Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks

Pranav Rajpurkar[*]   PRANAVSR@CS.STANFORD.EDU
Awni Y. Hannun[*]   AWNI@CS.STANFORD.EDU
Masoumeh Haghpanahi   MHAGHPANAHI@IRHYTHMTECH.COM
Codie Bourn   CBOURN@IRHYTHMTECH.COM
Andrew Y. Ng   ANG@CS.STANFORD.EDU

Abstract

We develop an algorithm which exceeds the performance of board certified cardiologists in detecting a wide range of heart arrhythmias from electrocardiograms recorded with a single-lead wearable monitor. We build a dataset with more than 500 times the number of unique patients than previously studied corpora. On this dataset, we train a 34-layer convolutional neural network which maps a sequence of ECG samples to a sequence of rhythm classes. Committees of board-certified cardiologists annotate a gold standard test set on which we compare the performance of our model to that of 6 other individual cardiologists. We exceed the average cardiologist performance in both recall (sensitivity) and precision (positive predictive value).

1. Introduction

We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist. Key to exceeding expert performance is a deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations along with a novel dataset two orders of magnitude larger than previous datasets of its kind.

Many heart diseases, including Myocardial Infarction, AV Block, Ventricular Tachycardia and Atrial Fibrillation can all be diagnosed from ECG signals with an estimated 300 million ECGs recorded annually (Hedén et al., 1996). We investigate the task of arrhythmia detection from the ECG record. This is known to be a challenging task for computers but can usually be determined by an expert from a single, well-placed lead.

---

[*]Authors contributed equally.
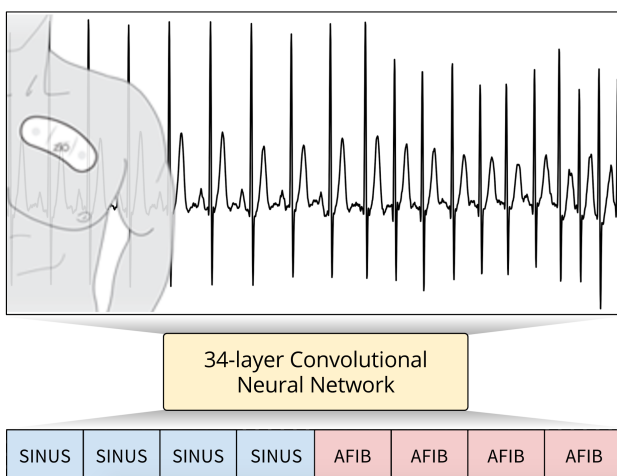   Project website at https://stanfordmlgroup.github.io/projects/ecg

Figure 1. Our trained convolutional neural network correctly detecting the sinus rhythm (SINUS) and Atrial Fibrillation (AFIB) from this ECG recorded with a single-lead wearable heart monitor.

Arrhythmia detection from ECG recordings is usually performed by expert technicians and cardiologists given the high error rates of computerized interpretation. One study found that of all the computer predictions for non-sinus rhythms, only about 50% were correct (Shah & Rubin, 2007); in another study, only 1 out of every 7 presentations of second degree AV block were correctly recognized by the algorithm (Guglin & Thatai, 2006). To automatically detect heart arrhythmias in an ECG, an algorithm must implicitly recognize the distinct wave types and discern the complex relationships between them over time. This is difficult due to the variability in wave morphology between patients as well as the presence of noise.

We train a 34-layer convolutional neural network (CNN) to detect arrhythmias in arbitrary length ECG time-series. Figure 1 shows an example of an input to the model. In addition to classifying noise and the sinus rhythm, the network learns to classify and segment twelve arrhythmia types present in the time-series. The model is trained end-

to-end on a single-lead ECG signal sampled at 200Hz and a sequence of annotations for every second of the ECG as supervision. To make the optimization of such a deep model tractable, we use residual connections and batch-normalization (He et al., 2016b; Ioffe & Szegedy, 2015). The depth increases both the non-linearity of the computation as well as the size of the context window for each classification decision.

We construct a dataset 500 times larger than other datasets of its kind (Moody & Mark, 2001; Goldberger et al., 2000). One of the most popular previous datasets, the MIT-BIH corpus contains ECG recordings from 47 unique patients. In contrast, we collect and annotate a dataset of about 30,000 unique patients from a pool of nearly 300,000 patients who have used the Zio Patch monitor[1] (Turakhia et al., 2013). We intentionally select patients exhibiting abnormal rhythms in order to make the class balance of the dataset more even and thus the likelihood of observing unusual heart-activity high.

