Representation Learning Approaches to Detect False Arrhythmia Alarms from ECG Waveforms

Eric P. Lehman * LEHMAN.E@HUSKY.NEU.EDU
College of Computer and Information Science
Northeastern University, Boston, MA

Rahul G. Krishnan RAHULGK@MIT.EDU
CSAIL & Institute for Medical Engineering & Science
Massachusetts Institute of Technology, Cambridge, MA

Xiaopeng Zhao † XZHAO9@UTK.EDU
Department of Mechanical, Aerospace, and Biomedical Engineering
University of Tennessee, Knoxville, TN

Roger G. Mark RGMARK@MIT.EDU
Institute for Medical Engineering & Science
Massachusetts Institute of Technology, Cambridge, MA

Liwei H. Lehman LILEHMAN@MIT.EDU
Institute for Medical Engineering & Science
Massachusetts Institute of Technology, Cambridge, MA

Abstract

The high rate of intensive care unit false arrhythmia alarms can lead to disruption of care and slow response time due to desensitization of clinical staff. We study the use of machine learning models to detect false ventricular tachycardia (v-tach) alarms using ECG waveform recordings. We propose using a Supervised Denoising Autoencoder (SDAE) to detect false alarms using a low-dimensional representation of ECG dynamics learned by minimizing a combined reconstruction and classification loss. We evaluate our algorithms on the PhysioNet Challenge 2015 dataset, containing over 500 records (over 300 training and 200 testing) with v-tach alarms. Our results indicate that using the SDAE on Fast Fourier Transformed (FFT) ECG at a beat-by-beat level outperforms several competitive baselines on the task of v-tach false alarm classification. We show that it is important to exploit the underlying known physiological structure using beat-by-beat frequency distribution from multiple cardiac cycles of the ECG waveforms to obtain competitive results and improvement over previous entries from the 2015 PhysioNet Challenge.

1. Introduction

Intensive care units (ICU) false arrhythmia alarm rates have been reported to be as high as 88.8%, and can lead to disruption of care and slow response time (Drew et al., 2014). Detecting and suppressing false arrhythmia alarms could potentially have high impact on

* Research performed while working as a research intern at MIT.
† Research performed while at MIT.

© 2018 E.P. Lehman, R.G. Krishnan, X. Zhao, R.G. Mark & L.H. Lehman.

the quality of patient care, reducing the chance of a life-threatening true alarm missed due to staff desensitization. Here, we investigate representation learning approaches to finding discriminative features of ECG dynamics for false alarm reduction. We focus on the problem of detecting false alarms for one of the life-threatening arrhythmias, ventricular tachycardia (v-tach), defined as five or more ventricular beats with a heart rate higher than 100 bpm (Clifford et al., 2015). Among all the life-threatening arrhythmia alarm types, false v-tach alarms have proven to be the hardest to detect (Clifford et al., 2015), and remain an open challenge.

This work investigates the utility of both linear and non-linear embeddings of ECG for detecting v-tach false alarms. We present a joint supervised generative technique, *Supervised Denoising Autoencoders (SDAE)*, to classify ventricular tachycardia alarms using non-linear embeddings of ECG dynamics. The model is learned using a combination of a discriminative and generative loss. As a comparison, we investigate linear-embedding techniques that use Principal Component Analysis (PCA) in combination with a non-linear classifier multi-layer Perceptrons (MLP).

Furthermore, we explore feature transformations that utilize the underlying known physiological structure within the ECG to enable learning under the constraints of limited labeled data. To this end, we propose an approach that utilizes FFT-transformed ECG at a (heart) beat-by-beat basis, and compare with several baseline approaches, using a wide range of time and frequency domain ECG features.

Technical Significance Application of machine learning techniques in arrhythmia analysis has had limited success, partly due to sparse availability of labeled data (Clifford et al., 2016, 2017). The best performing approaches for false v-tach alarm detection in the 2015 PhysioNet Challenge rely on a combination of expert-defined rule-based reasoning and simple machine-learning models (Kalidas and Tamil, 2016; Plesinger et al., 2016; Clifford et al., 2016).

In the current study, we explore a supervised generative model, SDAE, and feature transformation approaches to enable false v-tach alarm classification in a low labeled data setting. Our approach combines non-linear embeddings from SDAE, with ECG feature transformation from FFT at a beat-by-beat level. Tests on a real-world ICU dataset from 2015 PhysioNet Challenge containing over 500 records indicate that the proposed approach leads to improved performance over several baselines, including previous entries from the 2015 PhysioNet Challenge, and enables scalable learning that performs well even when labels are scarce.

Clinical Relevance In modern ICUs, where critically-ill patients are closely monitored, as many as 187 audible alarms have been reported per ICU bed per day (Drew et al., 2014), corresponding to an alarm every 7 to 8 minutes for each patient. Arrhythmia alarms contribute to approximately 45% of the overall ICU alarms, and have a high false alarm rate of 88%, which can lead to lower patient care quality (Drew et al., 2014; Clifford et al., 2015). In this work, we focus on detecting false arrhythmia alarms for v-tach. Detecting and minimizing false v-tach alarms could potentially reduce the chance of a life-threatening true alarm missed due to staff desensitization.

2. Methods

2.1. Approach Overview

Figure 1 illustrates our overall approach for the ECG data processing pipeline to derive FFT-transformed ECG features on a beat-by-beat basis as input to the SDAE for false alarm classification.

