Machine Learning for Healthcare

6.871, HST.956

Lecture 5: Risk stratification (continued) & Physiological time-series

David Sontag

CSAIL

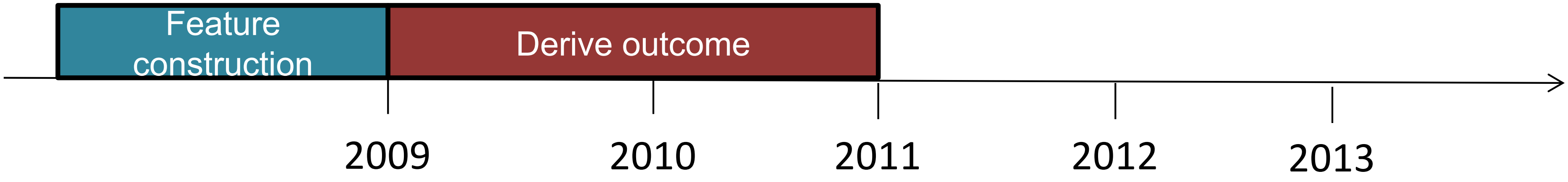
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Outline for today’s class

1. Using ML for risk stratification (continued)
   – Alternative framing: survival modeling
   – Evaluation: metrics, interpretability
2. Physiological time-series: application to detecting irregular heart arrythmias
   – Small data approach
   – Big data approach
   – Current research

Reminder: (One) framing for ML as binary classification



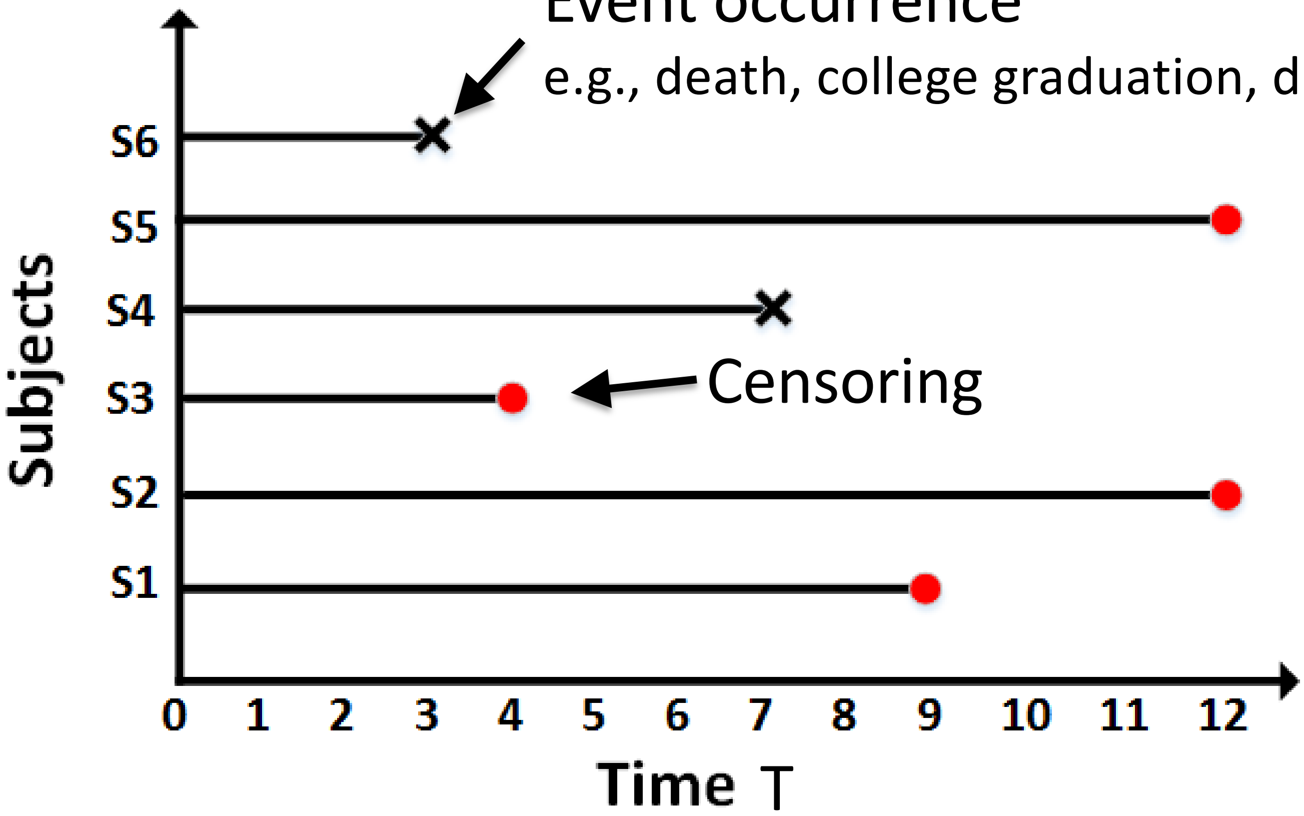
Exclusion criteria:

- Diabetes diagnosis (according to our rule) observed prior to January 1, 2009
- Less than 6 months of enrollment in feature construction window
- Member left health insurance prior to Jan. 1, 2011

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* '16]

Alternative framing: ML with survival models

- Regression (i.e., predict time to event) with right-censored data



Event occurrence
e.g., death, college graduation, diabetes onset

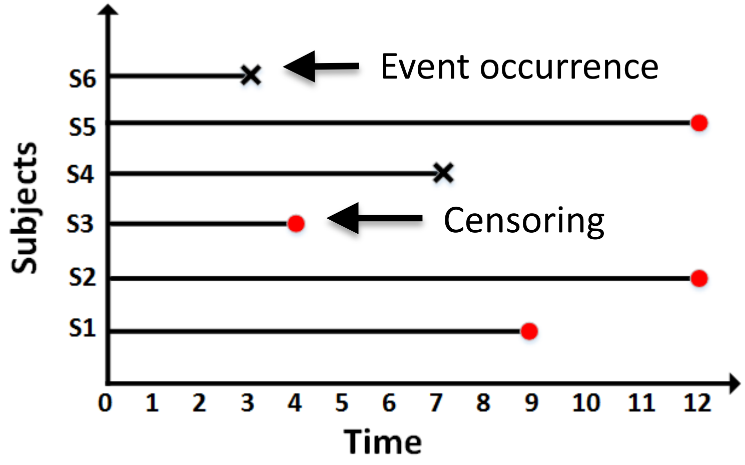
Censoring

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

Alternative framing: ML with survival models

- Advantages over window-based classification
  - More data for training (fewer exclusions)
  - Allows for more fine-grained metrics in evaluation

- Why not just minimize mean-squared error with observed events using least squares linear regression?
  - Time-to-event is non-negative (and non-Gaussian)
  - Naively removed censored events could introduce substantial bias



ML with survival models (more on this later in the semester)

- f(t) = P(t) be the probability of death at time t
- Learn (conditional) survival function: $S(t) = P(T > t) = \int_t^\infty f(x)dx$

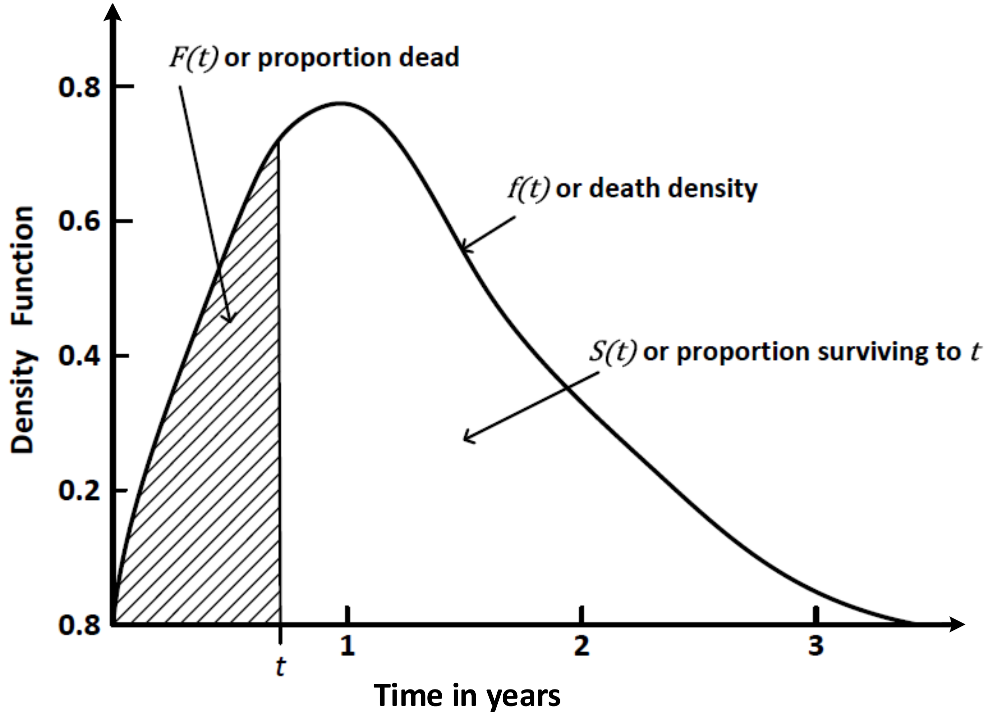


Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]
[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

Outline for today's class

1. Using ML for risk stratification
   - Alternative framing: survival modeling
   - **Evaluation: metrics, interpretability**
2. Physiological time-series: application to detecting irregular heart arrythmias
   - Small data approach
   - Big data approach
   - Current research

“Table 1” – who did this study include?

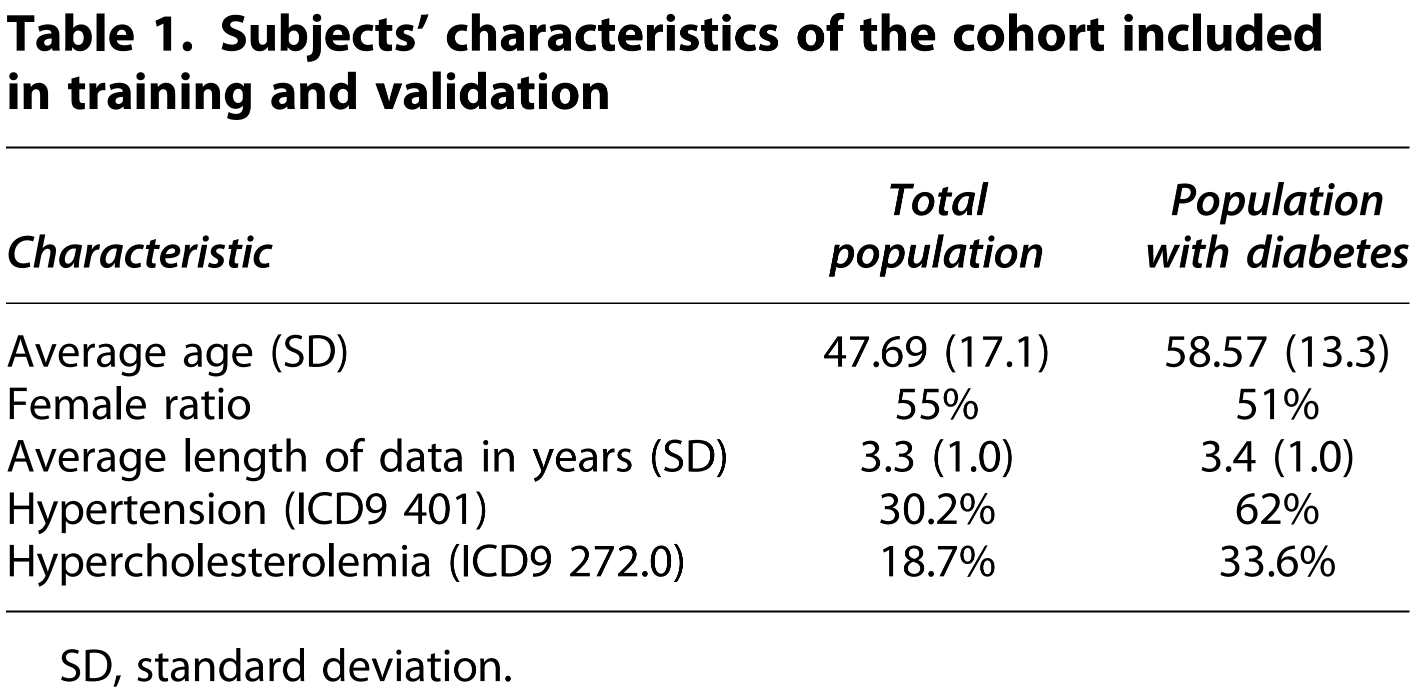


Table 1. Subjects’ characteristics of the cohort included in training and validation

Characteristic	*Total population*	*Population with diabetes*
Average age (SD)	47.69 (17.1)	58.57 (13.3)
Female ratio	55%	51%
Average length of data in years (SD)	3.3 (1.0)	3.4 (1.0)
Hypertension (ICD9 401)	30.2%	62%
Hypercholesterolemia (ICD9 272.0)	18.7%	33.6%

SD, standard deviation.

But what about… Past hospitalizations? Number of years of historical data? Race/ethnicity?