We test our model against board-certified cardiologists. A committee of three cardiologists serve as gold-standard annotators for the 336 examples in the test set. Our model exceeds the individual expert performance on both recall (sensitivity), and precision (positive predictive value) on this test set.

2. Model

Problem Formulation

The ECG arrhythmia detection task is a sequence-to-sequence task which takes as input an ECG signal $X = [x_1, ..x_k]$, and outputs a sequence of labels $r = [r_1, ...r_n]$, such that each r_i can take on one of m different rhythm classes. Each output label corresponds to a segment of the input. Together the output labels cover the full sequence.

For a single example in the training set, we optimize the cross-entropy objective function

$$\mathcal{L}(X, r) = \frac{1}{n} \sum_{i=1}^{n} \log p(R = r_i \mid X)$$

where $p(\cdot)$ is the probability the network assigns to the i-th output taking on the value r_i.

Model Architecture and Training

We use a convolutional neural network for the sequence-to-sequence learning task. The high-level architecture of the network is shown in Figure 2. The network takes as input a time-series of raw ECG signal, and outputs a sequence of label predictions. The 30 second long ECG signal is
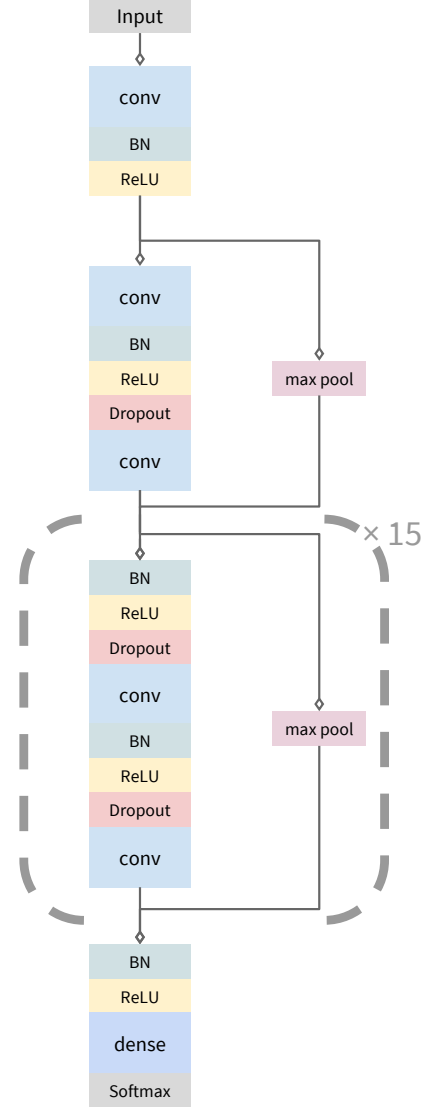
Figure 2. The architecture of the network. The first and last layer are special-cased due to the pre-activation residual blocks. Overall, the network contains 33 layers of convolution followed by a fully-connected layer and a softmax.

sampled at 200Hz, and the model outputs a new prediction once every second. We arrive at an architecture which is 33 layers of convolution followed by a fully connected layer and a softmax.

In order to make the optimization of such a network tractable, we employ shortcut connections in a similar manner to those found in the Residual Network architecture (He et al., 2015b). The shortcut connections between neural-network layers optimize training by allowing information to propagate well in very deep neural networks. Before the input is fed into the network, it is normalized using a
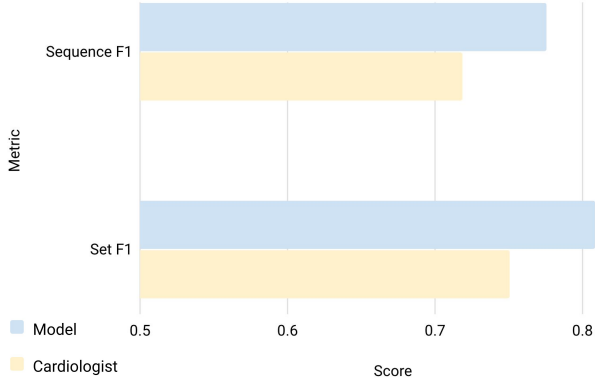
Figure 3. Evaluated on the test set, the model outperforms the average cardiologist score on both the Sequence and the Set F1 metrics.

robust normalization strategy. The network consists of 16 residual blocks with 2 convolutional layers per block. The convolutional layers all have a filter length of 16 and have $64k$ filters, where k starts out as 1 and is incremented every 4-th residual block. Every alternate residual block subsamples its inputs by a factor of 2, thus the original input is ultimately subsampled by a factor of 2^8. When a residual block subsamples the input, the corresponding shortcut connections also subsample their input using a Max Pooling operation with the same subsample factor.

