

Generalization Studies of Neural Network Models for Cardiac Disease Detection Using Limited Channel ECG

Deepta Rajan, David Beymer, Girish Narayan
IBM Research, San Jose, CA, USA

Abstract

Acceleration of machine learning research in health-care is challenged by lack of large annotated and balanced datasets. Furthermore, dealing with measurement inaccuracies and exploiting unsupervised data are considered to be central to improving existing solutions. In particular, a primary objective in predictive modeling is to generalize well to both unseen variations within the observed classes, and unseen classes. In this work, we consider such a challenging problem in machine learning driven diagnosis detecting a gamut of cardiovascular conditions (e.g. infarction, dysrhythmia etc.) from limited channel ECG measurements. Though deep neural networks have achieved unprecedented success in predictive modeling, they rely solely on discriminative models that can generalize poorly to unseen classes. We argue that unsupervised learning can be utilized to construct effective latent spaces that facilitate better generalization. This work extensively compares the generalization of our proposed approach against a state-of-the-art deep learning solution. Our results show significant improvements in F1-scores.

1. Introduction

Longitudinal patient records are central to the realm of healthcare, comprising a historical aggregation of diverse information such as diagnostic codes, lab measurements, imaging exams, text reports etc. Analyzing sequences of events and episodes are required for diagnosis and treatment planning. Harnessing the power of artificial intelligence tools and technologies for such analysis, e.g. deep learning, could potentially improve patient care quality and reduce costs by digitizing healthcare [1]. However, the success of data-driven solutions primarily depends on two aspects: infrastructure to manage big data and deploy solutions at scale; and strong generalization abilities of computational models to produce reliable predictions in new environments. However, leveraging large volumes of patient data being routinely collected is challenged by both availability of expert annotations and our ability to discern meaningful correlations from the vast number of predictor variables. Further, these datasets are plagued by discrepan-

cies in measurements and imbalances in disease distributions. Consequently, despite the community-wide efforts of curating representative datasets [2], we operate in small-data regimes that result in highly biased clinical models.

In this paper, we focus on detecting cardiac abnormalities, such as myocardial infarction, a disease causing over 8 million deaths annually [3], using limited channel ECG. The standard 12-channel ECG is a prevalent diagnostic modality and the primary screening exam for heart ailments, with over 300 million signals recorded every year [4]. However, in certain cases, only a subset of these leads can be accessed. Examples include inpatient telemetry [5], and ambulatory heart rhythm monitoring [6]. In such cases, the goal is to obtain meaningful clinical conclusions using only the limited measurements [7].

A popular solution for sequence modeling includes Recurrent Neural Networks (RNN) based on Long Short-Term Memory (LSTM) units, and have had proven success with clinical time-series classification [1]. Besides, a recent empirical study [8] deemed 1-D Convolutional Neural Networks (CNN) to be a powerful and more effective alternative to model sequences. Specifically, 1-D Residue Networks (ResNet) have become a popular solution for rhythm classification [9]. More recently, new learning paradigms such as deep attention models [10] and temporal convolutional networks have been shown to produce improved performances. Broadly, these methods fall under the category of discriminatory modeling, which can be ineffective when dealing with out-of-distribution samples. Recently, Rajan *et.al.* [11] argued that a generative modeling approach can enable inference of latent features to describe complex distributions. In this paper, we build on this idea and propose a novel neural network architecture referred as *ResNet++* for limited channel ECG classification. Our approach uses an unsupervised generative model to construct latent features, followed by a discriminative model (1-D ResNet [9]) to detect anomalies. The resulting latent feature representations implicitly exploit information from the missing channels and can predict the entire 12-channel ECG from a subset of channels. Results show that the proposed approach provides improved generalization to new diseases, and is less sensitive to hyperparameter settings even with small and imbalanced data.

| Seq2Seq Architecture | |
|----------------------|---------------------|
| Input | Filters |
| I: 196 x L | GRU → Dropout (0.4) |
| h1 | GRU → Dropout (0.4) |
| h2 | GRU → Dropout (0.4) |
| h3 | GRU → Dropout (0.4) |
| h4 | GRU → Dropout (0.4) |
| Z: latent rep. | GRU → Dropout (0.4) |
| h5 | GRU → Dropout (0.4) |
| h6 | GRU → Dropout (0.4) |
| h7 | GRU → Dropout (0.4) |
| h8 | GRU → Dropout (0.4) |
| h9 | O: 196 x C |