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

“Table 1”, better example

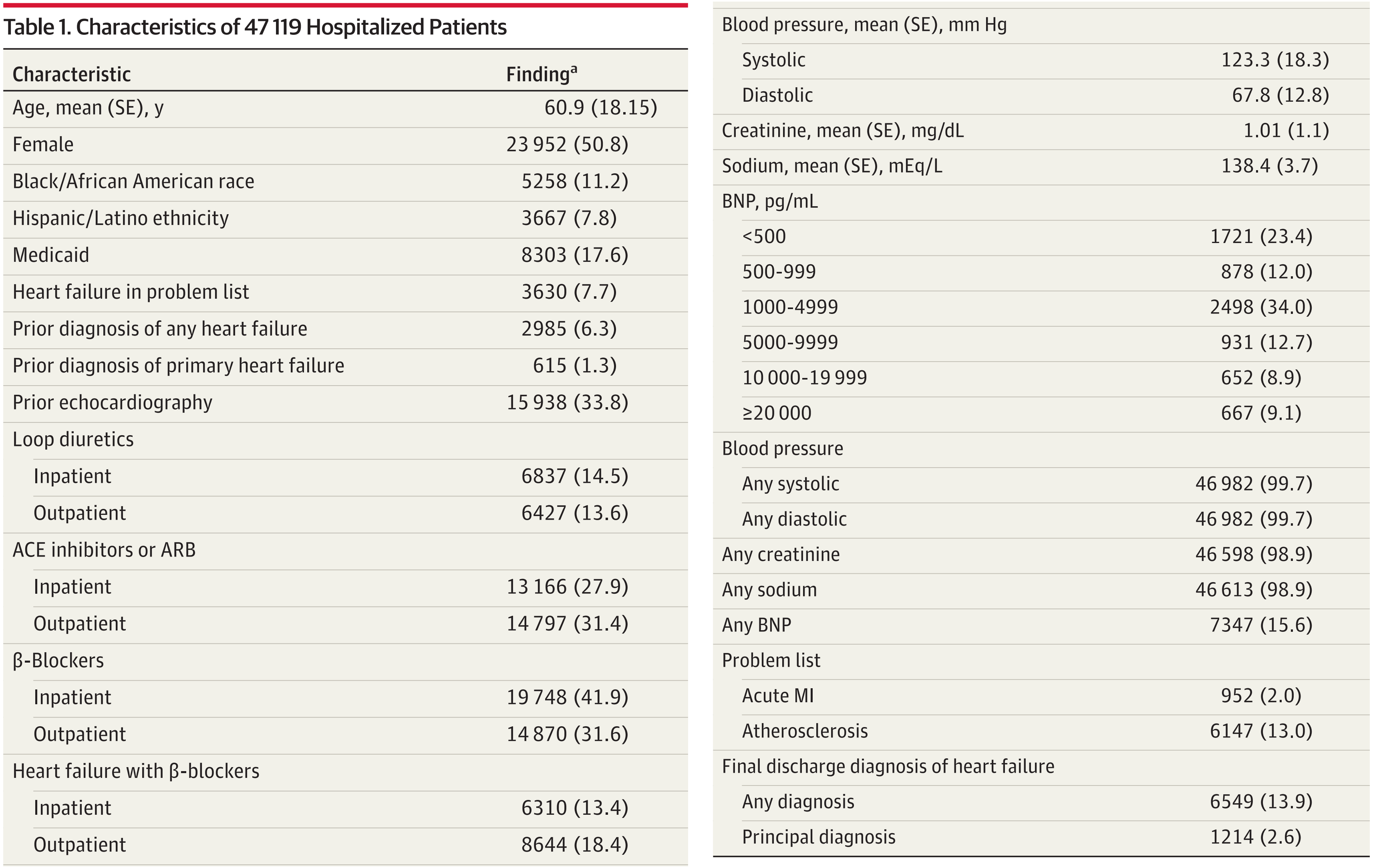


Table 1. Characteristics of 47 119 Hospitalized Patients

Characteristic	Finding[a]
Age, mean (SE), y	60.9 (18.15)
Female	23 952 (50.8)
Black/African American race	5258 (11.2)
Hispanic/Latino ethnicity	3667 (7.8)
Medicaid	8303 (17.6)
Heart failure in problem list	3630 (7.7)
Prior diagnosis of any heart failure	2985 (6.3)
Prior diagnosis of primary heart failure	615 (1.3)
Prior echocardiography	15 938 (33.8)
Loop diuretics	
Inpatient	6837 (14.5)
Outpatient	6427 (13.6)
ACE inhibitors or ARB	
Inpatient	13 166 (27.9)
Outpatient	14 797 (31.4)
β-Blockers	
Inpatient	19 748 (41.9)
Outpatient	14 870 (31.6)
Heart failure with β-blockers	
Inpatient	6310 (13.4)
Outpatient	8644 (18.4)
Blood pressure, mean (SE), mm Hg	
Systolic	123.3 (18.3)
Diastolic	67.8 (12.8)
Creatinine, mean (SE), mg/dL	1.01 (1.1)
Sodium, mean (SE), mEq/L	138.4 (3.7)
BNP, pg/mL	
<500	1721 (23.4)
500-999	878 (12.0)
1000-4999	2498 (34.0)
5000-9999	931 (12.7)
10 000-19 999	652 (8.9)
≥20 000	667 (9.1)
Blood pressure	
Any systolic	46 982 (99.7)
Any diastolic	46 982 (99.7)
Any creatinine	46 598 (98.9)
Any sodium	46 613 (98.9)
Any BNP	7347 (15.6)
Problem list	
Acute MI	952 (2.0)
Atherosclerosis	6147 (13.0)
Final discharge diagnosis of heart failure	
Any diagnosis	6549 (13.9)
Principal diagnosis	1214 (2.6)

[Blecker et al., Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data, JAMA Cardiology 2016]

Logistic regression with L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*

$$\min_w \sum_i \ell(x_i, y_i; w) + \lambda ||w||_1 \qquad ||\vec{w}||_1 = \sum_d |w_d|$$

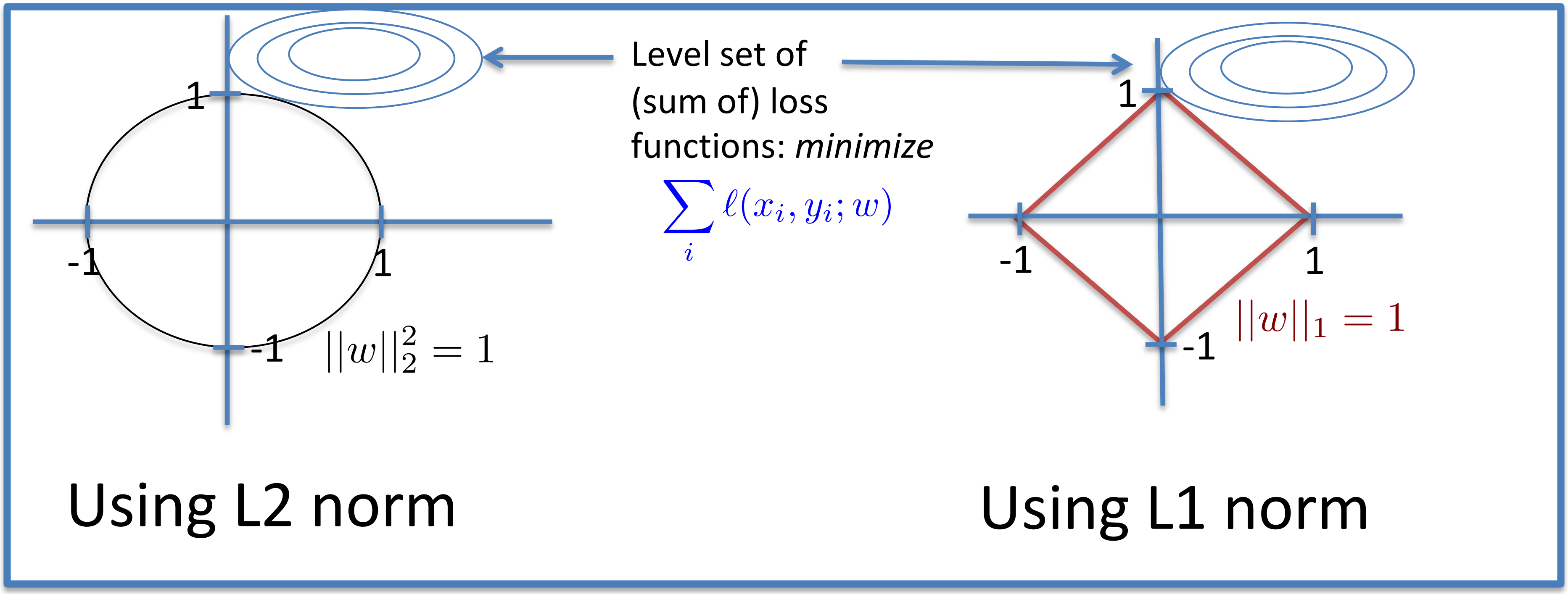
instead of

$$\min_w \sum_i \ell(x_i, y_i; w) + \lambda ||w||_2^2 \qquad ||\vec{w}||_2^2 = \sum_d w_d^2$$

- Let's understand why…

Logistic regression with L1 regularization

- Penalizing the L1 norm of the weight vector leads to *sparse* (read: many 0's) solutions for *w.*



1
-1
1
-1
$||w||_2^2 = 1$
Level set of (sum of) loss functions: *minimize*
$\sum_i \ell(x_i, y_i; w)$
1
-1
1
-1
$||w||_1 = 1$

Using L2 norm

Using L1 norm

Logistic regression with L1 regularization

- 769 variables have non-zero weight. Look at most positive & most negative
- Positively weighted diagnosis codes include

Pituitary dwarfism (253.3), Hepatomegaly(789.1), Chronic Hepatitis C (070.54), Hepatitis (573.3), Calcaneal Spur(726.73), Thyrotoxicosis without mention of goiter(242.90), Sinoatrial Node dysfunction(427.81), Acute frontal sinusitis (461.1), Hypertrophic and atrophic conditions of skin(701.9), Irregular menstruation(626.4), …

- Positively weighted laboratory features include

Albumin/Globulin (Increasing -Entire history), Urea nitrogen/Creatinine - (high - Entire History), Specific gravity (Increasing, Past 2 years), Bilirubin (high -Past 2 years), …

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

Interpreting high-dimensional linear models

- How do we interpret such high dimensional models…?
- A useful trick to build intuition: use higher value of λ (i.e., more regularization) than needed

$$\min_{w} \sum_{i} \ell(x_i, y_i; w) + \lambda ||w||_1$$

- What will the effect be?
- Intuition: often many predictive yet highly correlated features. Selects a representative set which still performs well

Features selected using model learned with more L1 regularization

History of Disease
Impaired Fasting Glucose (Code 790.21)
Abnormal Glucose NEC (790.29)
Hypertension (401)
Obstructive Sleep Apnea (327.23)
Obesity (278)
Abnormal Blood Chemistry (790.6)
Hyperlipidemia (272.4)
Shortness Of Breath (786.05)
Esophageal Reflux (530.81)

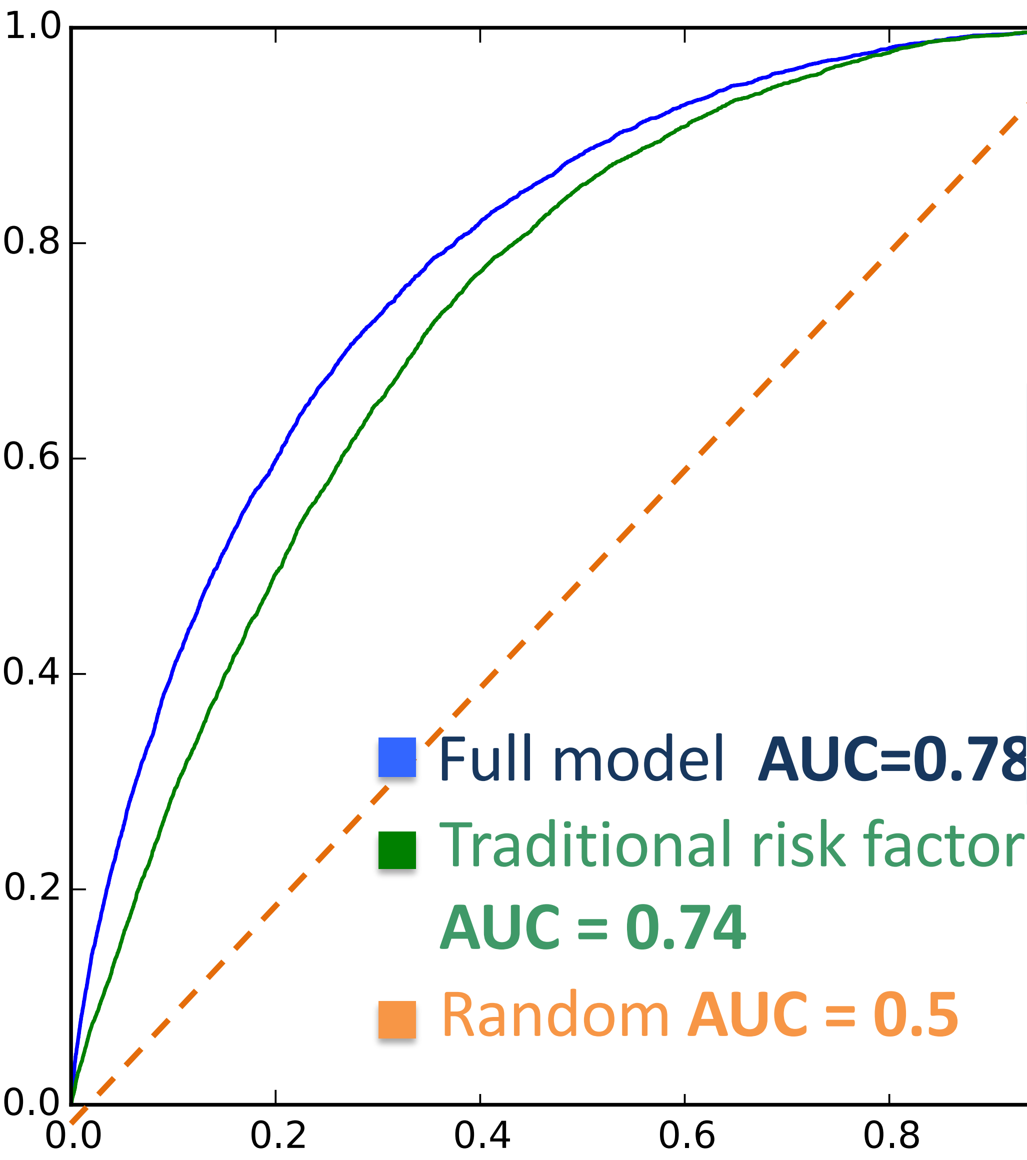
Top Lab Factors
Hemoglobin A1c /Hemoglobin.Total (High - past 2 years)
Glucose (High- Past 6 months)
Cholesterol.In VLDL (Increasing - Past 2 years)
Potassium (Low - Entire History)
Cholesterol.Total/Cholesterol.In HDL (High - Entire History)
Erythrocyte mean corpuscular hemoglobin concentration -(Low - Entire History)
Eosinophils (High - Entire History)
Glomerular filtration rate/1.73 sq M.Predicted (Low -Entire History)
Alanine aminotransferase (High Entire History)

[Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. *Big Data.* ‘16]

Debugging ML setup through model interpretation

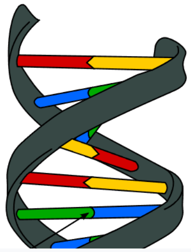
- Suppose a highly weighted positive feature is for “injection of aflibercept”, a treatment for diabetic macular edema. What could we infer?
- Suppose we see many features for flu vaccines with high positive and negative weights. Looking up the NDC code for one of them, we see it is “influenza A virus A/Hong Kong/4801/2014 (H3N2) antigen 0.03 MG”. What could we infer?
- Note, these would have been much harder to diagnose using the deep model

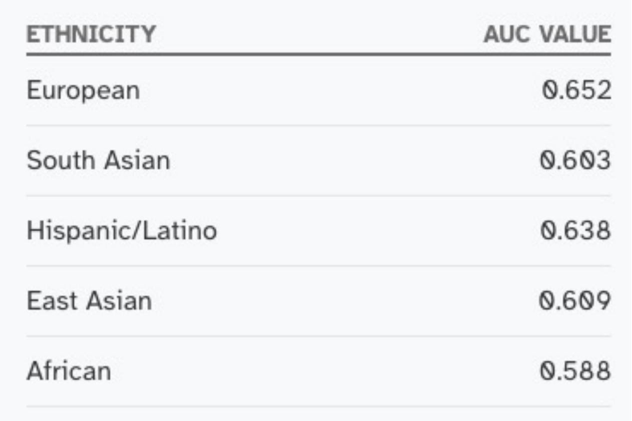
Receiver-operator characteristic curve



Diabetes 1-year gap

Recall the 23andme results:





ETHNICITY	AUC VALUE
European	0.652
South Asian	0.603
Hispanic/Latino	0.638
East Asian	0.609
African	0.588

DNA
Deoxyribonucleic acid

Comparison with the deep models

We consider three prediction tasks. **Over a horizon of 3 to 9 months into the future, predict:**

End of Life
(EOL)

Surgical
Procedures
(Surgery)

Likelihood of
Hospitalization
(LoH)

Evaluate using de-identified dataset of ~120K Medicare Advantage members

Kodialam et al., *Deep Contextual Clinical Prediction with Reverse Distillation*, AAAI ‘21

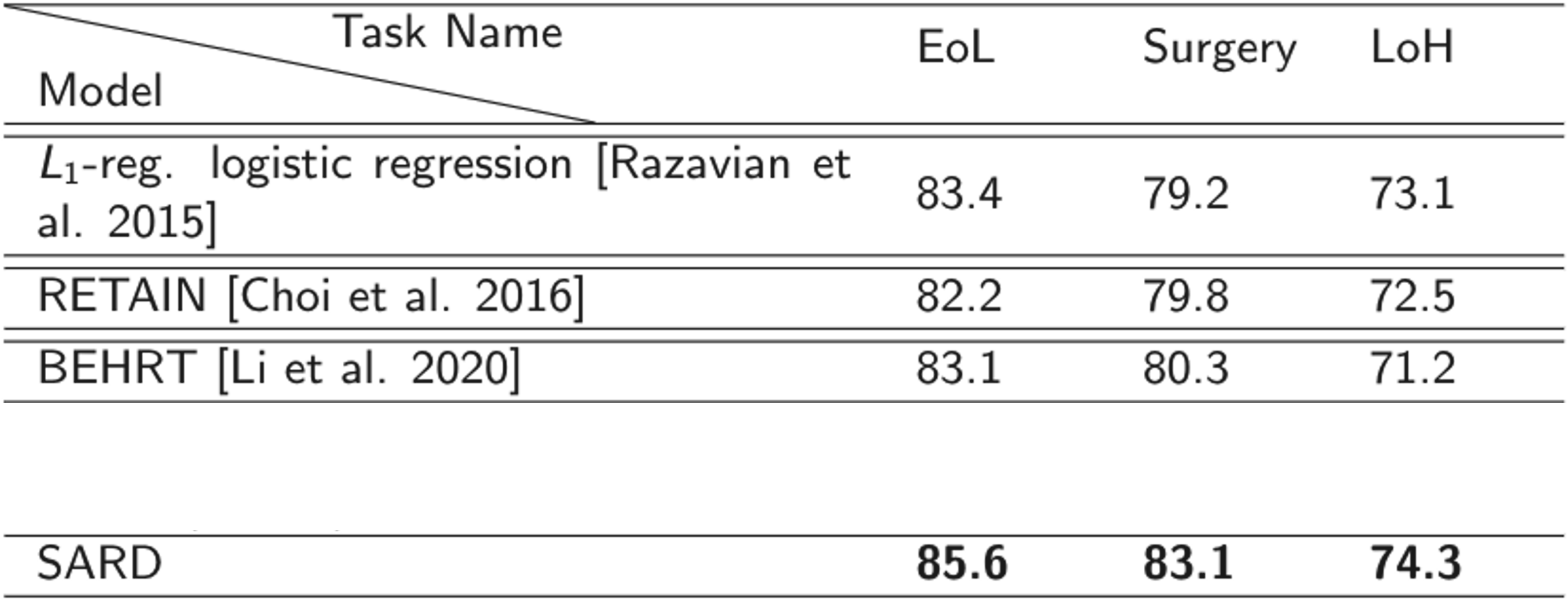
ML methods that we compare

- **SARD** (Kodialam et al. 2021)
- **BEHRT** (Li et al. 2020): another transformer-based neural network for claims data
- **RETAIN** (Choi et al. 2016): a recurrent neural network designed with interpretability in mind
- **Windowed linear model** (Razavian et al. 2015)

Kodialam et al., *Deep Contextual Clinical Prediction with Reverse Distillation*, AAAI ‘21

Results on the 3 prediction tasks

AUC-ROC scores on test set



Task Name Model	EoL	Surgery	LoH
L_1-reg. logistic regression [Razavian et al. 2015]	83.4	79.2	73.1
RETAIN [Choi et al. 2016]	82.2	79.8	72.5
BEHRT [Li et al. 2020]	83.1	80.3	71.2
SARD	**85.6**	**83.1**	**74.3**

SARD uses “reverse distillation” (RD) for pre-training (see Kodialam et al. ‘21)

Kodialam et al., *Deep Contextual Clinical Prediction with Reverse Distillation*, AAAI ‘21

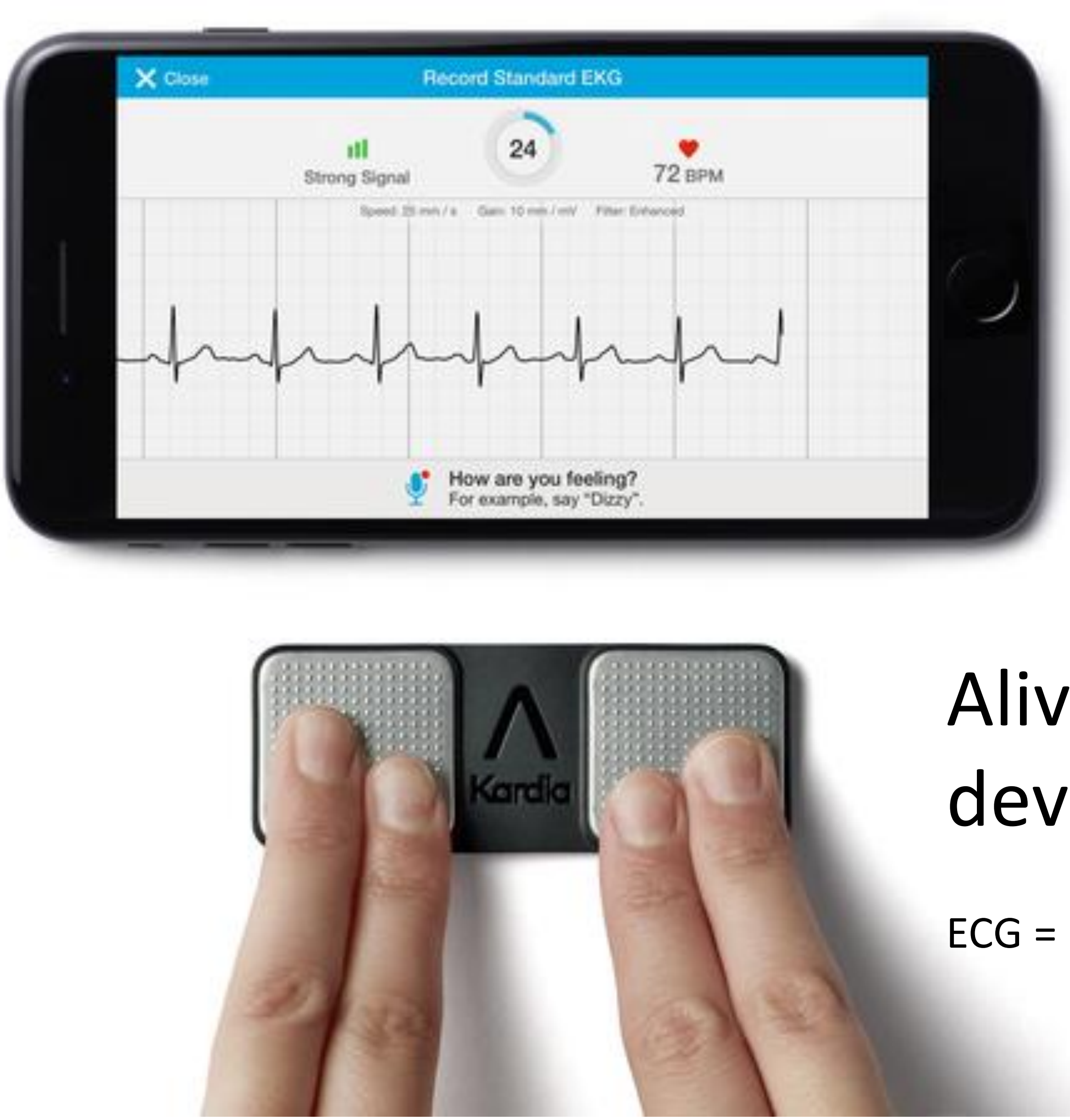
Closing reflections for risk stratification

- How can we build models that work with multi-modal data?
  - Multiple choices for neural network architectures
  - Will often be missing one or more modalities
- How do we choose which target to predict? Has implications for health equity
- What is a “good” result? Depends on the use case – high PPV for targeting interventions, high NPV (negative predictive value) for screening

Outline for today's class

1. Using ML for risk stratification
   - Alternative framing: survival modeling
   - Evaluation: metrics, interpretability
2. **Physiological time-series: application to detecting irregular heart arrythmias**
   - Small data approach
   - Big data approach
   - Current research

Detecting atrial fibrillation

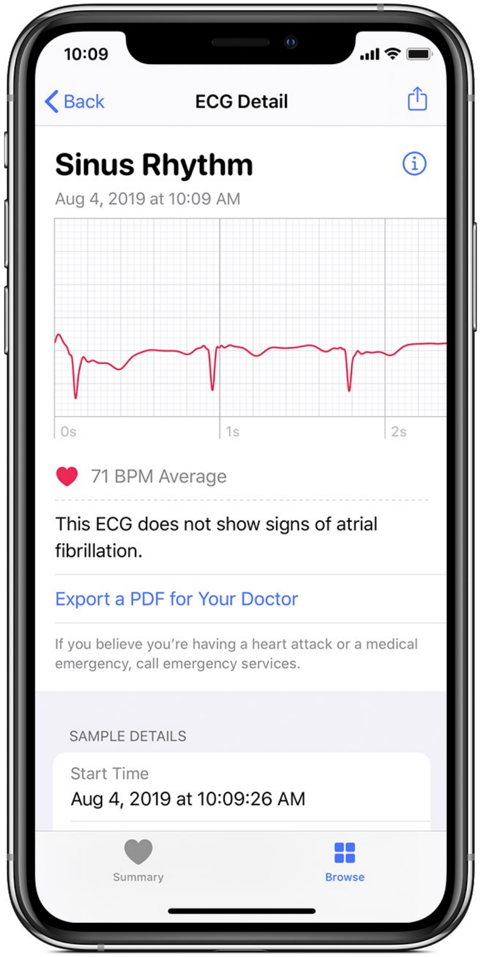

Close
Record Standard EKG
Strong Signal
24
72 BPM
Speed: 25 mm/s Gain: 10 mm/mV Filter: Enhanced
How are you feeling?
For example, say "Dizzy".
Kardia

AliveCore ECG device

ECG = electrocardiogram

Detecting atrial fibrillation

Apple Watch



Sinus Rhythm

Aug 4, 2019 at 10:09 AM

71 BPM Average

This ECG does not show signs of atrial fibrillation.