Before each convolutional layer we apply Batch Normalization (Ioffe & Szegedy, 2015) and a rectified linear activation, adopting the pre-activation block design (He et al., 2016a). The first and last layers of the network are special-cased due to this pre-activation block structure. We also apply Dropout (Srivastava et al., 2014) between the convolutional layers and after the non-linearity. The final fully connected layer and softmax activation produce a distribution over the 14 output classes for each time-step.

We train the networks from scratch, initializing the weights of the convolutional layers as in (He et al., 2015a). We use the Adam (Kingma & Ba, 2014) optimizer with the default parameters and reduce the learning rate by a factor of 10 when the validation loss stops improving. We save the best model as evaluated on the validation set during the optimization process. [ht]

3. Data

Training

We collect and annotate a dataset of 64,121 ECG records from 29,163 patients. The ECG data is sampled at a frequency of 200 Hz and is collected from a single-lead, non-invasive and continuous monitoring device called the Zio

Patch which has a wear period up to 14 days (Turakhia et al., 2013). Each ECG record in the training set is 30 seconds long and can contain more than one rhythm type. Each record is annotated by a clinical ECG expert: the expert highlights segments of the signal and marks it as corresponding to one of the 14 rhythm classes.

The 30 second records were annotated using a web-based ECG annotation tool designed for this work. Label annotations were done by a group of Certified Cardiographic Technicians who have completed extensive training in arrhythmia detection and a cardiographic certification examination by Cardiovascular Credentialing International. The technicians were guided through the interface before they could annotate records. All rhythms present in a strip were labeled from their corresponding onset to offset, resulting in full segmentation of the input ECG data. To improve labeling consistency among different annotators, specific rules were devised regarding each rhythm transition.

We split the dataset into a training and validation set. The training set contains 90% of the data. We split the dataset so that there is no patient overlap between the training and validation sets (as well as the test set described below).

Testing

We collect a test set of 336 records from 328 unique patients. For the test set, ground truth annotations for each record were obtained by a committee of three board-certified cardiologists; there are three committees responsible for different splits of the test set. The cardiologists discussed each individual record as a group and came to a consensus labeling. For each record in the test set we also collect 6 individual annotations from cardiologists not participating in the group. This is used to assess performance of the model compared to an individual cardiologist.

Rhythm Classes

We identify 12 heart arrhythmias, sinus rhythm and noise for a total of 14 output classes. The arrhythmias are characterized by a variety of features. Table 2 in the Appendix shows an example of each rhythm type we classify. The noise label is assigned when the device is disconnected from the skin or when the baseline noise in the ECG makes identification of the underlying rhythm impossible.

The morphology of the ECG during a single heart-beat as well as the pattern of the activity of the heart over time determine the underlying rhythm. In some cases the distinction between the rhythms can be subtle yet critical for treatment. For example two forms of second degree AV Block, Mobitz I (Wenckebach) and Mobitz II (here referred to as AVB_TYPE2) can be difficult to distinguish. Wenckebach is considered benign and Mobitz II is considered patholog-

	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	0.632	**0.774**	0.720
JUNCTIONAL	0.670	**0.684**	**0.783**	0.674
NOISE	**0.823**	0.768	**0.704**	0.689
SINUS	**0.879**	0.847	**0.939**	0.907
SVT	**0.477**	0.449	**0.658**	0.556
TRIGEMINY	**0.908**	0.843	**0.870**	0.816
VT	0.506	**0.566**	0.694	**0.769**
WENCKEBACH	**0.709**	0.593	**0.806**	0.736
Aggregate Results				
Precision (PPV)	**0.800**	0.723	**0.809**	0.763
Recall (Sensitivity)	**0.784**	0.724	**0.827**	0.744
F1	**0.776**	0.719	**0.809**	0.751

Table 1. The top part of the table gives a class-level comparison of the expert to the model F1 score for both the Sequence and the Set metrics. The bottom part of the table shows aggregate results over the full test set for precision, recall and F1 for both the Sequence and Set metrics.

ical, requiring immediate attention (Dubin, 1996).