C: Total number of channels; L: Limited subset
GRU: Gated Recurrent Unit Cell
h's: Hidden representations

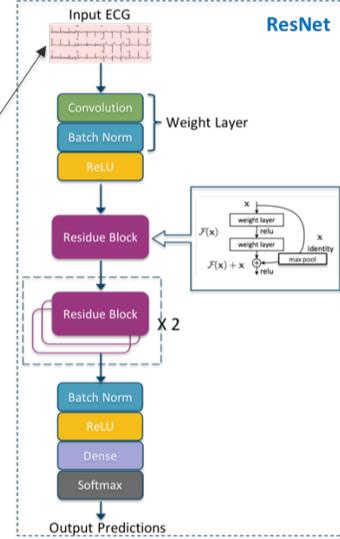


Figure 1. Proposed ResNet++ architecture, with Stage 1 depicting the unsupervised Seq2Seq model used to construct latent representation from limited channels, and Stage 2 depicts the ResNet model for disease detection.

2. Problem Formulation

In this section, we provide a formal definition of the problem and introduce the notations used in the rest of this paper. In limited channel problems, ECG signals from only 1 to 3 leads are assumed to be available, and the goal is to build predictive models for disease detection. In this paper, we consider inferior myocardial infarction (MI) leads (II, III, AvF) in order to build the model. From previous studies, it is known that these leads provide information to adequately localize inferior wall ischemia and infarction [12]. However, the ability of models trained on datasets dominated by subjects with infero MI conditions to generalize to other cardiac conditions is unknown. Such generalization studies emphasize the importance of choosing the optimal subset of leads that can be more broadly applicable in diagnosis. However, by restricting the algorithm to use only the inferior leads, disease generalization capacity of even sophisticated models suffers. To this end, we propose a novel approach that assumes access to full channel training data, builds meaningful latent spaces that can compensate for the missing channel data, and finally construct a deep neural network that performs predictions based on these latent features. Note that, in the rest of this paper, the terms leads and channels are used interchangeably. The multi-variate sequence dataset is represented as $X \in \mathbb{R}^{N \times T \times K}$, where N denotes the number of training samples, T denotes the number of time-steps in each measurement and K indicates the total number of channels.

Based on the limited channel configuration denoted by the set $\mathcal{C} = \{\text{II, III, aVF}\}$, whose cardinality $|\mathcal{C}| \ll K$, we extract the matrix $\hat{X} \in \mathbb{R}^{N \times T \times |\mathcal{C}|}$. In order to perform implicit completion of the missing data, we propose to build a generative model that attempts to recover X using \hat{X} . In this process, it infers a latent space that defines an effective metric to compare different samples.

3. Proposed Approach

Stage 1: As shown in Figure 1, the first stage of our algorithm employs unsupervised representation learning for predicting the complete ECG using only the 3-channel measurements. More specifically, we build an encoder-decoder architecture, commonly referred as Seq2Seq [13], with an optional attention mechanism. Though originally developed for machine translation, they are applicable to more general sequence to sequence transformation tasks. The architecture is comprised of two RNNs (based on Gated Recurrent Unit (GRU)), one each for encoder and decoder. The encoder transforms an input sequence from \hat{X} into a fixed length vector, either from the last time step of the sequence or by concatenating hidden representations from all time steps. The decoder then predicts the output sequence, in our case X , using the encoder output. Optionally, the decoder can also attend to a certain part of the encoder states through an attention mechanism. The attention mechanism often uses both content from the encoder states, and also context from the sequence generated

so far at the decoder. Our RNNs are designed using GRU cells, which are capable of learning long-term dependencies. Each GRU cell is comprised of the following operations, implemented using fully connected networks:

$$\begin{aligned} \text{(update gate): } z &= \sigma(x_t U^z + s_{t-1} W^z) \\ \text{(reset gate): } r &= \sigma(x_t U^r + s_{t-1} W^r) \\ \text{(hidden state): } h &= \tanh(x_t U^h + (s_{t-1} \circ r) W^h) \quad (1) \\ \text{(final output): } s_t &= (1 - z) \circ h + z \circ s_{t-1} \end{aligned}$$

A GRU has two gates, a reset gate r , and an update gate z . While the reset gate determines how to combine the current input with the previous memory state, the update gate defines how much of the previous memory to retain as we proceed to the next step. In the simplest case, when we set the reset gate to all 1s and the update gate to all 0s it simplifies into a plain RNN model. While the underlying idea of using gating mechanisms to capture long-term dependencies is the same as in a LSTM, they are different in terms of the number of gates and the absence of an internal memory that is different from the hidden state. The generative model is trained with an ℓ_2 loss at the decoder output. Note that, our architecture attempts to reconstruct the observed channels as well as predict the missing channel measurements.