Export a PDF for Your Doctor

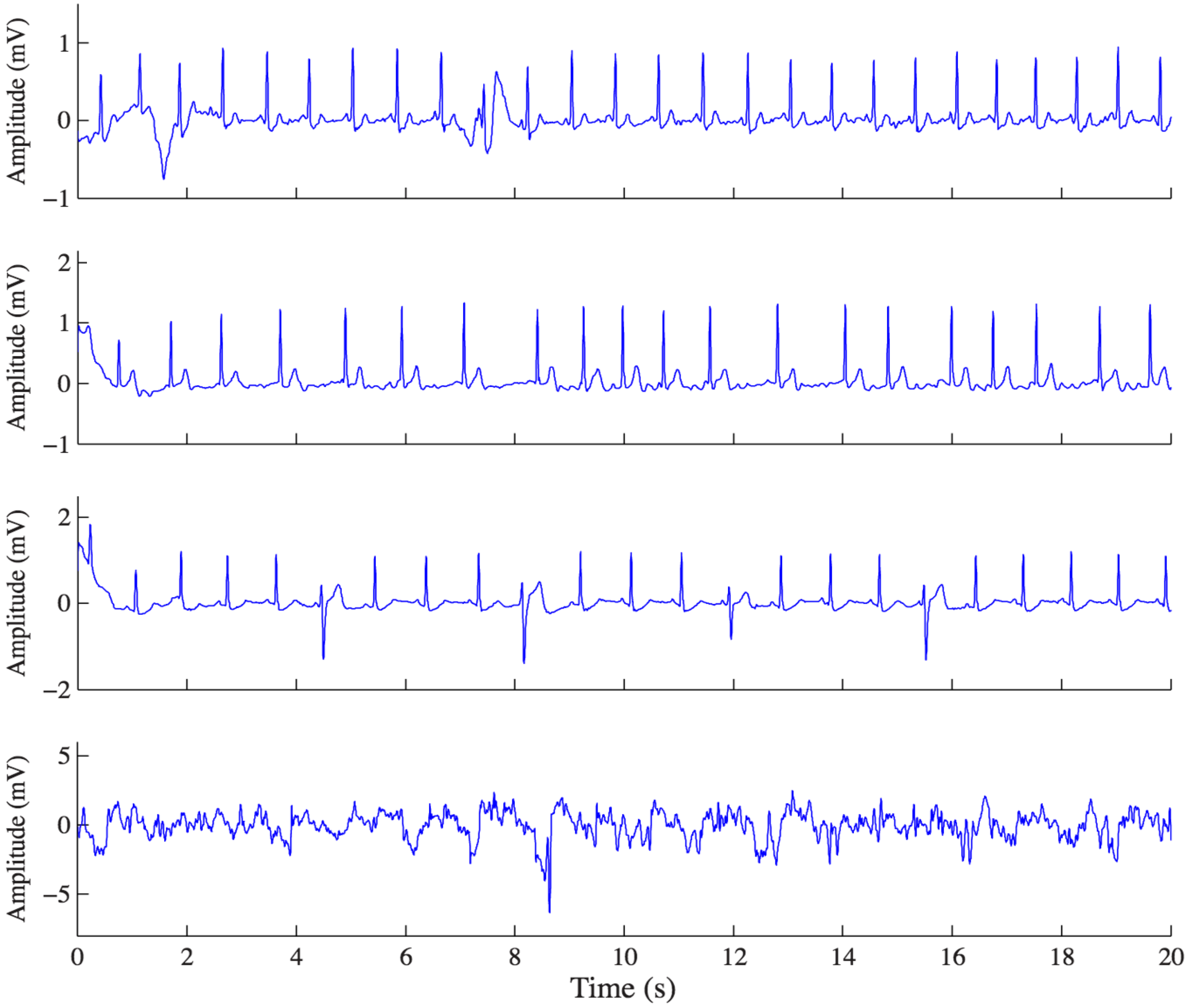
If you believe you're having a heart attack or a medical emergency, call emergency services.

SAMPLE DETAILS

Start Time

Aug 4, 2019 at 10:09:26 AM

What type of heart rhythm?



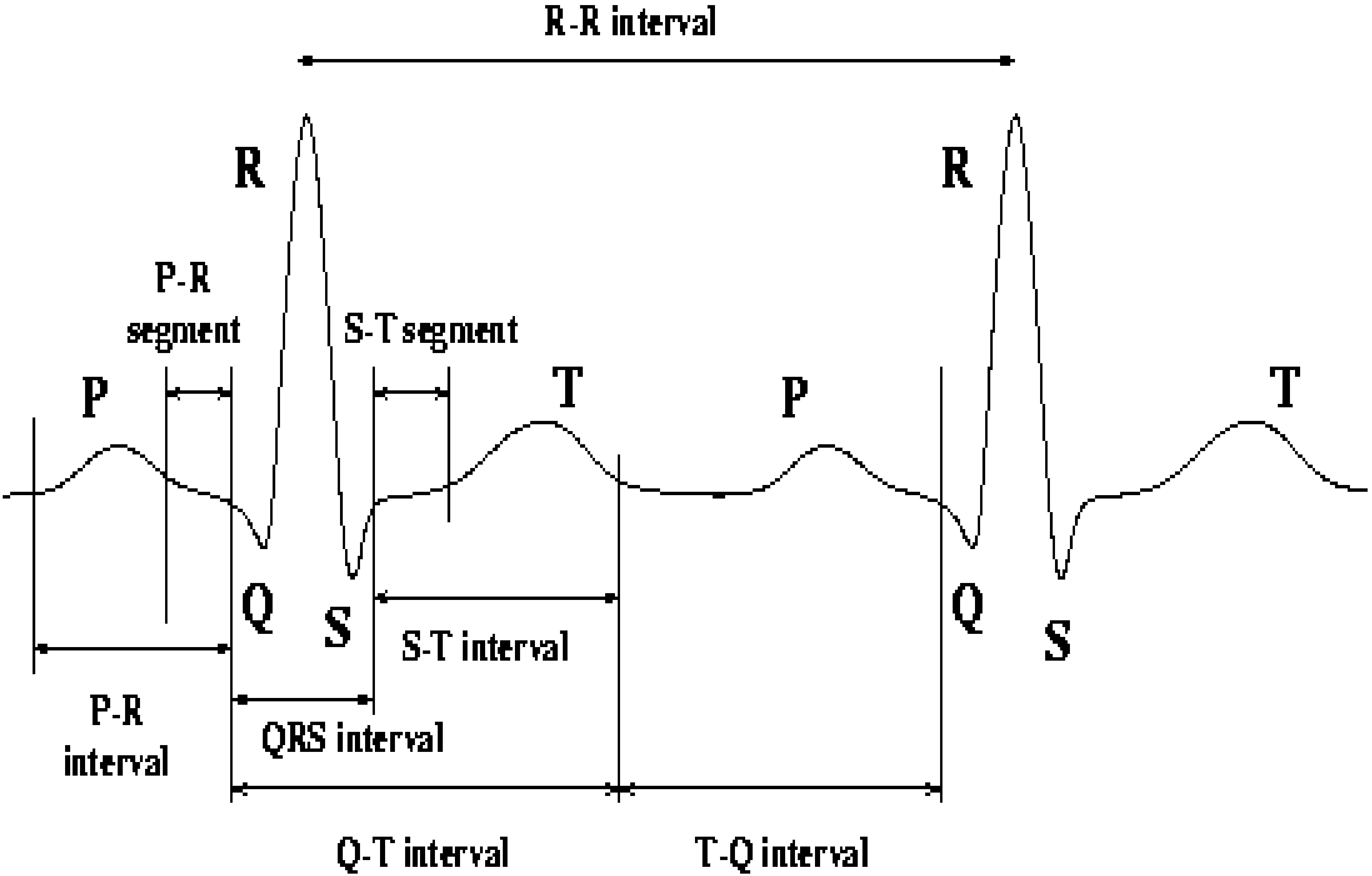
Normal rhythm

AF rhythm

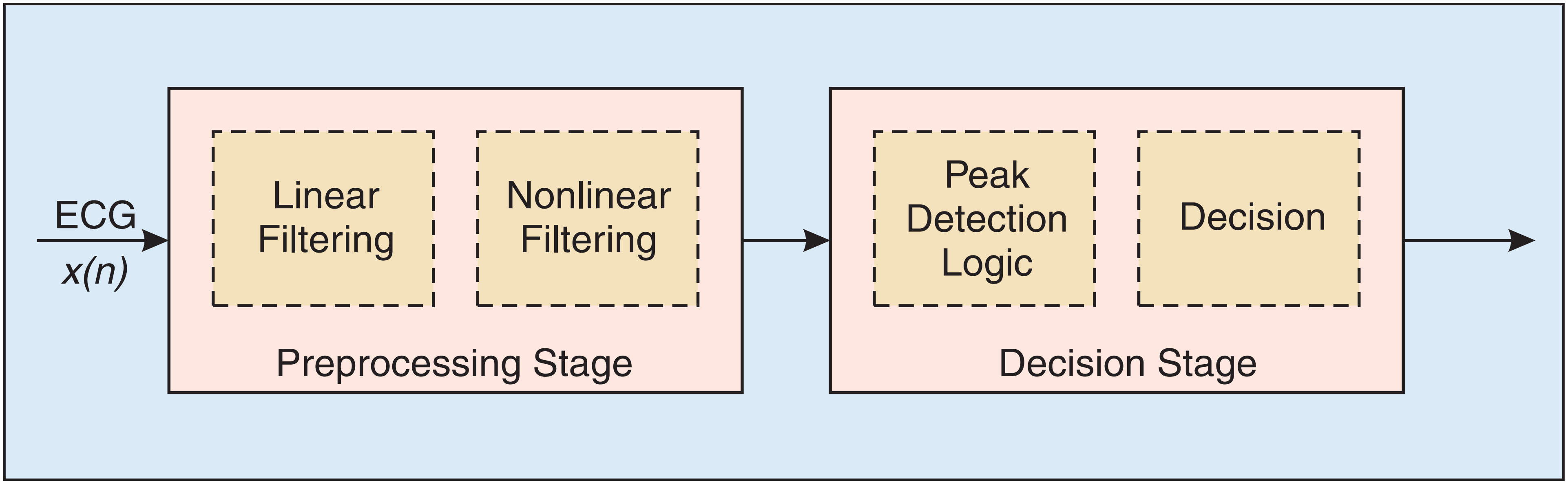
Other rhythm

Noisy recording

[Clifford, Liu, Moody, Mark. PhysioNet Computing in Cardiology Challenge 2017]

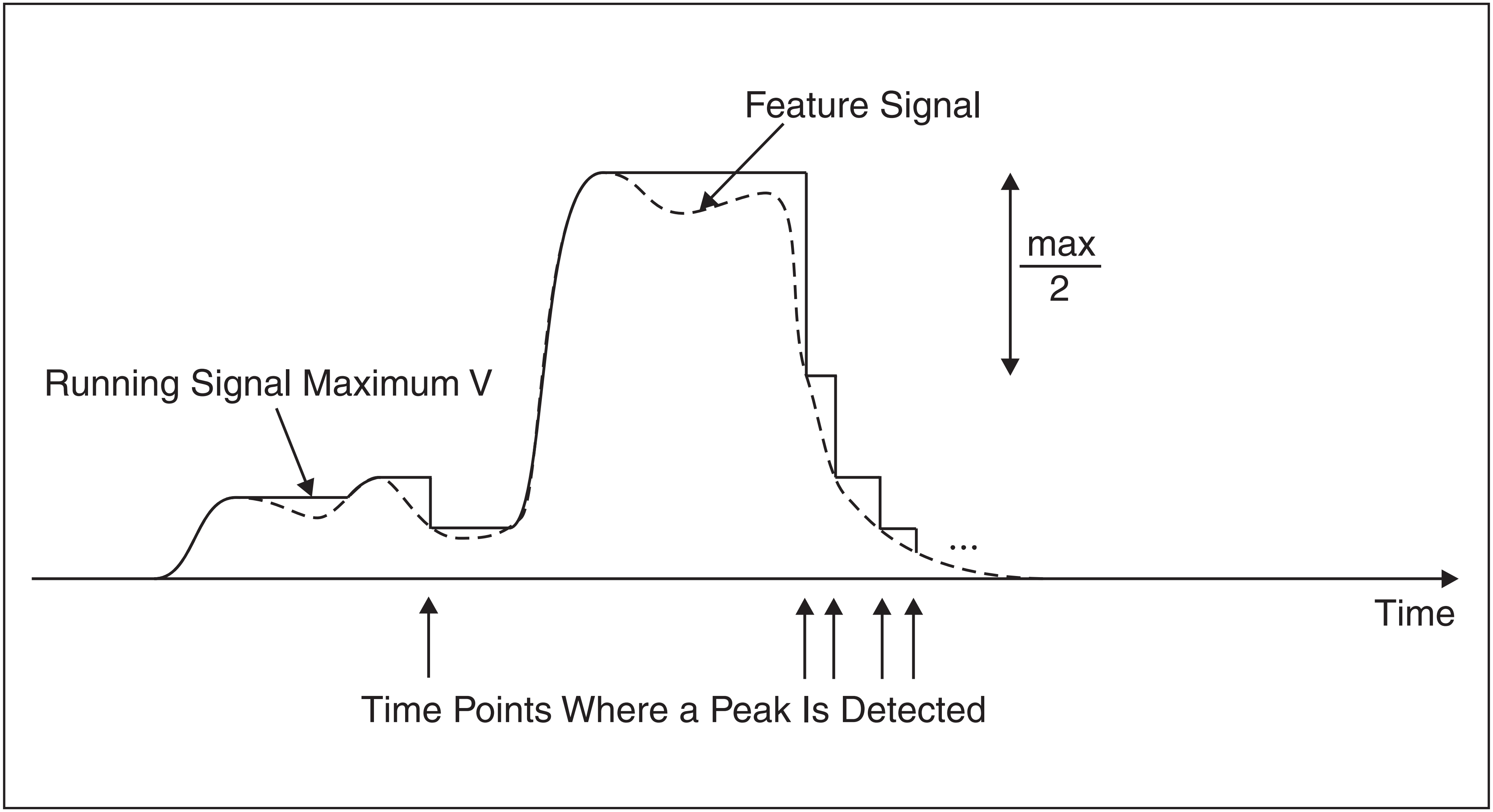

R-R interval
R
R
P-R
segment
S-T segment
P
T
P
T
Q
S
S-T interval
Q
S
P-R
interval
QRS interval
Q-T interval
T-Q interval