Table 2 in the Appendix also shows the number of unique patients in the training (including validation) set and test set for each rhythm type.

4. Results

Evaluation Metrics

We use two metrics to measure model accuracy, using the cardiologist committee annotations as the ground truth.

Sequence Level Accuracy (F1): We measure the average overlap between the prediction and the ground truth sequence labels. For every record, a model is required to make a prediction approximately once per second (every 256 samples). These predictions are compared against the ground truth annotation.

Set Level Accuracy (F1): Instead of treating the labels for a record as a sequence, we consider the set of unique arrhythmias present in each 30 second record as the ground truth annotation. Set Level Accuracy, unlike Sequence Level Accuracy, does not penalize for time-misalignment within a record. We report the F1 score between the unique class labels from the ground truth and those from the model prediction.

In both the Sequence and the Set case, we compute the F1 score for each class separately. We then compute the overall F1 (and precision and recall) as the class-frequency weighted mean.

Model vs. Cardiologist Performance

We assess the cardiologist performance on the test set. Recall that each of the records in the test set has a ground truth label from a committee of three cardiologists as well as individual labels from a disjoint set of 6 other cardiologists. To assess cardiologist performance for each class, we take the average of all the individual cardiologist F1 scores using the group label as the ground truth annotation.

Table 1 shows the breakdown of both cardiologist and model scores across the different rhythm classes. The model outperforms the average cardiologist performance on most rhythms, noticeably outperforming the cardiologists in the AV Block set of arrhythmias which includes Mobitz I (Wenckebach), Mobitz II (AVB_Type2) and complete heart block (CHB). This is especially useful given the severity of Mobitz II and complete heart block and the importance of distinguishing these two from Wenckebach which is usually considered benign.

Table 1 also compares the aggregate precision, recall and F1 for both model and cardiologist compared to the ground truth annotations. The aggregate scores for the cardiologist are computed by taking the mean of the individual cardiologist scores. The model outperforms the cardiologist average in both precision and recall.

5. Analysis

The model outperforms the average cardiologist score on both the sequence and the set F1 metrics. Figure 4 shows a confusion matrix of the model predictions on the test set. Many arrhythmias are confused with the sinus rhythm. We expect that part of this is due to the sometimes ambiguous location of the exact onset and offset of the arrhythmia in the ECG record.

Often the mistakes made by the model are understandable. For example, confusing Wenckebach and AVB_Type2 makes sense given that the two rhythms in general have very similar ECG morphologies. Similarly, Supraventricular Tachycardia (SVT) and Atrial Fibrillation (AFIB) are often confused with Atrial Flutter (AFL) which is understandable given that they are all atrial arrhythmias. We also note that Idioventricular Rhythm (IVR) is sometimes mistaken as Ventricular Tachycardia (VT), which again makes sense given that the two only differ in heart-rate and are difficult to distinguish close to the 100 beats per minute delineation.
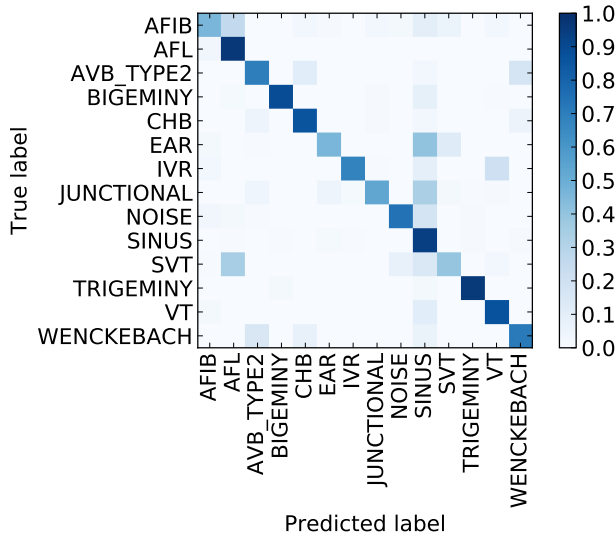
Figure 4. A confusion matrix for the model predictions on the test set. Many of the mistakes the model makes are not surprising. For example, confusing second degree AV Block (Type 2) with Wenckebach makes sense given the often similar expression of the two arrhythmias in the ECG record.

