M., Al-Rifaie, M. M. & Nicolaou, M. A. Deep neuroevolution: Training deep neural networks for false alarm detection in intensive care units. In *2018 26th European Signal Processing Conference (EUSIPCO)*, 1157–1161 (IEEE, 2018).
30. Mousavi, S., Fotoohinasab, A. & Afghah, F. Single-modal and multi-modal false arrhythmia alarm reduction using attention-based convolutional and recurrent neural networks. *PLoS ONE* **15**, e0226990 (2020).
31. Yu, Q. *et al.* Intensive care unit false alarm identification based on convolution neural network. *IEEE Access* **9**, 81841–81854. https://doi.org/10.1109/ACCESS.2021.3086862 (2021).
32. Zihlmann, M., Perekrestenko, D. & Tschannen, M. Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, 1–4 (IEEE, 2017).
33. Clifford, G. D. *et al.* Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, 1–4 (IEEE, 2017).
34. Hong, S. *et al.* Encase: An ensemble classifier for ecg classification using expert features and deep neural networks. In *2017 Computing in Cardiology (CinC)*, 1–4 (IEEE, 2017).
35. Hyvarinen, A. & Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Adv. Neural Inf. Process. Syst.* **29**, 3765–3773 (2016).
36. Kiyasseh, D., Zhu, T. & Clifton, D. A. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615 (PMLR, 2021).
37. Pei, W., Tax, D. M. & van der Maaten, L. Modeling time series similarity with siamese recurrent networks. Preprint at http://arxiv.org/abs/1603.04713 (2016).
38. Wu, X., Kimura, A., Iwana, B. K., Uchida, S. & Kashino, K. Deep dynamic time warping: End-to-end local representation learning for online signature verification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1103–1110 (IEEE, 2019).
39. Szegedy, C. *et al.* Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1–9 (2015).
40. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings* (eds. Bengio, Y. & LeCun, Y.) (2015).
41. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 3104–3112 (2014).
42. Gamboa, J. C. B. Deep learning for time-series analysis. Preprint at http://arxiv.org/abs/1701.01887 (2017).
43. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440 (2015).
44. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
45. for the Advancement of Medical Instrumentation, A. *et al.* Cardiac monitors, heart rate meters, and alarms. *American National Standard (ANSI/AAMI EC13: 2002) Arlington, VA*, 1–87 (2002).
46. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Preprint at http://arxiv.org/abs/1502.03167 (2015).
47. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *ICML*, 807–814 (2010).
48. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2014).
49. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 8026–8037 (2019).

Acknowledgements

This work is in part funded by the NSFC, China under Grants 61902309; in part by the Fundamental Research Funds for the Central Universities, China (xxj022019003); in part by China Postdoctoral Science Foundation (2020M683496); and in part by the National Postdoctoral Innovative Talents Support Program, China (BX20190273); and in part funded by the NIH Grants R01EB030362 and R01 EB017205.

Author contributions

L.H.L. and R.G.M. conceived the study. G.Z., L.H.L., Y.Z., J.L., and G.S. designed the study. Y.Z., G.Z., and B.M. conducted the experiments. Y.Z., G.Z. and L.H.L. wrote the manuscript. All authors reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07761-9.

Correspondence and requests for materials should be addressed to G.Z. or L.H.L.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.