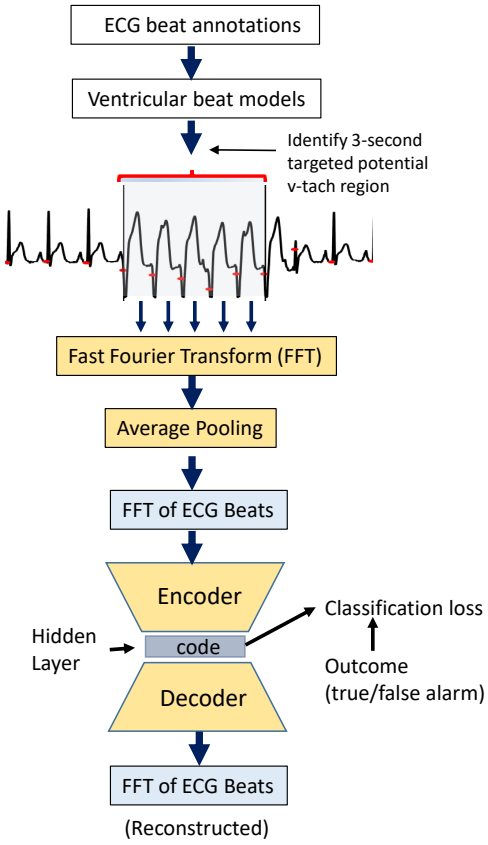


Figure 1: Approach Overview: ECG data processing pipeline and a SDAE model with FFT-transformed ECG beats from a single-lead ECG as input for v-tach false alarm classification.

- We use the MIT-BIH database, with annotated beat labels, to build a ventricular beat (v-beat) classifier that classifies whether the FFT transform of a beat is a ventricular beat or a non-ventricular beats.
- We first apply a peak-detection algorithm to multi-channel ECG signals in the PhysioNet 2015 Challenge dataset. For each beat in the PhysioNet Challenge dataset, we apply the FFT transform to derive the frequency content on a per-beat basis.
- On the FFT transformed data in the PhysioNet Challenge (last 25 seconds before alarm onset), we apply the v-beat classifier to estimate the probability of a ventricular beat. We aim to obtain a small region in the ECG recordings to focus on when

predicting alarm outcomes. This is done by identifying a 3-second target interval with the highest v-beat probability (averaged over consecutive beats) among beats where the heart rate exceeds 100bpm.

- From the targeted region, we extract the following features: (i) FFT of beats: obtained by performing an FFT transform on each ECG beat from all ECG beats in the three-second target interval prior to the alarm onset. Each beat is represented as a 41-dimensional FFT-transformed feature vector, and (ii) FFT transform on the entire 3-second targeted ECG segment. Figure 1 illustrates the data processing and FFT-transformation pipeline.

- Finally, we use the data from the previous step to classify the probability of a true or false alarm. In Figure 1, SDAE is used to learn a low-dimensional representation of the FFT-transformed features extracted from the target intervals of all training records.

2.2. Datasets

2.2.1. MIT-BIH Arrhythmia Dataset

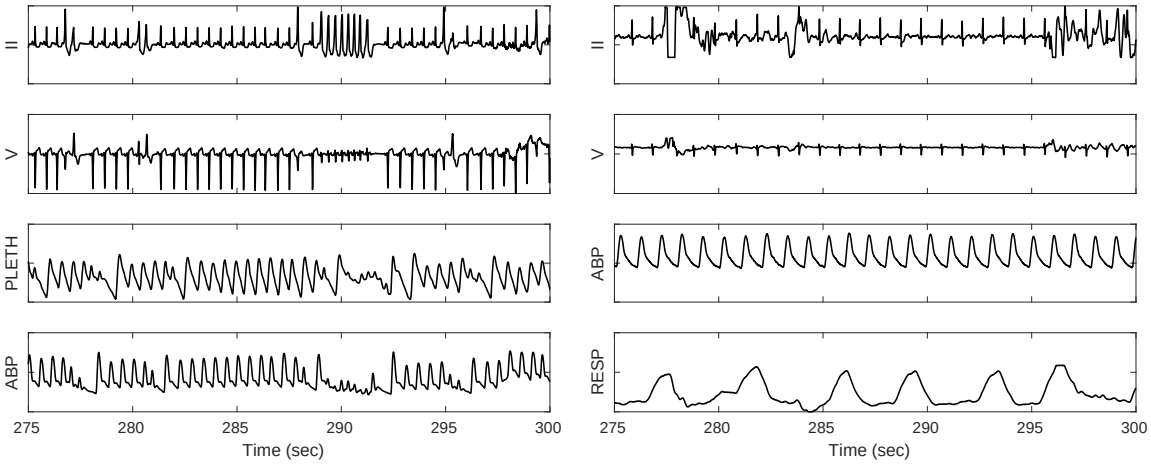
We used the MIT-BIH Arrhythmia Database (Moody and Mark, 2001; Goldberger et al., 2000) to train a ventricular beat identification model. The MIT-BIH dataset contains 48 half-hour excerpts of two-channel ambulatory ECG recordings (360 Hz), obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. All data are re-sampled to 250 Hz to match the sampling frequency of the PhysioNet Challenge dataset. We extracted a total of 107,129 annotated beats from the MIT-BIH database (lead II), among which 7,104 were ventricular beats, and the remaining 100,025 were non-ventricular beats.

2.2.2. PhysioNet Challenge 2015 Dataset

The 2015 PhysioNet Challenge event (Clifford et al., 2015, 2016) focused on five types of life-threatening arrhythmias, including ventricular tachycardia (v-tach), asystole, extreme bradycardia, extreme tachycardia, and ventricular fibrillation/flutter. The goal of the challenge was to reduce the number of false alarms, while avoiding suppression of true alarms.