One of the most common confusions is between Ectopic Atrial Rhythm (EAR) and sinus rhythm. The main distinguishing criteria for this rhythm is an irregular P wave. This can be subtle to detect especially when the P wave has a small amplitude or when noise is present in the signal.

6. Related Work

Automatic high-accuracy methods for R-peak extraction have existed at least since the mid 1980's (Pan & Tompkins, 1985). Current algorithms for R-peak extraction tend to use wavelet transformations to compute features from the raw ECG followed by finely-tuned threshold based classifiers (Li et al., 1995; Martínez et al., 2004). Because accurate estimates of heart rate and heart rate variability can be extracted from R-peak features, feature-engineered algorithms are often used for coarse-grained heart rhythm classification, including detecting tachycardias (fast heart rate), bradycardias (slow heart rate), and irregular rhythms. However, such features alone are not sufficient to distinguish between most heart arrhythmias since features based on the atrial activity of the heart as well as other features pertaining to the QRS morphology are needed.

Much work has been done to automate the extraction of other features from the ECG. For example, beat classification is a common sub-problem of heart-arrhythmia classification. Drawing inspiration from automatic speech recognition, Hidden Markov models with Gaussian observation probability distributions have been applied to the task of

beat detection (Coast et al., 1990). Artificial neural networks have also been used for the task of beat detection (Melo et al., 2000). While these models have achieved high-accuracy for some beat types, they are not yet sufficient for high-accuracy heart arrhythmia classification and segmentation. For example, (Artis et al., 1991) train a neural network to distinguish between Atrial Fibrillation and Sinus Rhythm on the MIT-BIH dataset. While the network can distinguish between these two classes with high-accuracy, it does not generalize to noisier single-lead recordings or classify among the full range of 15 rhythms available in MIT-BIH. This is in part due to insufficient training data, and because the model also discards critical information in the feature extraction stage.

The most common dataset used to design and evaluate ECG algorithms is the MIT-BIH arrhythmia database (Moody & Mark, 2001) which consists of 48 half-hour strips of ECG data. Other commonly used datasets include the MIT-BIH Atrial Fibrillation dataset (Moody & Mark, 1983) and the QT dataset (Laguna et al., 1997). While useful benchmarks for R-peak extraction and beat-level annotations, these datasets are too small for fine-grained arrhythmia classification. The number of unique patients is in the single digit hundreds or fewer for these benchmarks. A recently released dataset captured from the AliveCor ECG monitor contains about 7000 records (Clifford et al., 2017). These records only have annotations for Atrial Fibrillation; all other arrhythmias are grouped into a single bucket. The dataset we develop contains 29,163 unique patients and 14 classes with hundreds of unique examples for the rarest arrhythmias.

Machine learning models based on deep neural networks have consistently been able to approach and often exceed human agreement rates when large annotated datasets are available (Amodei et al., 2016; Xiong et al., 2016; He et al., 2015c). These approaches have also proven to be effective in healthcare applications, particularly in medical imaging where pretrained ImageNet models can be applied (Esteva et al., 2017; Gulshan et al., 2016). We draw on work in automatic speech recognition for processing time-series with deep convolutional neural networks and recurrent neural networks (Hannun et al., 2014; Sainath et al., 2013), and techniques in deep learning to make the optimization of these models tractable (He et al., 2016b;c; Ioffe & Szegedy, 2015).

7. Conclusion

We develop a model which exceeds the cardiologist performance in detecting a wide range of heart arrhythmias from single-lead ECG records. Key to the performance of the model is a large annotated dataset and a very deep convolutional network which can map a sequence of ECG samples

to a sequence of arrhythmia annotations.

On the clinical side, future work should investigate extending the set of arrhythmias and other forms of heart disease which can be automatically detected with high-accuracy from single or multiple lead ECG records. For example we do not detect Ventricular Flutter or Fibrillation. We also do not detect Left or Right Ventricular Hypertrophy, Myocardial Infarction or a number of other heart diseases which do not necessarily exhibit as arrhythmias. Some of these may be difficult or even impossible to detect on a single-lead ECG but can often be seen on a multiple-lead ECG.