Stage 2: We now design a classifier stage that exploits the latent space from the generative model trained for missing channel prediction. Interestingly, compared to discriminative models, this approach utilizes additional channel information from the training stage and builds a more effective metric for the whole data space instead of discriminating the normal/abnormal classes. Furthermore, since the first stage is unsupervised, we can use even unlabeled data to construct a more robust latent space. The classifier we use is a 1-D ResNet architecture with convolution, ReLU, batch normalization and dropout layers, as illustrated in Figure 1.

4. Experiment Setup

In this section, we describe the dataset used for all empirical analysis that compares the performance of ResNet++ over ResNet for the problem of cardiac disease detection using limited channel ECG. We also provide details on the choice of ECG channel configurations, training parameters, and evaluation metrics employed.

Dataset: The widely adopted Physionet [2] repository hosts the PTB database (PTBDB) [14] that comprises a set of 30 seconds long, 12 channel ECG records, sampled at 1000 Hz. It includes ECGs collected from both healthy patients and those diagnosed with diseases such as bundle branch block, valvular heart disease etc. Every record also contains a header file with demographics, medical history and diagnosis related information curated

Table 1. ECG class labels and sample sizes from the Physionet PTBDB used in our disease generalization study.

| Class Label | Sample Size |
|---|-------------|
| Myocardial Infarction: inferior | 3222 |
| Myocardial Infarction: antero-septal | 2855 |
| Myocardial Infarction: infero-lateral | 1933 |
| Myocardial Infarction: anterior | 1685 |
| Myocardial Infarction: antero-lateral | 1603 |
| Myocardial Infarction: no | 628 |
| Myocardial Infarction: infero-postero-lateral | 573 |
| Myocardial Infarction: postero-lateral | 185 |
| Myocardial Infarction: posterior | 148 |
| Myocardial Infarction: infero-poster-lateral | 111 |
| Myocardial Infarction: lateral | 111 |
| Myocardial Infarction: infero-lateral | 86 |
| Myocardial Infarction: antero-septo-lateral | 74 |
| Myocardial Infarction: infero-posterior | 12 |
| Bundle Branch Block | 623 |
| Cardiomyopathy | 603 |
| Dysrhythmia | 411 |
| Valvular Heart Disease | 122 |
| Healthy Control | 3055 |

from echocardiography exams. In our study, a total of 504 records from 250 patients were used. The ECG records were pre-processed to remove noise, reduce sampling rate, and divided into frames of about 3 seconds long prior to being parsed through the neural network, as described in [15]. The resulting dataset contains 18,040 ECG frames downsampled to 64 Hz .

ECG leads: As depicted in Table 1, the dataset is dominated by samples representing myocardial infarction (MI) occurring in various regions of the heart, namely: inferior, anterior, lateral, septal, and posterior. These localized abnormal activities are captured in different combinations of ECG leads. For example, inferior leads (II, III, aVF) provide information to adequately localize inferior wall ischemia and infarction while anterior leads (V1, V2, V3) are effective in detecting anterior wall infarction. Similarly, lead II and V1 provide excellent assessment of atrial activity enabling detection of atrial arrhythmias, and leads V1 and V2 can often be used to differentiate left from right bundle branch block (BBB) patterns [12]. To study the generalization of deep models across diseases, we create two scenarios defined by choice of ECG leads which in turn determines the disease that the model can be trained to detect. A train-test pair dataset is created for each scenario, with the train data containing samples from one disease class, and the healthy class. The model is then tested on its ability to classify data it was trained to detect, in addition to the rest of the unobserved disease samples. In the first scenario, we select the leads II, III, aVF and create

| Channel Configuration: II, III, aVF | | | |
|--|--------|-------------|--------|
| Input Classes: Healthy Control, Infero Myocardial Infarction | | | |
| Cardiac Disease | ResNet | ResNet++ | % Gain |
| Infero MI | 0.84 | 0.87 | 3.57 |
| Antero MI | 0.87 | 0.89 | 2.30 |
| Bundle Branch Block | 0.59 | 0.66 | 11.86 |
| Dysrhythmia | 0.67 | 0.71 | 5.97 |
| Cardiomyopathy | 0.77 | 0.81 | 5.19 |
| Valvular Heart Disease | 0.32 | 0.33 | 3.13 |