Traditional approach



2. Common structure of the QRS detectors.

[Kohler, Hennig, Orglmeister. The Principles of Software QRS Detection, IEEE Engineering in Medicine & Biology, 2002]



3. Peak detector proposed in [41].

[Kohler, Hennig, Orglmeister. The Principles of Software QRS Detection, IEEE Engineering in Medicine & Biology, 2002]

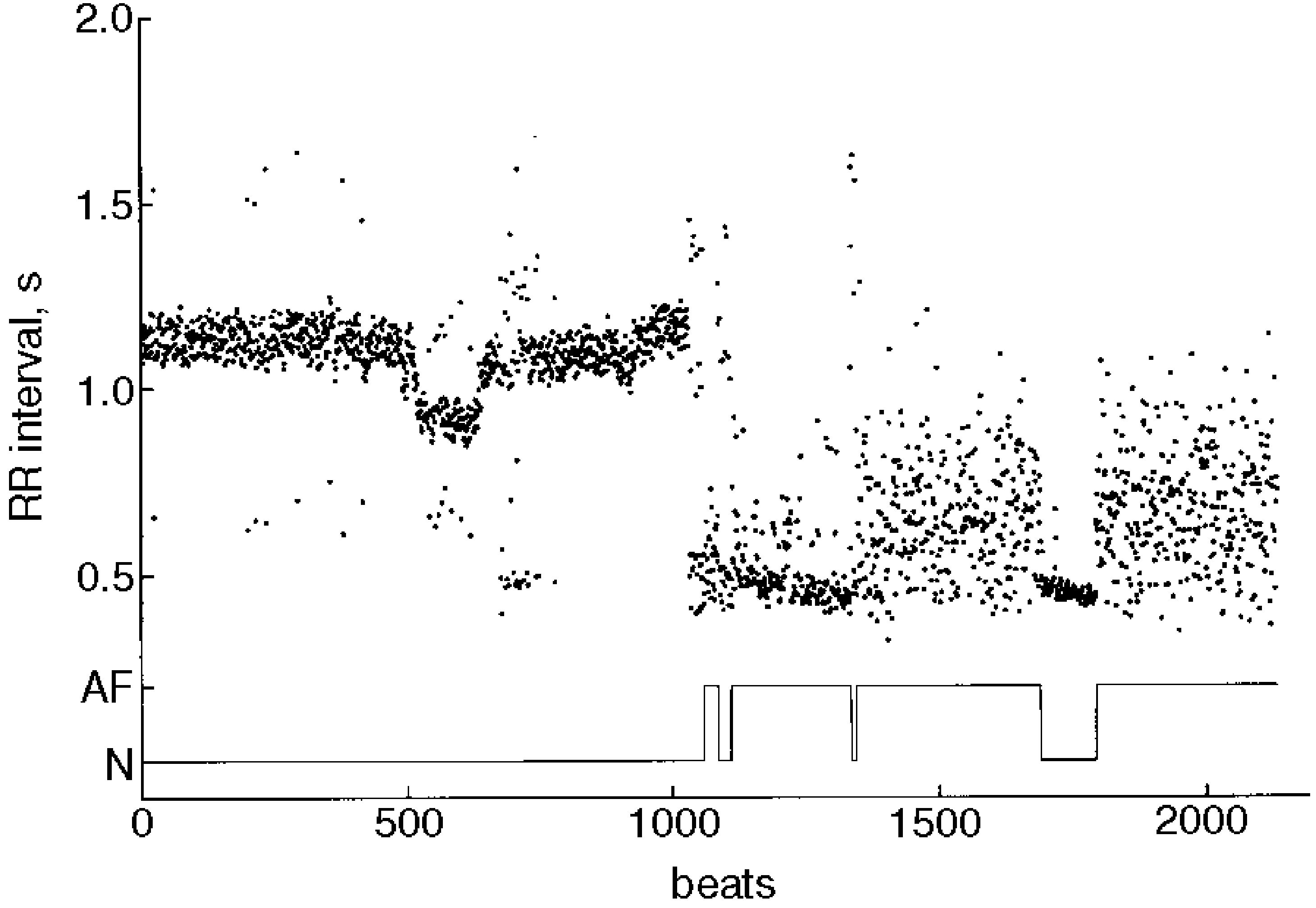


Fig. 1 *Time series showing RR intervals from subject 202 from MIT-BIH arrhythmia database. (——) Assessment of atrial fibrillation (AF) or non-atrial fibrillation (N) as reported in database*

[Tateno & Glass, Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals. MBEC, 2001]

COMPUTERS AND BIOMEDICAL RESEARCH 4, 385-392 (1970)

Cardiac Arrhythmia Classification: A Heart-Beat Interval-Markov Chain Approach *

WILL GERSCH,† DAVID M. EDDY,‡ AND EUGENE DONG, JR.§

Division of Cardiovascular Surgery, Department of Surgery, Stanford University Medical Center, Stanford, California 94305

Received March 2, 1970

A sequence of heart-beat intervals (R-R wave intervals) is automatically transformed into a three-symbol Markov chain sequence. For convenience the symbols used may be thought of as S-R-L for short, regular, and long heart-beat intervals, respectively. The probability that the observed sequence was generated by each of a set of prototype models characteristic of different cardiac disorders is computed. That prototype corresponding to the largest probability of observed sequence generation is designated as the disorder. This procedure is the equivalent of Kullback's classification by the minimization of directed divergence procedure.

In a preliminary experiment primarily using data sequences of 100 heart-beat intervals, 35 different known cases were automatically classified into six cardiac disorders without error. The disorders considered were atrial fibrillation, APC and VPC, bigeminy, sinus tachycardia with occasional bigeminy. sinus tachycardia, and ventricular tachycardia.

An automatic procedure to classify cardiac arrhythmias using a Markov chain interpretation of heart-beat interval data is reported. A sequence of heart-beat

Proceedings Computers in Cardiology (1991)

Detection of Atrial Fibrillation Using Artificial Neural Networks

SG Artis, RG Mark, GB Moody

Harvard-MIT
Division of Health Sciences and Technology, Cambridge, MA

Abstract

Artificial neural networks (ANNs) were used as pattern detectors to detect atrial fibrillation (AF) in the MIT-BIH Arrhythmia Database. ECG data was represented using generalized interval transition matrices, as in Markov model AF detectors[1]. A training file was developed, using these transition matrices, for a back-propagation ANN. This file consisted of approximately 15 minutes each of AF and non-AF data. The ANN was succesfully trained using this data. Three standard databases were used to test network performance. Post-processing of the ANN output yielded an AF sensitivity of 92.86% *and an AF positive predictive accuracy of* 92.34%.

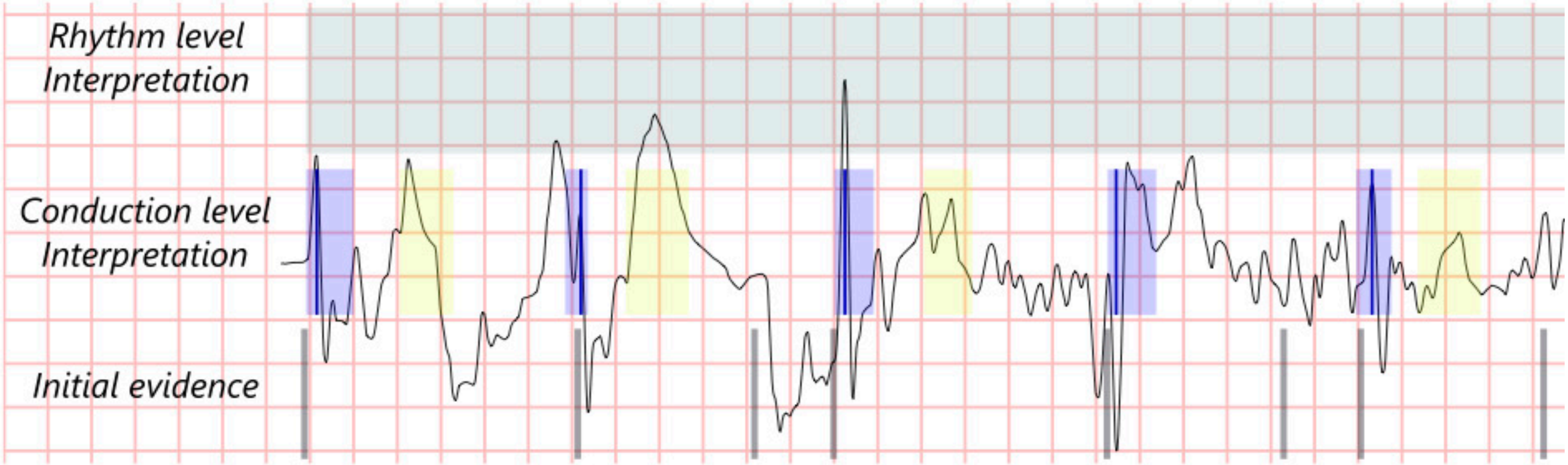
1 Introduction

on R-R interval sequences using a variety of statistical methods [1] but there is room for improvement in these techniques.

Pattern classifiers exist in many forms, and artificial neural networks (ANNs) represent an important subset of these classifiers. ANNs are attractive for solving pattern recognition problems because few assumptions about the underlying data need to be made. The task of the operator of an ANN is to separate the data into subsets. The network will be able classify these subsets according to type as long as they are distinct. Neural network training requires appropriate training data, pre-processing and post-processing algorithms, an appropriate network topology, and a training algorithm, as well as evaluation databases. This document will present the design and evaluation of a technique which detects AF in the presence of other cardiac arrhythmias using a backpropagation artificial neural network.

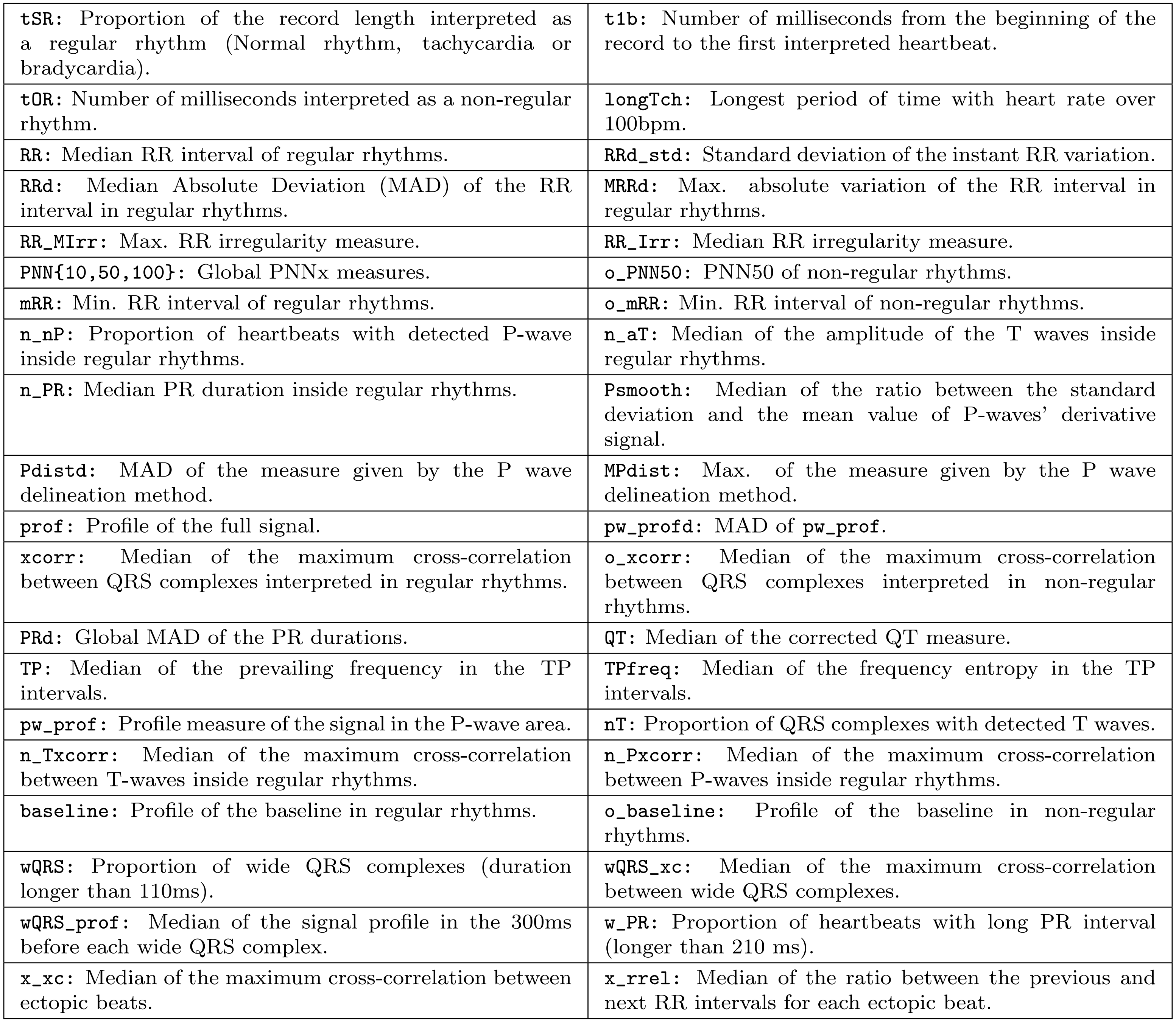
Winning approach in 2017 Physionet challenge