The PhysioNet 2015 Challenge data (Clifford et al., 2015; Goldberger et al., 2000) contains a total of 1250 records (750 train, 500 hidden test), each containing one of the five life-threatening arrhythmia alarms. Each record contains 3 to 4 channels, including two channels of ECG and one or two of the following channels: arterial blood pressure (ABP), respiration, or photoplethysmography (PPG). Each record contains 5-minute recordings of multi-channel physiological waveforms (250 Hz) immediately prior to the alarm onset. A subset of the records also contain 10-second recordings of the waveforms after the alarm onset. The Challenge consists of two events: (1) real-time classification using only data up to the alarm onset; (2) retrospective analysis in which the contestants are allowed to use the 10-second data after the alarm onset for classification. Here, we focus strictly on the real-time setting where only data prior to the alarm onset time is used. Figure 2 shows an example of true and false v-tach alarm each from the PhysioNet Challenge training dataset.

Among the 1250 records, 562 records contain v-tach alarms, and overall more than 75% of these v-tach alarms are false. The training and test sets of our study are extracted from the 562 records (train N=341, test N=221) corresponding to the v-tach alarms from the PhysioNet Challenge 2015 dataset. Among the 341 v-tach alarms in the training records, 73.3% are false (91 true alarms, and 250 false alarms). Among the 221 v-tach alarms in the hidden test set 79.6% are false (45 true alarm, 176 false alarms). For training, we used 337 records from the training set containing either lead II (N=331) or V (N=312) ECGs. The hidden test set contains all 221 records with v-tach alarms. We extracted the last 25-seconds ECG segments from each record prior to the alarm onset for analysis.



(a) True alarm (Record v254s) (b) False alarm (Record v360s)

Figure 2: Example true vs. false v-tach alarms (from PhysioNet Challenge 2015). Each plot shows data in the 25-second interval immediately prior to the v-tach alarm onset (at the end of the interval).

2.3. Data Preprocessing

2.3.1. Selection of Targeted Segments for Feature Extraction

Ventricular Beat Identification Model

We used the a peak detection algorithm developed by Johnson et al. (2015) and the Pan Tompkins algorithm (Pan and Tompkin, 1985) to identify each ECG beat. A model for ventricular beat identification was trained using the MIT-BIH database. We then applied the v-beat classifier learned from the MIT-BIH v-beat classification model to identify a potential v-tach episode as a 3-second ECG segment from each record in the 2015 PhysioNet Challenge. We detail the algorithm used to identify the 3-second targeted interval in Appendix A.

2.3.2. Data Transformation: from Time-Domain to Frequency-Domain

Due to heart rate variability, durations of RR-intervals (i.e. intervals between R-peaks) are different. It is a common approach to nondimentionalize time so that all the beats are

aligned. Here, we adopt a different approach using frequency distribution of the ECG beat to avoid rescaling the time axis. FFT is applied to every beat within the selected 3-second range in order to obtain the frequency spectrum between 0.1Hz and 40Hz. The beat-level FFT of all beats from a 3-second targeted window of each record are then averaged. Since we are interested in the frequency distribution, the averaged FFT transform is scaled to be between 0 and 1 to form the final feature vector for each record.

2.4. Model

Multi-layer Perceptron (MLP)

As baselines, we predict the probability of a true alarm using (a) the Fast Fourier Transform (FFT) of the ECG waveforms and (b) the projection of the FFT data onto its principal components. In both cases the features are passed into a three-layer MLP with Rectified Linear Units as non-linearities between intermediate layers. The model parameters were selected using a grid search performed over the layer size (16, 32, 64, 128, 256), and dropout (0.2, 0.4, 0.5).

PCA/MLP

The structure of the neural network used for the PCA output is identical to the previously mentioned MLP architecture. A grid search over neural network parameters was performed, but there was no statistically significant difference between architectures. However, the dimensionality of the PCA showed significant impact on performance; thus a grid search PCA dimensionality was performed (5, 10, 15, 25), and the optimal size was selected. PCA can be viewed as a linear equivalent of the autoencoder. Specifically, when using a linear activation function with a single layer, the two can be considered identical (Bourlard and Kamp, 1988).

Supervised Denoising Autoencoder (SDAE)

To study how *generative* approaches affect this domain, we use a supervised denoising autoencoder on FFT transformed ECG waveforms. Denoising autoencoders (Vincent et al., 2010, 2008; Lovedeep and Gondara, 2016; Bengio et al., 2013) are a class of autoencoders where the goal of the model is to reconstruct a noisy (typically lower dimensional) transformation of the input. Here, we also use the denoising autoencoder to predict the label of whether or not the beats represented in the input corresponds to a true alarm. The model's input is comprised of 2 different ECG channels. Figure 3 shows the SDAE architecture. The parameters and architecture of the model are discussed in more detail in Appendix C.

The SDAE is trained using the ADAM optimizer (Kingma and Ba, 2014) with two loss functions: (1) a reconstruction loss (in our case the mean-squared error) that encourages the model to reconstruct the input and (2) a prediction loss that encourages the intermediate hidden state of the autoencoder to predict the likelihood of a true alarm.

The model parameters were selected using a grid search performed over the layer size (16, 32, 64, 128, 256), dropout (0.2, 0.4, 0.5), and variance of the noise added to the hidden state (0, 0.001, 0.01). For each one, a ten-fold-cross-validation was used to compute the validation accuracy, which in turn was used to determine the best model parameters.