Figure 2. F1-Scores demonstrating the improvement of ResNet++ over ResNet on the Physionet PTB dataset.

a training set of size 7, 246 with samples belonging to infero MI and healthy control group only, while keeping the rest of the disease classes such as posterior MI, BBB, dysrhythmia etc. along with a portion of infero MI and healthy samples in the test set giving rise to a sample size 10, 794. Correspondingly, in the second scenario we select ECG leads V1, V2, V3, and create a train-test dataset pair with only antero MI and healthy classes in the training set.

Training parameters: Stage 1 of ResNet++ uses 5 GRU layers, while stage 2 uses 7 convolutional layers in comparison to the 13 layers used by ResNet. Further, a batch size of 32 samples, the Adam optimizer with a learning rate of 0.001, and the categorical cross entropy loss function was used for training.

Evaluation metric: In order to evaluate the predictive models, we compute the F1-score, a summary metric typically used to trade-off between precision and recall when datasets are imbalanced.

5. Results

Our results on the PTB dataset show that, the proposed *ResNet++* model trained to detect only inferior Myocardial Infarction (MI) using leads II, III, and aVF, is more effective than ResNet in detecting other MI variants (anterior, posterior, lateral), bundle branch block, dysrhythmia etc. Overall we achieve a 2 to 11% improvement in F1-scores when generalizing to new unobserved diseases as shown in Figure 2. More importantly, *ResNet++* leverages unlabeled data in its unsupervised pre-training stage. In addition, it is robust to out-of-distribution samples, less sensitive to training hyperparameters and produces significant improvements in generalization capabilities over ResNet even in highly imbalanced data scenarios.

References

[1] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference. 2016; 301–318.

[2] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* 2000 (June 13);101(23):e215–e220. *Circulation Electronic Pages*: <http://circ.ahajournals.org/content/101/23/e215.full> PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[3] Ansari S, Farzaneh N, Duda M, Horan K, Andersson HB, Goldberger ZD, Nallamothu BK, Najarian K. A review of automated methods for detection of myocardial ischemia and infarction using electrocardiogram and electronic health records. *IEEE reviews in biomedical engineering* 2017;10:264–298.

[4] Hedén B, Ohlin H, Rittner R, Edenbrandt L. Acute myocardial infarction detected in the 12-lead ecg by artificial neural networks. *Circulation* 1997;96(6):1798–1802.

[5] Sandau KE, Funk M, Auerbach A, Barsness GW, Blum K, Cvach M, Lampert R, May JL, McDaniel GM, Perez MV, et al. Update to practice standards for electrocardiographic monitoring in hospital settings: a scientific statement from the american heart association. *Circulation* 2017;CIR–0000000000000527.

[6] Kennedy HL. The evolution of ambulatory ecg monitoring. *Progress in cardiovascular diseases* 2013;56(2):127–132.

[7] Atoui H, Fayn J, Rubel P. A novel neural-network model for deriving standard 12-lead ecgs from serial three-lead ecgs: application to self-care. *IEEE transactions on information technology in biomedicine* 2010;14(3):883–890.

[8] Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv180301271* 2018;.

[9] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv170701836* 2017;.

[10] Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend and Diagnose: Clinical Time Series Analysis using Attention Models. *Proceedings of AAAI 2018* 2018;.

[11] Rajan D, Thiagarajan JJ. A generative modeling approach to limited channel ecg classification. *arXiv preprint arXiv180206458* 2018;.

[12] Dubin D. *Rapid Interpretation of EKG's*, volume 200. Cover Publishing Company Tampa (FL), 1996.

[13] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 2014; 3104–3112.

[14] Boussejot R, Kreiseler D, Schnabel A. Nutzung der ekg-signal-datenbank cardiodat der ptb über das internet. *Biomedizinische Technik/Biomedical Engineering* 1995; 40(s1):317–318.

[15] Reasat T, Shahnaz C. Detection of inferior myocardial infarction using shallow convolutional neural networks. *arXiv preprint arXiv171001115* 2017;.

Address for correspondence:

Deeptra Rajan - 650, Harry Road, San Jose, CA, USA - 95120