- Training data: ~8500 ECGs
- Best algorithms use a combination of expert-derived features and machine learning



[Teijeiro, Garcia, Castro, Felix. arXiv:1802.05998, 2018]

Table 1: Set of features used to train the global classifier



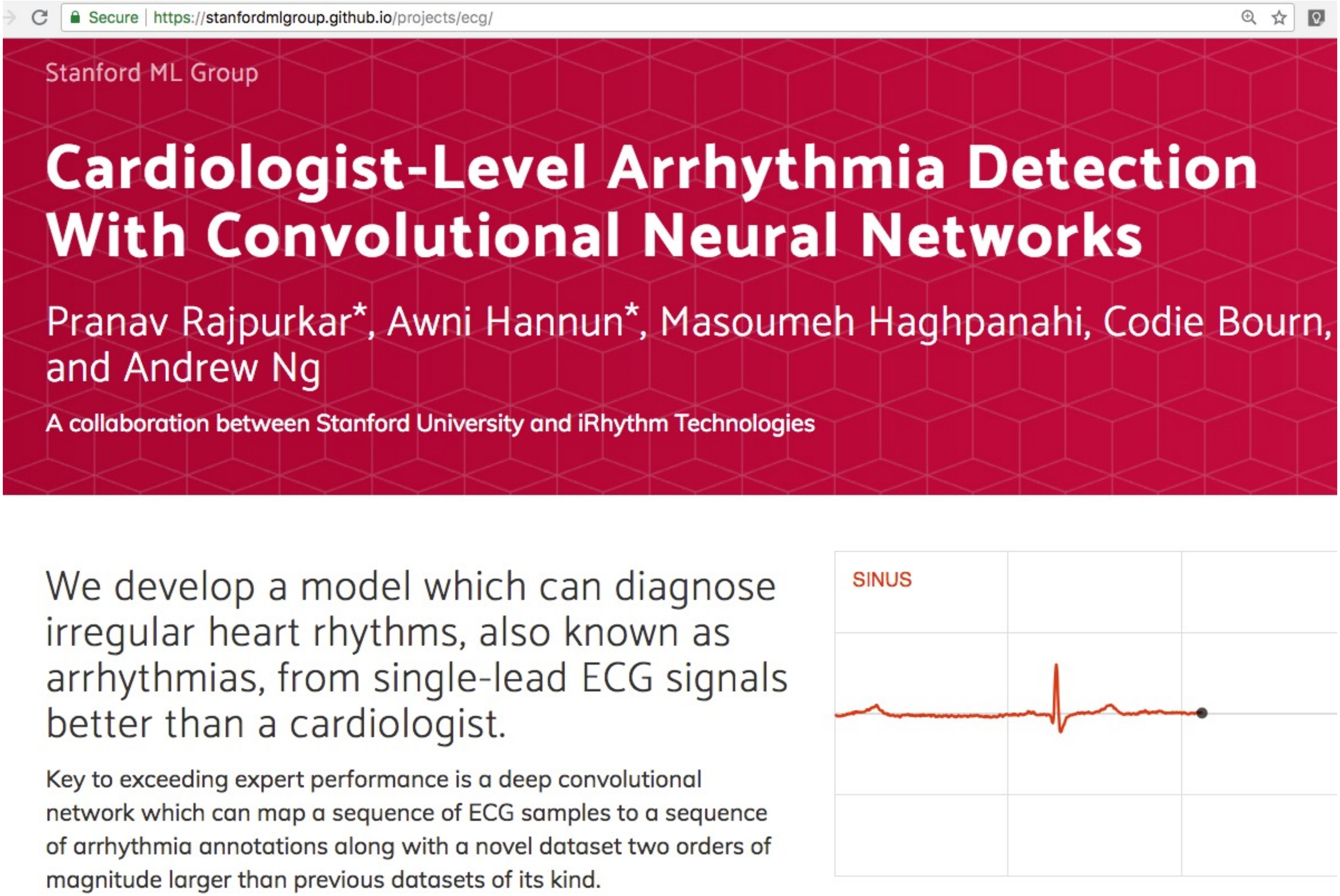
tSR: Proportion of the record length interpreted as a regular rhythm (Normal rhythm, tachycardia or bradycardia).	t1b: Number of milliseconds from the beginning of the record to the first interpreted heartbeat.
tOR: Number of milliseconds interpreted as a non-regular rhythm.	longTch: Longest period of time with heart rate over 100bpm.
RR: Median RR interval of regular rhythms.	RRd_std: Standard deviation of the instant RR variation.
RRd: Median Absolute Deviation (MAD) of the RR interval in regular rhythms.	MRRd: Max. absolute variation of the RR interval in regular rhythms.
RR_MIrr: Max. RR irregularity measure.	RR_Irr: Median RR irregularity measure.
PNN{10,50,100}: Global PNNx measures.	o_PNN50: PNN50 of non-regular rhythms.
mRR: Min. RR interval of regular rhythms.	o_mRR: Min. RR interval of non-regular rhythms.
n_nP: Proportion of heartbeats with detected P-wave inside regular rhythms.	n_aT: Median of the amplitude of the T waves inside regular rhythms.
n_PR: Median PR duration inside regular rhythms.	Psmooth: Median of the ratio between the standard deviation and the mean value of P-waves' derivative signal.
Pdistd: MAD of the measure given by the P wave delineation method.	MPdist: Max. of the measure given by the P wave delineation method.
prof: Profile of the full signal.	pw_profd: MAD of pw_prof.
xcorr: Median of the maximum cross-correlation between QRS complexes interpreted in regular rhythms.	o_xcorr: Median of the maximum cross-correlation between QRS complexes interpreted in non-regular rhythms.
PRd: Global MAD of the PR durations.	QT: Median of the corrected QT measure.
TP: Median of the prevailing frequency in the TP intervals.	TPfreq: Median of the frequency entropy in the TP intervals.
pw_prof: Profile measure of the signal in the P-wave area.	nT: Proportion of QRS complexes with detected T waves.
n_Txcorr: Median of the maximum cross-correlation between T-waves inside regular rhythms.	n_Pxcorr: Median of the maximum cross-correlation between P-waves inside regular rhythms.
baseline: Profile of the baseline in regular rhythms.	o_baseline: Profile of the baseline in non-regular rhythms.
wQRS: Proportion of wide QRS complexes (duration longer than 110ms).	wQRS_xc: Median of the maximum cross-correlation between wide QRS complexes.
wQRS_prof: Median of the signal profile in the 300ms before each wide QRS complex.	w_PR: Proportion of heartbeats with long PR interval (longer than 210 ms).
x_xc: Median of the maximum cross-correlation between ectopic beats.	x_rrel: Median of the ratio between the previous and next RR intervals for each ectopic beat.

[Teijeiro, Garcia, Castro, Felix. arXiv:1802.05998, 2018]

Not enough data for deep learning? Wrong architectures?

"However, the fact that a standard random forest with well chosen features performed as well as more complex approaches, indicates that perhaps a set of 8,528 training patterns was not enough to give the more complex approaches an advantage. With so many parameters and hyperparameters to tune, the search space can be enormous and significant overtraining was seen…"

[Clifford et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge, Computing in Cardiology 2017]



Stanford ML Group

Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks

Pranav Rajpurkar*, Awni Hannun*, Masoumeh Haghpanahi, Codie Bourn, and Andrew Ng

A collaboration between Stanford University and iRhythm Technologies

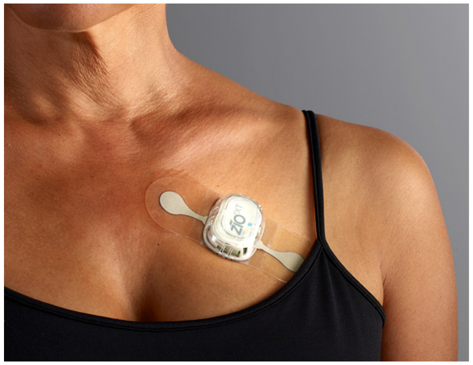
We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist.

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.

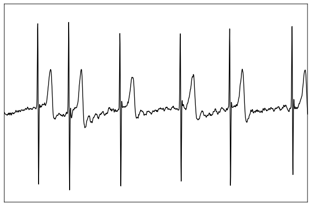
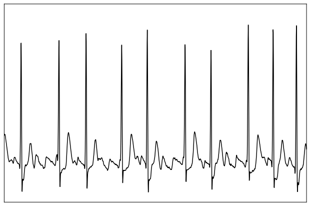
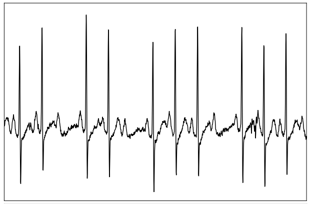
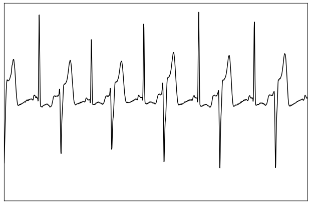
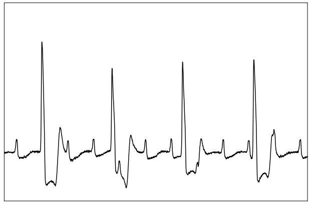
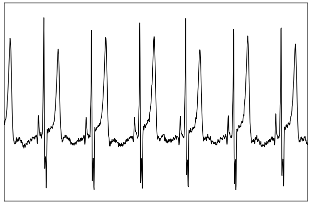
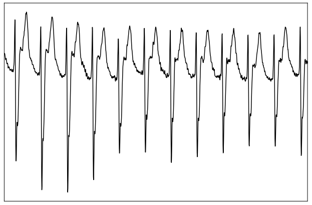
[Rajpurkar et al., arXiv:1707.01836, 2017; Hannun et al. Nature Medicine ‘19]

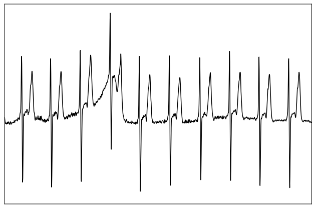
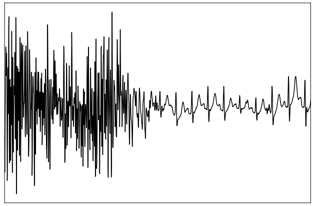
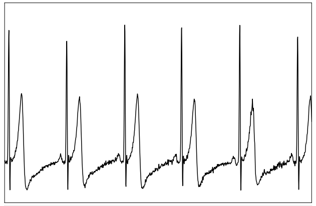
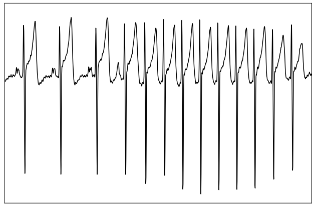
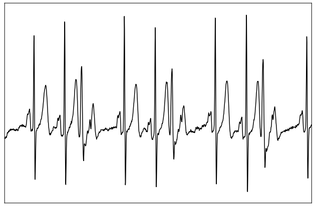
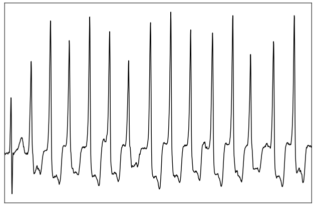
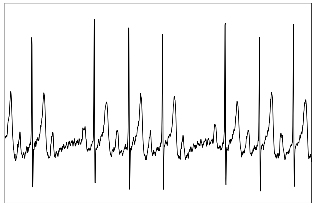
Differences with previous work

- Sensor is a Zio patch – conceivably much less noisy:



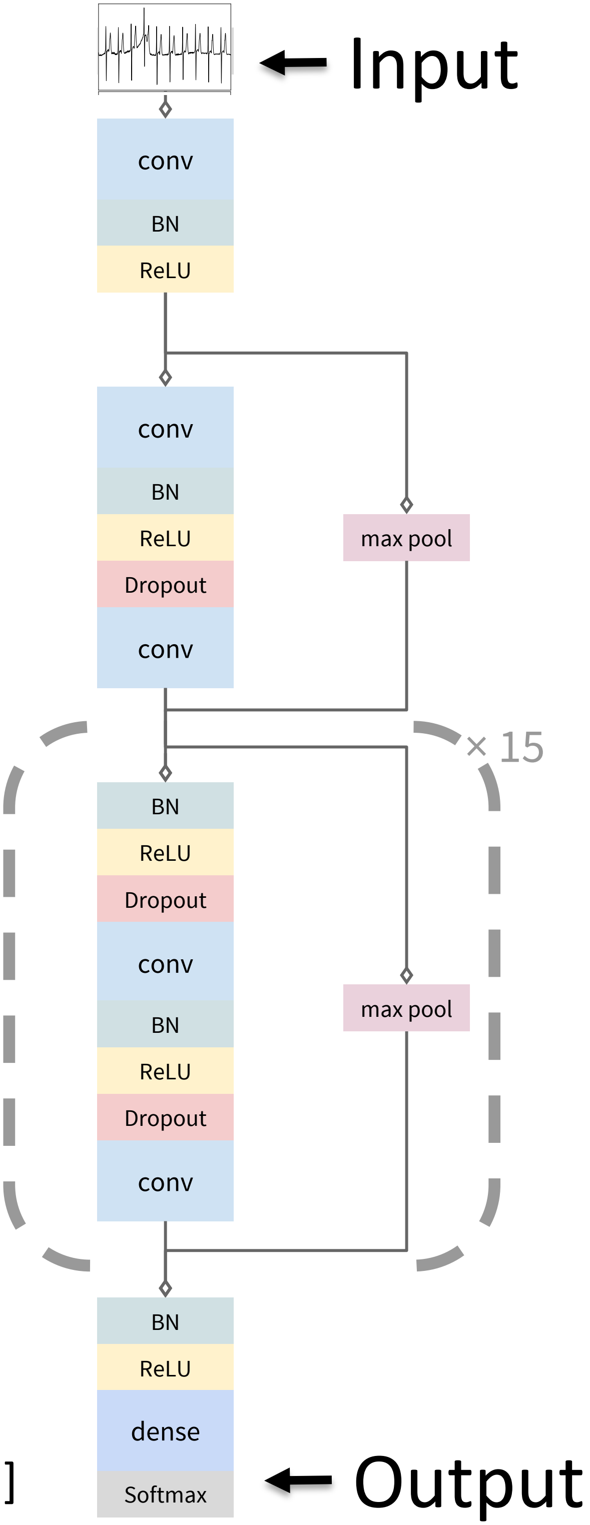
- ~90K ECG records annotated (from ~50K patients)
- Identify 12 heart arrhythmias, sinus rhythm and noise for a total of 14 output classes

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

Deep convolutional network

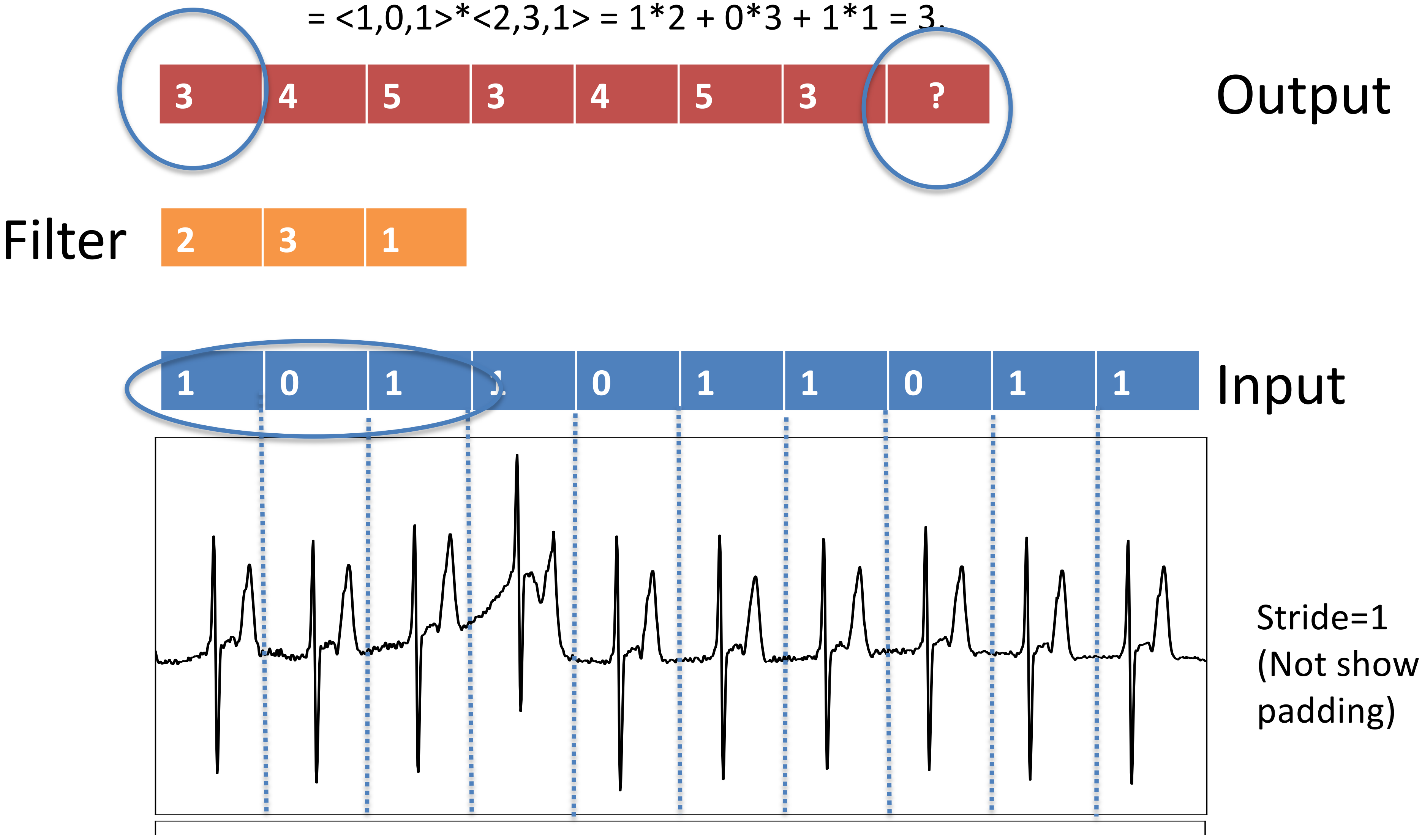
- 1-D signal sampled at 200Hz, labeled at 1 sec intervals
- 34 layers
- Shortcut connections (ala residual networks) with max-pooling
- Subsampled every other layer (2^8 in total)



[Rajpurkar et al., arXiv:1707.01836, 2017; Nature Medicine ‘19]

Example of 1D convolution

= <1,0,1>*<2,3,1> = 1*2 + 0*3 + 1*1 = 3.



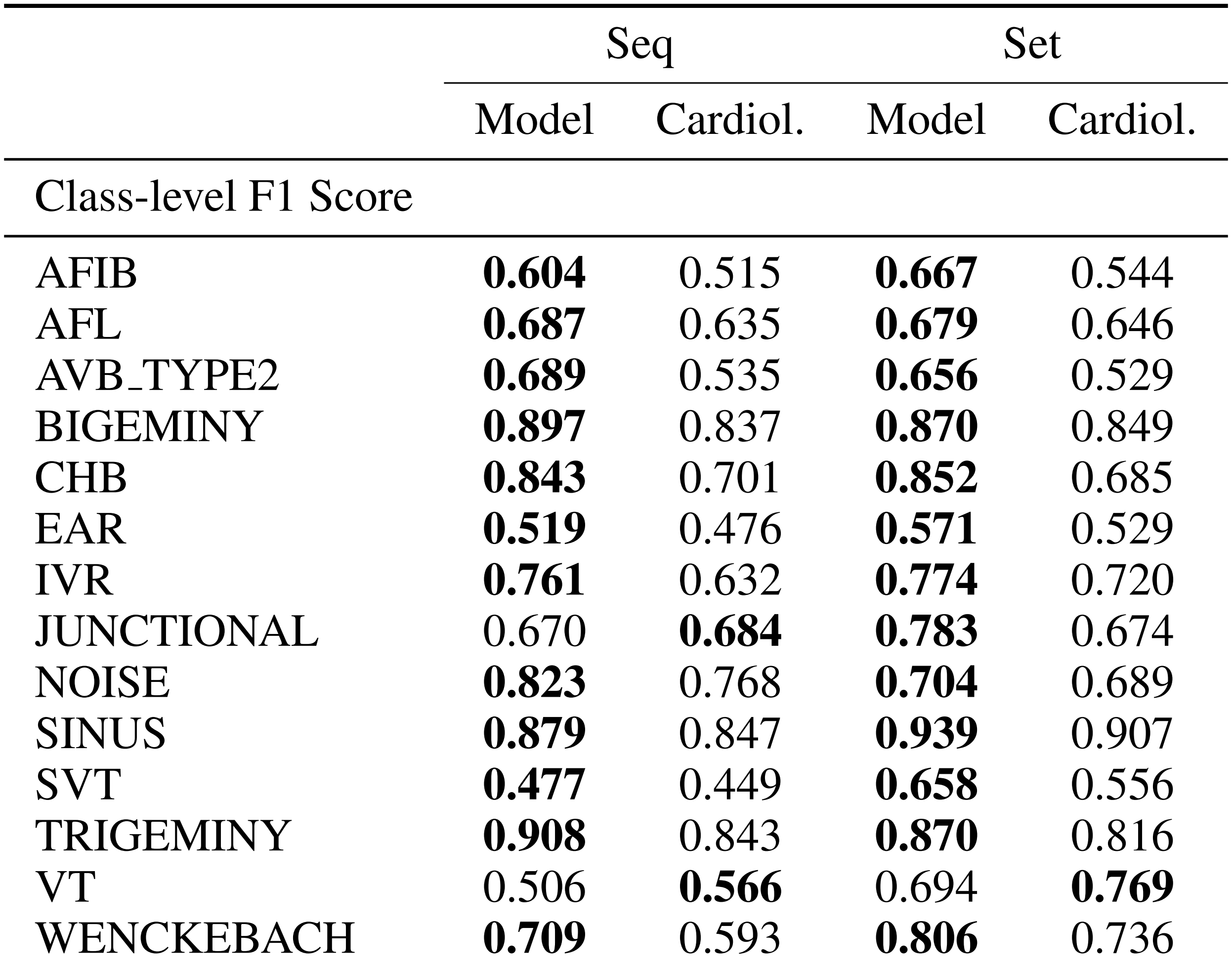
Output: 3 | 4 | 5 | 3 | 4 | 5 | 3 | ?

Filter: 2 | 3 | 1

Input: 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1

Stride=1
(Not showing
padding)

Evaluation: beat-level (‘Seq’) vs. patient-level (‘Set’)

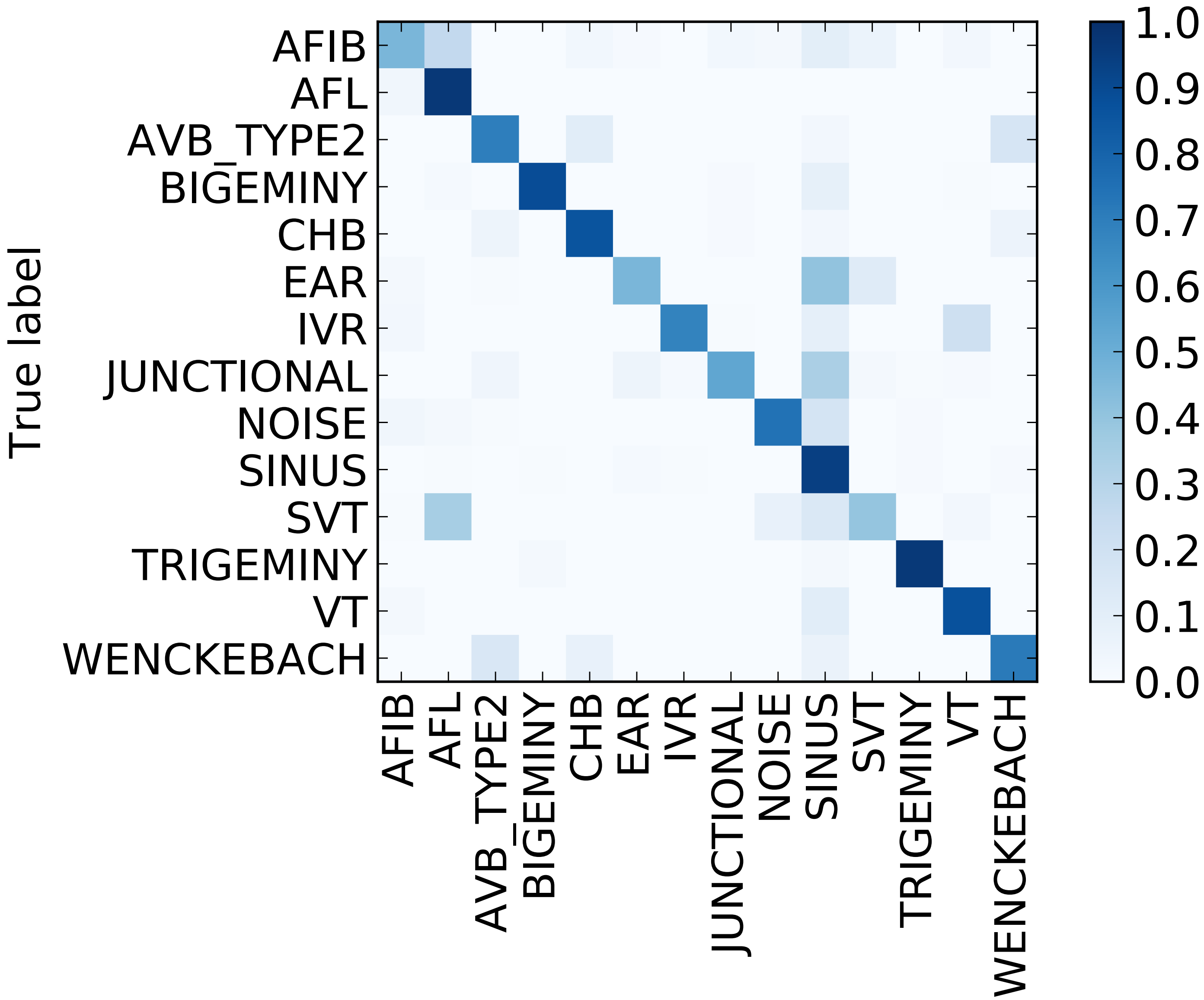


	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	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