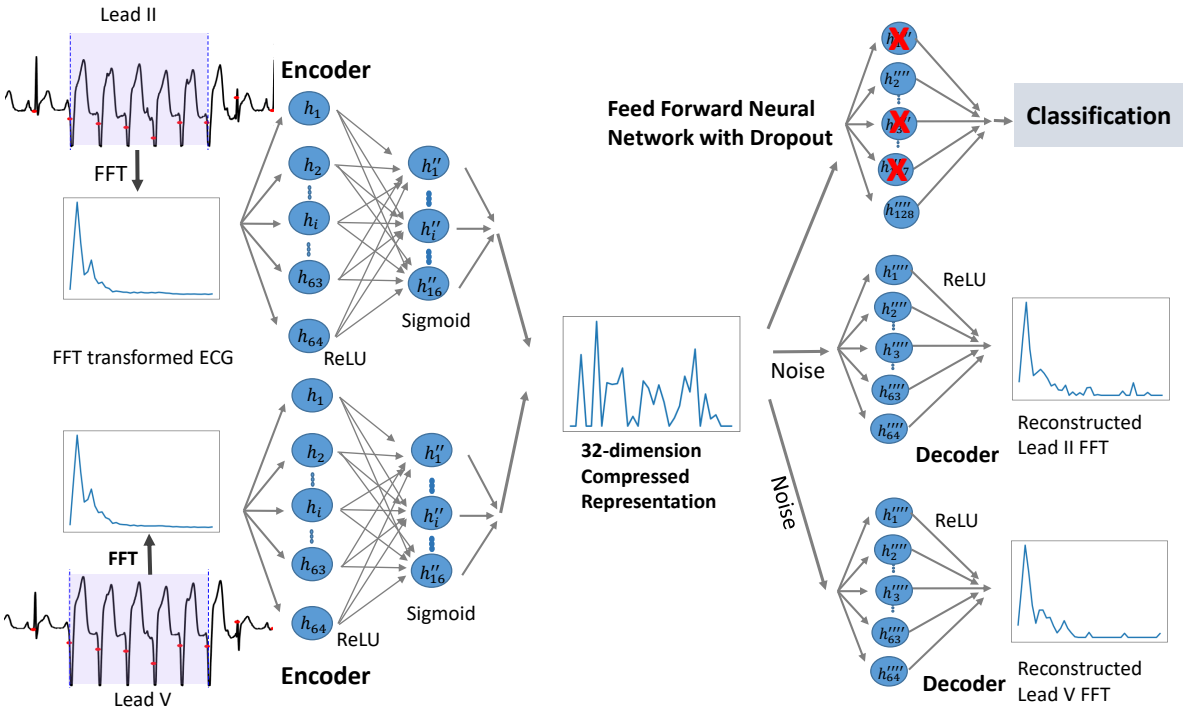


Figure 3: SDAE model diagram for v-tach false alarm detection using two leads of ECG. Each of these input ECG channels is run through two perception layers and then combined to form a 32-dimension compressed hidden-layer representation. The data from individual leads are then reconstructed through separate decoding layers, while an outcome and classification loss are gathered using the combined layer.

2.5. Experimental Setting

Representation Learning from ECG waveforms: we explore the following types of ECG features i) default 10-second raw ECG waveform intervals (extracted from 10-seconds prior to the alarm onset) ii) targeted 3-second raw ECG waveform segments identified using the MIT-BIH ventricular beat model and iii) their respective spectral content, and iv) the beat-level spectral representation from the targeted region.

For each feature, we explore the following models for representation learning and prediction modeling: i) Logistic Regression (LR), ii) Feed-forward neural network on original feature space (without representation learning), iii) PCA-embedding with feedforward neural network, and iv) joint model of Autoencoder and feedforward neural network.

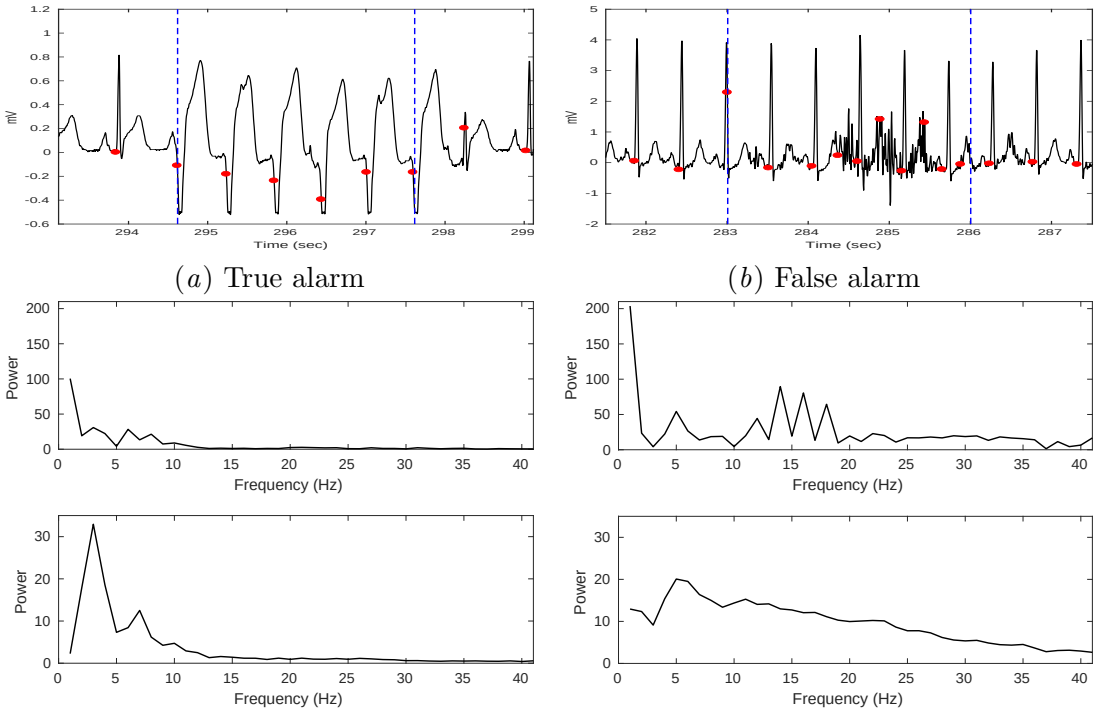
2.6. Evaluation

The training data was split 10 different times, each resulting in 70% used for training and 30% used for validation. We report the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) and Area under the Curve (AUC) from the receiver operating characteristic (ROC) curve. Confidence intervals (CI) for area under the receiver operating curves (AUCs) were based on the method described in DeLong et al.