Given that more than 300 million ECGs are recorded annually, high-accuracy diagnosis from ECG can save expert clinicians and cardiologists considerable time and decrease the number of misdiagnoses. Furthermore, we hope that this technology coupled with low-cost ECG devices enables more widespread use of the ECG as a diagnostic tool in places where access to a cardiologist is difficult.
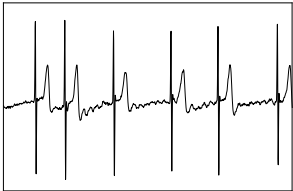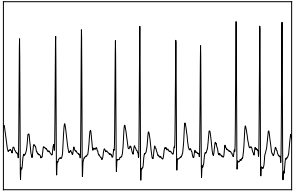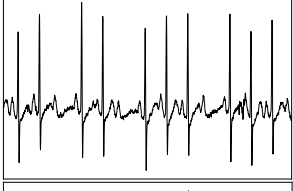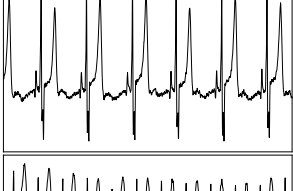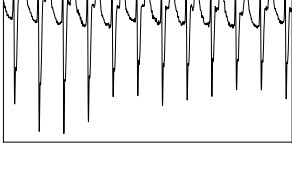
Acknowledgements

References

Amodei, Dario, Anubhai, Rishita, Battenberg, Eric, Case, Carl, Casper, Jared, Catanzaro, Bryan, Chen, JingDong, Chrzanowski, Mike, Coates, Adam, Diamos, Greg, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 173–182, 2016.

Artis, Shane G, Mark, RG, and Moody, GB. Detection of atrial fibrillation using artificial neural networks. In *Computers in Cardiology 1991, Proceedings.*, pp. 173–176. IEEE, 1991.

Clifford, GD, Liu, CY, Moody, B, Lehman, L, Silva, I, Li, Q, Johnson, AEW, and Mark, RG. Af classification from a short single lead ecg recording: The physionet computing in cardiology challenge 2017. 2017.

Coast, Douglas A, Stern, Richard M, Cano, Gerald G, and Briller, Stanley A. An approach to cardiac arrhythmia analysis using hidden markov models. *IEEE Transactions on biomedical Engineering*, 37(9):826–836, 1990.

Dubin, Dale. *Rapid Interpretation of EKG's*. USA: Cover Publishing Company, 1996, 1996.

Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M, and Thrun, Sebastian. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Goldberger, Ary L, Amaral, Luis AN, Glass, Leon, Hausdorff, Jeffrey M, Ivanov, Plamen Ch, Mark, Roger G, Mietus, Joseph E, Moody, George B, Peng, Chung-Kang, and Stanley, H Eugene. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Guglin, Maya E and Thatai, Deepak. Common errors in computer electrocardiogram interpretation. *International journal of cardiology*, 106(2):232–237, 2006.

Gulshan, Varun, Peng, Lily, Coram, Marc, Stumpe, Martin C, Wu, Derek, Narayanaswamy, Arunachalam, Venugopalan, Subhashini, Widner, Kasumi, Madams, Tom, Cuadros, Jorge, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316 (22):2402–2410, 2016.

Hannun, Awni Y., Case, Carl, Casper, Jared, Catanzaro, Bryan, Diamos, Greg, Elsen, Erich, Prenger, Ryan, Satheesh, Sanjeev, Sengupta, Shubho, Coates, Adam, and Ng, Andrew Y. Deep speech: Scaling up end-to-end speech recognition. abs/1412.5567, 2014. URL http://arxiv.org/abs/1412.5567.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015a. URL http://arxiv.org/abs/1502.01852.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015b. URL http://arxiv.org/abs/1512.03385.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015c.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016a. URL http://arxiv.org/abs/1603.05027.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016b.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016c.

Hedén, Bo, Ohlsson, Mattias, Holst, Holger, Mjöman, Mattias, Rittner, Ralf, Pahlm, Olle, Peterson, Carsten, and Edenbrandt, Lars. Detection of frequently overlooked electrocardiographic lead reversals using artificial neural networks. *The American journal of cardiology*, 78 (5):600–604, 1996.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

Laguna, Pablo, Mark, Roger G, Goldberg, A, and Moody, George B. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in Cardiology 1997*, pp. 673–676. IEEE, 1997.