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

Recall = sensitivity
Precision = PPV

Evaluation: confusion matrix



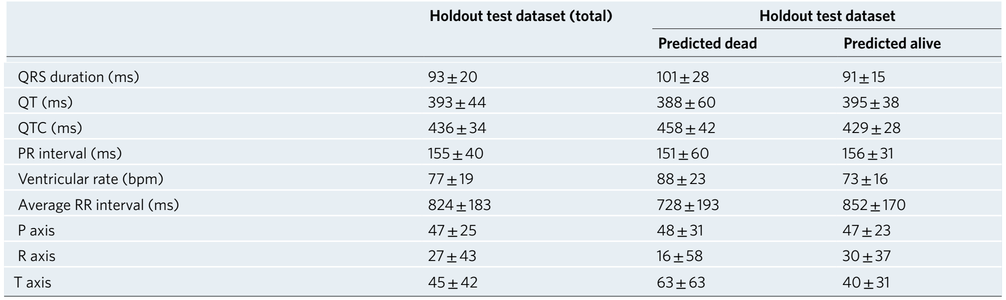
Outline for today's class

1. Using ML for risk stratification
   - Alternative framing: survival modeling
   - Evaluation: metrics, interpretability
2. Physiological time-series: application to detecting irregular heart arrythmias
   - Small data approach
   - Big data approach
   - **Current research**

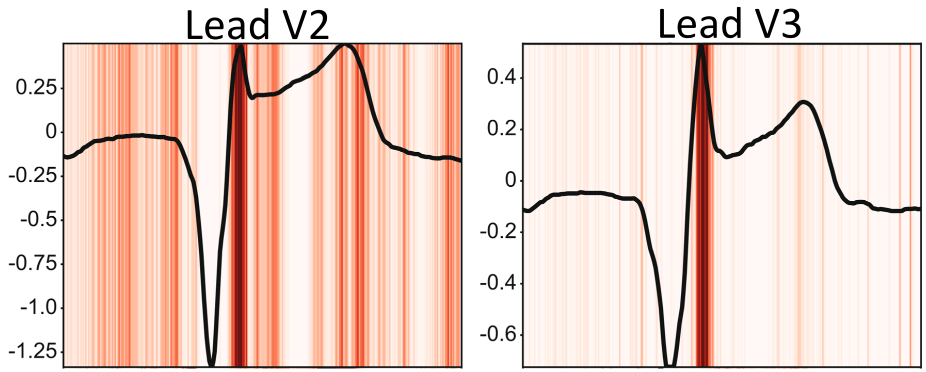
Predicting 1-year mortality using 12-lead ECGs

- >2 million ECGs from 500k patients seen at Geisinger (in Pennsylvania) over 30 years
- Comparison of predictive performance (AUC):
  - .876 – Deep model ECG + age, gender
  - .86 – XGBoost using ECG measures + age, gender
  - .816 - Charlson comorbidity index

"Table 1"



	Holdout test dataset (total)	Holdout test dataset	
		Predicted dead	Predicted alive
QRS duration (ms)	93 ± 20	101 ± 28	91 ± 15
QT (ms)	393 ± 44	388 ± 60	395 ± 38
QTC (ms)	436 ± 34	458 ± 42	429 ± 28
PR interval (ms)	155 ± 40	151 ± 60	156 ± 31
Ventricular rate (bpm)	77 ± 19	88 ± 23	73 ± 16
Average RR interval (ms)	824 ± 183	728 ± 193	852 ± 170
P axis	47 ± 25	48 ± 31	47 ± 23
R axis	27 ± 43	16 ± 58	30 ± 37
T axis	45 ± 42	63 ± 63	40 ± 31

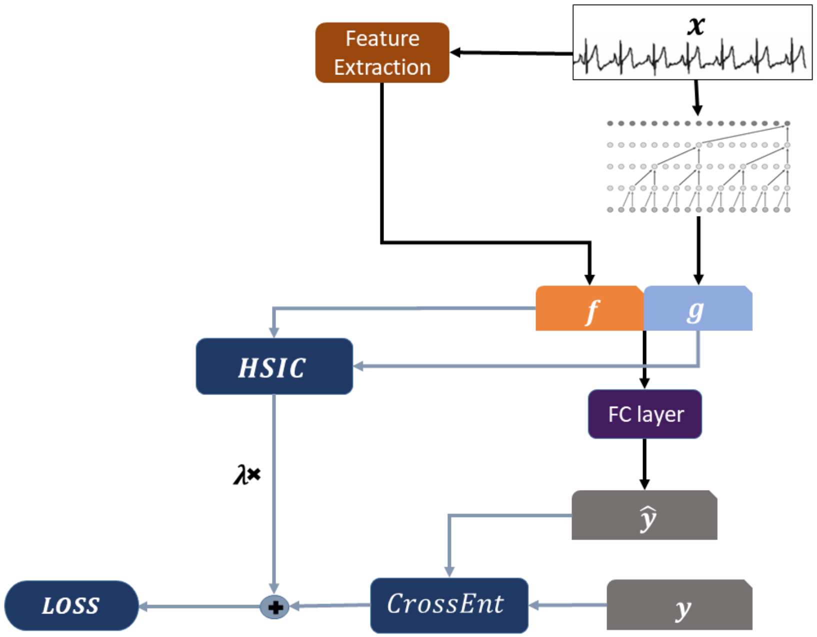


Patient with anterior STEMI who died within 1 year

Gradient-CAM (Selvaraju et al., IJCV '19)

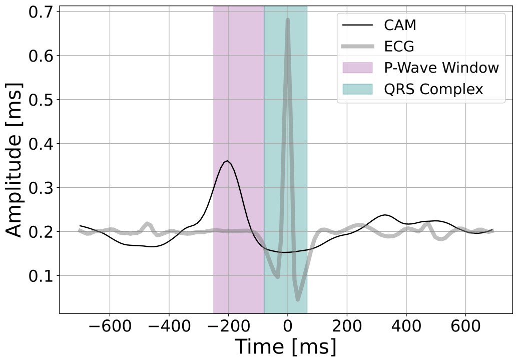
[Figures from: Raghunath et al., Prediction of mortality from 12-lead electro-cardiogram voltage data using a deep neural network, Nature Medicine 2020.
For related work, see also: Ribeiro et al., Automatic diagnosis of the 12-lead ECG using a deep neural network, Nature Communications 2020]

Can we ‘push’ deep networks to discover new features?



Feature Set	Accuracy	F1
RR feature set	93.9%	0.91
P-Wave feature set	87.3%	0.86
All feature set	95.5%	0.95

Model	Accuracy	F1	Avg. R^2 (Independence)	Rep2Label Accuracy
Baseline Model	89.8%	0.90	(0.51, 0.1)	94%
RR Model	94.5%	0.94	0.018	57%
P-Wave Model	89.7%	0.90	-0.082	58%

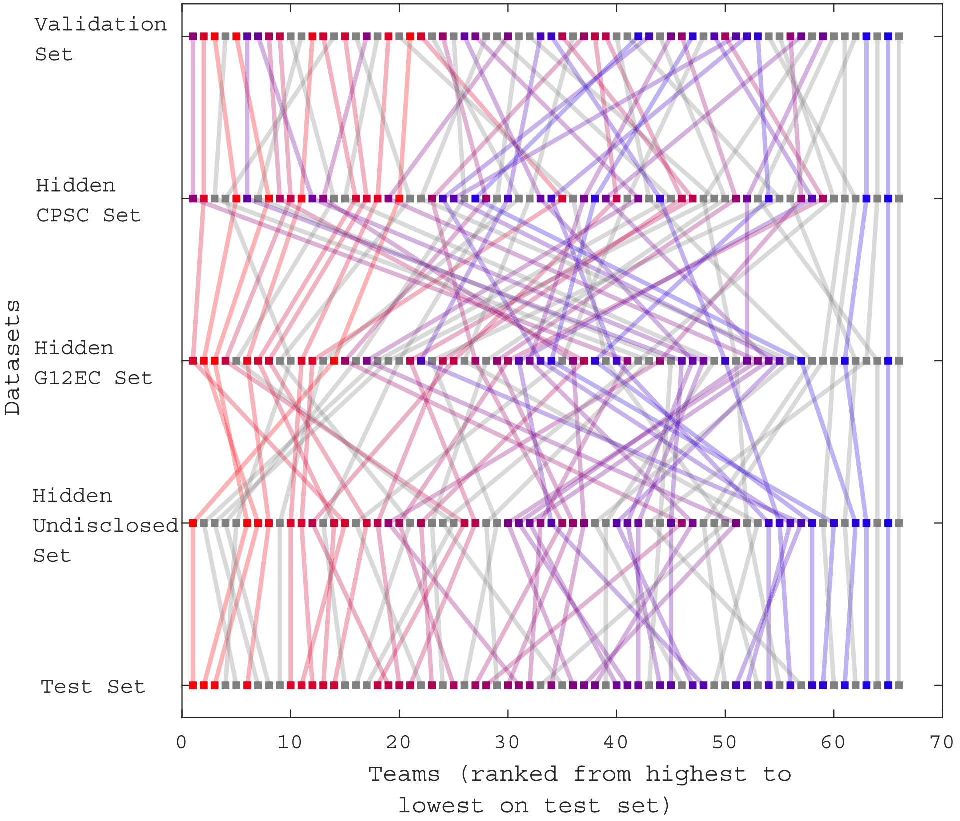


(d) RR constrained

Class activation maps (Zhou et al. 2016) followed by alignment & averaging

[Beer, Eini-Porat, Goodfellow, Eytan, Shalit. Using deep networks for scientific discovery in physiological signals, Machine Learning in Healthcare 2020]

Do models generalize across institutions?



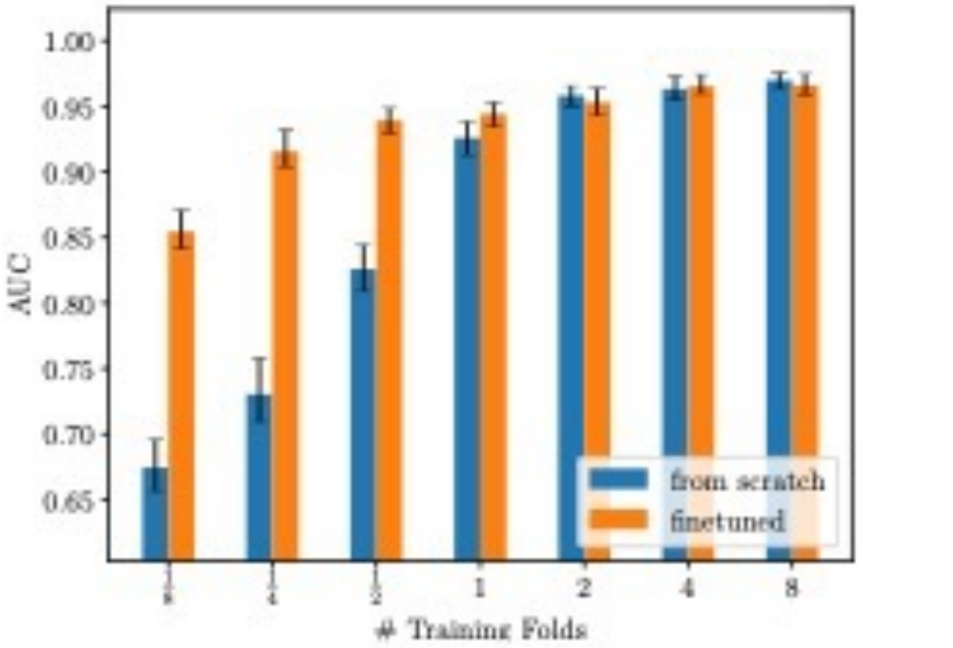


Fig. 4: Effect of transfer learning from PTB-XL to ICBEB2018 upon varying the size of the ICBEB2018 training set.

[Alday et al., Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020, Physiological Measurement, 2020]

[Strodthoff, Wagner, Schaeffter, Samek. Deep learning for ECG Analysis: Benchmarks and Insights from PTB-XL, IEEE Journal of Biomedical and Health Informatics, 2020]

Closing reflections for ML on physiological data

- We are often in realm of “not enough data”
  - Modeling and incorporating prior knowledge can be critical to good performance
- Is machine learning actually picking up new features?
- How can we improve the interpretability and generalizability of the learned models?