(1988). For our final model, our primary performance metric when comparing with previous approaches is the PhysioNet Challenge 2015 scores, defined as $(TP + TN)/(TP + TN + FP + 5 * FN)$, a function of the following variables: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

3. Evaluation

3.1. PhysioNet Challenge Dataset Statistics/Characteristics



(a) True alarm

(b) False alarm

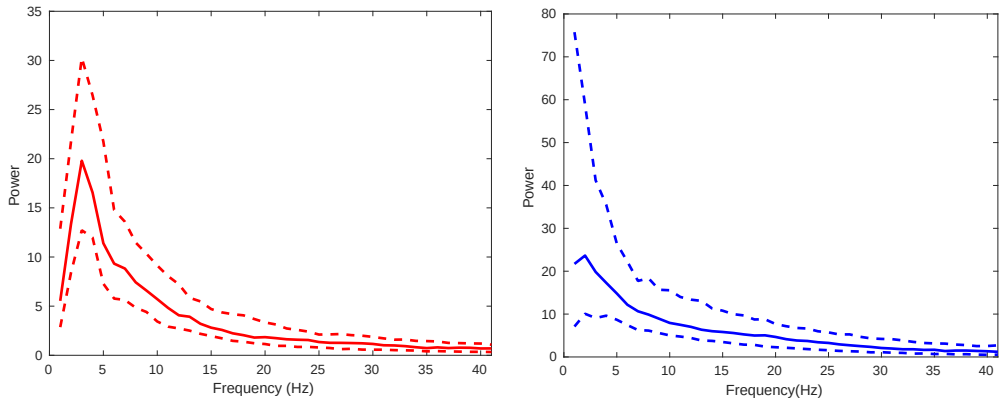
(c) Top: FFT of 3-sec targeted interval from (a). Bottom: average FFT of beats from the 3-sec interval from (a).

(d) Top: FFT of 3-sec targeted interval from (b). Bottom: average FFT of beats from the 3-sec interval from (b).

Figure 4: Example ECG segments which triggered true (a) and false (b) v-tach alarms. Two vertical blue dashed lines demarcate the targeted 3-second segment identified by our algorithm. (c) and (d) show FFT from an entire 3-sec targeted segment (top) vs. average FFT from individual beats in the 3-second targeted interval (bottom).

Figure 4 shows example plots of ECG waveform segments and their respective beat-level FFT-transformed features (averaged over 3-seconds) from a true and false v-tach alarm records from the PhysioNet Challenge 2015 training set.

Figure 5 compares the spectral content of ECG beats extracted from the targeted region of the true versus false alarm records from the PhysioNet Challenge training dataset (lead II) at the population-level. Population median in each frequency bin is plotted as a solid line, with the interquartile range (IQR) as dashed lines. Note that the power spectral density (PSD) of the true alarms peaks at the 4 Hz location. It is known that peaks in the PSD at 1,



(*a*) True alarm (FFT) (*b*) False alarm (FFT)

Figure 5: FFT of beats (lead II) - population median and IQR.

4, 7 and 10 Hz, in a normal ECG (with QRS width 80-100 msec), correspond approximately to the heart rate (at 60 bpm), T wave, P wave, and the QRS complex respectively (Clifford, 2006). The differences in the PSD content of the true and false alarm may be due to the ventricular beat's wider QRS complexes.

3.2. Classification Using Single-Lead ECG Waveform

Table 1 compares the performance of various representation learning techniques and classifiers when using different feature representations of single-lead ECG. In particular, our goal is to characterize and compare the performance of classifiers built on linear (PCA) vs. non-linear (autoencoder-based) representations of various time and frequency domain features of ECG signals. We experiment with (i) the raw ECG waveform from targeted 3-second segments, (ii) frequency content from targeted 3-second ECG segments, and (iii) averaged beat-level frequency content from targeted ECG segments. Additionally, we also evaluated the performance of classifiers that predict false alarms using the 10-second waveform (immediately prior to the alarm onset); the performance (AUC) of all classifiers (in Table 1) were in the 0.50's to 0.60's range, significantly worse than using the 3-second targeted waveform.

We find that FFT-transforming the ECG signal prior to predicting false alarms performs significantly better than using the raw ECG waveforms directly. Furthermore, we find that learning a non-linear representation of the ECG frequency content using SDAE achieves even better performance, with an AUC of 0.87 (0.80, 0.94) and 0.86 (0.77, 0.93) using lead II and V ECG channels respectively, than other baselines that include LR, MLP and PCA/MLP. Using FFT-transformed **beat-level** representations from single lead ECG, we note that MLP and PCA-MLP performed better than when FFT features from entire waveform segments were used (Table 1). Table 1 demonstrates that exploiting beat-level information in ECG dynamics generally leads to a boost in performance results.

3.3. Classification with Two-Channels of ECG Waveforms

Table 2 compares the performance of different classifiers using two-lead ECGs. We list the best known results from the 2015 PhysioNet Challenge as a comparison. The first-place

Table 1: Single lead ECG. Lead II (N=214), V (N=203). AUC (95% CI) shown.