Li, Cuiwei, Zheng, Chongxun, and Tai, Changfeng. Detection of ECG characteristic points using wavelet transforms. *IEEE Transactions on biomedical Engineering*, 42(1):21–28, 1995.

Martínez, Juan Pablo, Almeida, Rute, Olmos, Salvador, Rocha, Ana Paula, and Laguna, Pablo. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering*, 51(4): 570–581, 2004.

Melo, SL, Caloba, LP, and Nadal, J. Arrhythmia analysis using artificial neural network and decimated electrocardiographic data. In *Computers in Cardiology 2000*, pp. 73–76. IEEE, 2000.

Moody, George B and Mark, Roger G. A new method for detecting atrial fibrillation using RR intervals. *Computers in Cardiology*, 10(1):227–230, 1983.

Moody, George B and Mark, Roger G. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

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

Sainath, Tara N, Mohamed, Abdel-rahman, Kingsbury, Brian, and Ramabhadran, Bhuvana. Deep convolutional neural networks for lvcsr. In *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*, pp. 8614–8618. IEEE, 2013.

Shah, Atman P and Rubin, Stanley A. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of electrocardiology*, 40(5):385–390, 2007.

Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Turakhia, Mintu P, Hoang, Donald D, Zimetbaum, Peter, Miller, Jared D, Froelicher, Victor F, Kumar, Uday N, Xu, Xiangyan, Yang, Felix, and Heidenreich, Paul A. Diagnostic utility of a novel leadless arrhythmia monitoring device. *The American journal of cardiology*, 112 (4):520–524, 2013.

Xiong, Wayne, Droppo, Jasha, Huang, Xuedong, Seide, Frank, Seltzer, Mike, Stolcke, Andreas, Yu, Dong, and Zweig, Geoffrey. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

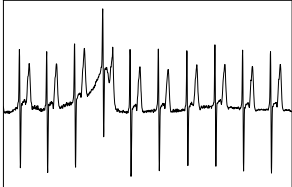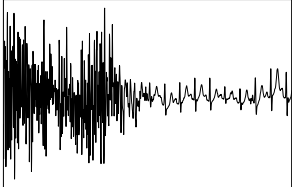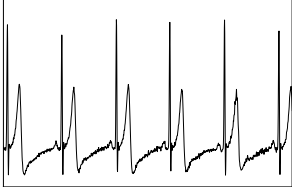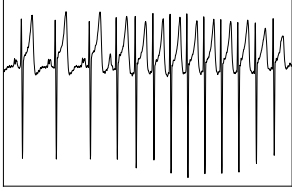| Class	Description	Example	Train + Val Patients	Test Patients
AFIB	Atrial Fibrillation		4638	44
AFL	Atrial Flutter		3805	20
AVB_TYPE2	Second degree AV Block Type 2 (Mobitz II)		1905	28
BIGEMINY	Ventricular Bigeminy		2855	22
CHB	Complete Heart Block		843	26
EAR	Ectopic Atrial Rhythm		2623	22
IVR	Idioventricular Rhythm		1962	34
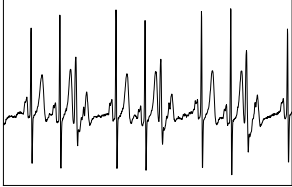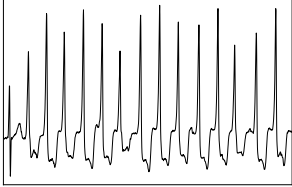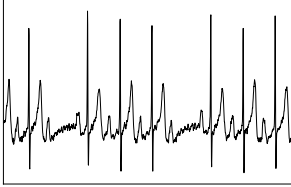
Class	Description	Example	Train + Val Patients	Test Patients
JUNCTIONAL	Junctional Rhythm		2030	36
NOISE	Noise		9940	41
SINUS	Sinus Rhythm		22156	215
SVT	Supraventricular Tachycardia		6301	34
TRIGEMINY	Ventricular Trigeminy		2864	21
VT	Ventricular Tachycardia		4827	17
WENCKEBACH	Wenckebach (Mobitz I)		2051	29

Table 2. A list of all of the rhythm types which the model classifies. For each rhythm we give the label name, a more descriptive name and an example chosen from the training set. We also give the total number of patients with each rhythm for both the training and test sets.