			LR	MLP	PCA+MLP	SDAE
1	3-sec waveform	II	0.54 (0.44, 0.64)	0.77 (0.67, 0.86)	0.83 (0.75, 0.92)	0.77 (0.68, 0.86)
2	3-sec waveform	V	0.57 (0.47, 0.67)	0.74 (0.65, 0.84)	0.78 (0.69, 0.87)	0.76 (0.67, 0.85)
3	FFT 3-sec interval	II	0.70 (0.60, 0.80)	0.84 (0.76, 0.92)	0.85 (0.77, 0.93)	0.87 (0.80, 0.94)
4	FFT 3-sec interval	V	0.73 (0.63, 0.82)	0.81 (0.73, 0.90)	0.85 (0.77, 0.93)	0.86 (0.77, 0.93)
5	FFT of beats	II	0.76 (0.67, 0.86)	0.89 (0.86, 0.98)	0.88 (0.84, 0.97)	0.87 (0.80, 0.94)
6	FFT of beats	V	0.69 (0.60, 0.79)	0.87 (0.79, 0.94)	0.87 (0.79, 0.94)	0.88 (0.80, 0.94)

entry by Kalidas and Tamil (2016) achieved a Challenge score of 75.10 in the real-time event and 76.70 when considering both real-time and retrospective events. The second-place entry in v-tach alarm by Plesinger et al. (2016) used ECG, ABP, and PPG waveforms, and achieved a Challenge score of 72.73 in real-time event, and 75.07 when considering both real-time and retrospective events. The performance is generally better in retrospective events, since data post alarm onset can be used for classification.

We find that the SDAE achieved a better F1-score and Challenge score than the MLP and MLP-PCA. The SDAE achieved an AUC of 0.91 (0.85, 0.97) and a higher F1 score (0.73 vs 0.65) and Challenge score (77.6 vs 73.8) than PCA/MLP. Using data only from prior to the alarm onset, SDAE with beat-level FFT-transformed ECG (from the targeted 3-second interval) achieved a Challenge score of 77.59, outperforming other baselines, including the winning Challenge 2015 entry (score 75.10) by Kalidas and Tamil in the real-time event. SDAE reduced the v-tach false alarm rate of the test set from 79.64% to 11.3% (suppressed 151 false v-tach alarms), at the cost of missing 11.1% (5 out of 45) true alarms. We report the specificity of SDAE when sensitivity is set at 0.89; SDAE has a higher specificity (0.86) than the other baselines for the real-time event.

Table 2: Performance Using Two-Channel ECG (N=221). RT (realtime), Retro (retrospective).

		Event	Features	Sens.	Spec.	Prec.	F1	AUC	Score
1	MLP	RT	FFT 3-sec interval	0.89	0.73	0.45	0.60	0.87 (0.80, 0.94)	67.98
2	PCA-MLP	RT	FFT 3-sec interval	0.87	0.78	0.48	0.62	0.89 (0.83, 0.96)	69.29
3	SDAE	RT	FFT 3-sec interval	0.84	0.79	0.51	0.63	0.89 (0.82, 0.95)	68.49
4	MLP	RT	FFT of beats	0.89	0.67	0.52	0.64	0.89 (0.83, 0.96)	71.65
5	PCA-MLP	RT	FFT of beats	0.89	0.80	0.50	0.65	0.88 (0.82, 0.95)	73.82
6	SDAE	RT	FFT of beats	0.89	0.86	0.62	0.73	0.91 (0.85, 0.97)	77.59
7	Challenge,1st	RT	ECG	0.89	0.80	-	-	-	75.10
8	Challenge, 2nd	RT	ECG/ABP	0.82	0.84	-	-	-	72.73
9	Challenge, 1st	RT/Retro	ECG	0.90	0.82	-	-	-	76.75
10	Challenge, 2nd	RT/Retro	ECG/ABP	0.85	0.84	-	-	-	75.07

3.4. Varying the number of training examples

We investigate how the SDAE behaves under different feature representations (beat-by-beat vs. FFT-transformed 3-second waveform segments) as a function of the labeled data size. Table 3 shows the classification performance of SDAE as the training data size varies from 25 to over 300 using FFT transformation of 3-second segment versus beat-by-beat

data respectively. Using FFT-transformed beat-by-beat data, SDAE scales gracefully as the training sample size is reduced to 25.

Table 3: SDAE performance as a function of training size using two ECG channels. (Test set N=221)

	Training Size	25	50	150	250	337
1	FFT waveform	0.63 (0.53, 0.72)	0.63 (0.53, 0.72)	0.87 (0.80, 0.94)	0.88 (0.81, 0.95)	0.89 (0.82, 0.95)
2	FFT B2B	0.86 (0.79, 0.93)	0.87 (0.80, 0.94)	0.91 (0.85, 0.97)	0.91 (0.85, 0.97)	0.91 (0.85, 0.97)

Using 25 training samples, SDAE with beat-by-beat data achieved an AUC of 0.86 (0.79, 0.93) which is only slightly worse than when using the full training data size (N=337) ($p = 0.035$). Using 50 or more training samples, SDAE with beat-by-beat data achieved similar performance as using the full data set (AUC of 0.87 [0.80, 0.94] trained with 50 samples vs. 0.91 [0.85, 0.97] with full dataset, p-val = 0.058). This showcases that leveraging beat level information is crucial for ensuring good predictive performance when labels are scarce.

4. Discussion and Related Work

Human heart beats generate a wide-range of complex ECG dynamics, which have been studied extensively under both healthy and pathological conditions. Classical ECG analyses are based on hand-crafted features obtained from temporal and/or frequency analyses, which are then used as inputs in a machine learning classifier. Recent advances on deep learning inspire new models where features are learned from segments of ECG signals. While deep learning has made significant advances in the domains of image and voice analysis, the application of deep learning in physiological waveform analysis has had limited success, partly due to limited availability of labeled data. Expert-defined rule-based approaches or simple machine learning models (such as gradient boosting, or random forest) combined with hand-crafted features often outperform more complex models, such as deep neural networks (Clifford et al., 2017).

Here, our results indicate that direct application of several machine learning techniques on raw waveforms performed poorly (with the current training sample size). We show that learning representations of the FFT-transformed ECG waveforms results in significantly better performance than using raw waveforms. We study both linear and non-linear embeddings of ECG for the purpose of detecting true v-tach alarms. When comparing the performance of PCA/MLP and SDAE using various ECG feature transformation, we observe that both achieved similar AUCs in most settings. When beat-level frequency features from two-channels of ECG were used, SDAE achieved a slightly higher AUC and a better Challenge score than PCA/MLP. Further investigation is required to characterize the performance of these approaches as the sample size increases. One avenue of future work is to leverage large amounts of unlabeled data to improve the quality of the nonlinear embeddings in the SDAE.

Tests on 2015 PhysioNet Challenge dataset indicate that, for both linear and non-linear embeddings, representation learning approaches that exploit the underlying known physiological structure, specifically, using FFT-transformed ECG beats, averaged over multiple cardiac cycles, lead to higher Challenge scores compared to models that use the entire ECG waveform segments. In the case of SDAE, this procedure is key to out-performing previ-

ous known best results from PhysioNet 2015 Challenge. We conjecture that this averaged fourier feature representation significantly simplifies the learning task (relative to models that use the entire ECG waveform segments) under the constraint of limited labeled data.

The 2015 PhysioNet Challenge event focused on five types of life-threatening arrhythmias, including asystole, extreme bradycardia, extreme tachycardia, ventricular tachycardia, and ventricular fibrillation/flutter. More than 200 entries were submitted for this event (see the collection of articles in the special issue of Physiological Measurement (Clifford et al., 2015)). Among all the arrhythmia alarm types, v-tach has proven to be the hardest to classify, and received the lowest prediction accuracy (Clifford et al., 2015). While machine learning approaches have been proposed (Eerikinen et al., 2016; Ansari et al., 2016), the winning entry in v-tach alarm classification by Kalidas and Tamil (2016) used logical analysis to improve classification results from SVM classification, and Plesinger et al. (2016) used signal processing and rule-based reasoning to achieve the second best performance in v-tach alarm classification in the 2015 PhysioNet Challenge.

Early work on v-tach false alarm reduction by Aboukhalil et al. (2008) used a data fusion approach with rule-based logic to reduce 5 types of false arrhythmia alarms from 42.7% to 17.2% (on average) when simultaneous ECG and arterial blood pressure waveform were available. They reported false v-tach alarm suppression as the most challenging task (with the lowest suppression rate among all alarm categories tested), with a reduced false v-tach alarm rate from 46.6% to 30.8%, at the cost of suppressing 14.5% and 4% of the true alarms in the train and test set respectively. In (Schwab et al., 2018), a supervised multi-task learning approach was proposed to reduce the number of training labels required to suppress general ICU false alarms. Our work, in contrast, focuses on v-tach false alarms, and uses FFT-transformation of individual ECG beats for scalable learning. Rajpurkar et al. (2016) used a deep convolutional neural network for arrhythmia detection, and obtained good performance with a significantly larger dataset.

5. Conclusion

We developed a supervised representation learning approach to detect false v-tach alarms from two leads of ECG waveforms, and obtained improved performance over several baselines, including previous results in the same task using the PhysioNet 2015 Challenge dataset. Our final best-performing model used a supervised denoising autoencoder, SDAE, to learn non-linear embeddings of spectral dynamics, averaged over multiple cardiac cycles. These results suggest that generative modeling may play a role in tackling the problem of v-tach false alarm detection. Future work will extend the current approach to combine information from multi-channel physiological waveforms (PPG, ABP) for false arrhythmia alarm reduction. We expect the full potential of such representational learning methods will lead to more significant results when the sample sizes are greatly increased. We also expect that representation learning will play an important role in analyzing other biomedical waveforms, such as time series of blood pressure, respiration, electroencephalogram (EEG), and photoplethysmogram (PPG).

ACKNOWLEDGMENTS

The authors thank Dr. Gari Clifford and the 2015 PhysioNet/CinC Challenge organizing team; we especially thank Mr. Benjamin Moody, from for his help in calculating and providing the realtime Challenge scores for Challenge entries. We thank Dr. Wei Zong for insightful discussion; Dr. Alistair Johnson for his valuable contribution in the ECG beat annotation software. This work was in part supported by the National Institutes of Health (NIH) grant R01GM104987 and the National Science Foundation (NSF) grant CMMI-1661615. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, or the NSF.

References

A Aboukhalil, L Nielsen, M Saeed, RG Mark, and GD Clifford. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *J Biomed Inform*, 41 (3):442–51, 2008.

Sardar Ansari, Ashwin Belle, Hamid Ghanbari, Mark Salamango, and Kayvan Najarian. Suppression of false arrhythmia alarms in the icu: a machine learning approach. *Physiological Measurements*, 37(8):1186–203, 2016.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 899–907. Curran Associates, Inc., 2013.

H Bourlard and Y Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cyernetics*, 59(4-5):291–294, 1988.

Gari Clifford. *Advanced Methods and Tools for ECG Data Analysis*. Artech House, 2006.

Gari Clifford, Ikaro Silva, Benjamin Moody, Qiao Li, Danesh Kella, Abdullah Shahin, Tristan Kooistra, Diane Perry, and Roger Mark. The physionet/computing in cardiology physionet challenge 2015: Reducing false arrhythmia alarms in the icu. In *Proceedings of Computing in Cardiology*, 2015.

Gari Clifford, Chengyu Liu, Benjamin Moody, Liwei Lehman, Ikaro Silva, Qiao Li, Alistair Johnson, and Roger Mark. Af classification from a short single lead ecg recording: the physionet computing in cardiology challenge 2017. In *Proceedings of Computing in Cardiology*, 2017.

GD Clifford, I Silva, B Moody, Q Li, D Kella, A Chahin, Kooistra T, D Perry, and RG. Mark. False alarm reduction in critical care. *Physiological Measurements*, 37(8):E5–E23, 2016.

Elizabeth R DeLong, David M DeLong, and Daniel Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

Barbara J Drew, Patricia Harris, Jessica K. Zgre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. *PLOSONE*, 9(10), 2014.

Linda Eerikinen, Joaquin Vanschoren, Michael J. Rooijakkers, Rik Vullings, and Ronald M. Aarts. Reduction of false arrhythmia alarms using signal selection and machine learning. *Physiological Measurements*, 2016.

AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, PCh Ivanov, RG Mark, JE Mietus, GB Moody, C-K Peng, and HE Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101 (23):e215–e220, June 2000.

A. E. W. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster. Multimodal heart beat detection using signal quality indices. *Physiological Measurements*, 36:1665–1677, 2015. Code available at https://github.com/alistairewj/peak-detector/.

V. Kalidas and L.S. Tamil. Cardiac arrhythmia classification using multi-modal signal analysis. *Physiological Measurements*, 37(8):1253–72, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lovedeep and Gondara. Medical image denoising using convolutional denoising autoencoders. *IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246, 12 2016.

GB Moody and RG Mark. The impact of the mit-bih arrhythmia database. *IEEE Eng in Med and Biol*, 20(3):45–50, 2001.

Jiapu Pan and Willis Tompkin. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering*, 3(4-5):230–236, 1985.

F Plesinger, P Klimes, J Halamek, and P Jurak. Taming of the monitors: reducing false alarms in intensive care units. *Physiological Measurements*, 37:1313–1325, 2016.

Pranav Rajpurkar, Awni Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. In *https://arxiv.org/pdf/1707.01836.pdf*, 2016.

Patrick Schwab, Emanuela Keller, Carl Muroi, David J. Mack, Christian Strässle, and Walter Karlen. Not to cry wolf: Distantly supervised multitask learning in critical care. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Pascal Vincent, Hugo Larochelle, Y Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 01 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010.

Appendix A: Identify Potential V-Tach Episodes

A model for ventricular beats identification was trained using a total of 107,129 annotated beats from lead II of the MIT-BIH database (7,104 ventricular beats). A model for lead V was also trained using 96,189 annotated beats from lead V in MIT-BIH database (7,010 ventricular beats). We extracted the FFT features from all beats, and trained a ventricular-beat logistic classifiers to classify a beat as either a ventricular or non-ventricular beat based on the FFT transformed features. The 10-fold cross validated median training AUC (and inter-quartile range) of this ventricular beat classifier is 0.94 (0.93, 0.94) using the MIT-BIH arrhythmia dataset.

For each beat (in the last 25-seconds of each record) in the PhysioNet Challenge dataset, we used the MIT-BIH ventricular beat model to estimate the probability of each beat being a ventricular beat based on the FFT of individual beats. At each beat j, with onset time t_j, we estimated the probability of the next five beats starting at time t_j being ventricular beats, by averaging v-beat probabilities of the following five beats. The onset time T_i of the potential v-tach region of each record i is identified as the onset time of the beat with the highest running 5-beat ventricular-beat probability (from among the beats with RR-intervals smaller than 600 ms). For each record i, we extracted its FFT features by averaging FFT of all individual beats in a K-second window starting at time T_i. We chose K to be 3 (seconds), since the v-tach is defined as 5 consecutive ventricular beats with heart rate of over 100 bpm, and therefore a 3-second of ECG interval should be sufficient in length to identify potential v-tach episodes (5 beats at less than 600 ms RR-interval per beat).

Appendix B: MLP Parameter Settings

The first layer of both input channels is a 64 neuron layer, with 'ReLU'-activation, followed by a dropout of 0.5. The channels are then added together and given to 128-neuron 'ReLU' layer. Finally, a final sigmoid layer is applied for prediction.

Appendix C: SDAE Parameter Settings

The SDAE uses the 'adam' optimizer in attempt to minimize 'mean-squared-error' and 'binary cross-entropy' loss. Each input channel sequentially leads into a layer of size 64, each using a 'ReLU' activation function. Afterwards, the output of the size 64 hidden layer is given to a layer size of 32, which uses a 'sigmoid' activation function. The output of both of the 32 unit hidden layers are added together in order to find an underlying representation of the combined data. Depending on the results of the grid search, Gaussian noise is applied to the summed layer. The leads are once again separated and sent into a decoding 64 unit 'ReLU' layer, and a 41 unit 'ReLU' layer, which attempts to reconstruct the original signal from its respective lead. The output of summed layer is fed into a simple feed forward neural network. This simple feed forward neural network consisting of only 1 hidden layer with a hidden unit size of 128 and a 'ReLU' activation function; the output is then fed into a sigmoid layer to obtain a prediction. The model utilizes the ADAM optimizer, and a batch size of 64.