Multiple Instance Learning for Cardiovascular Risk Stratification

Anonymous Authors

Abstract

Risk stratification of cardiovascular patients is of clinical importance in ensuring patients at high risk for an adverse outcomes receive the attention they need. ECG signals contain important predictive information, but using them is challenging. In this paper, we introduce the multiple instance learning paradigm as a solution to signal-based risk stratification for cardiovascular outcomes. In contrast to methods that require hand-crafted features or domain knowledge, our method operates directly on the raw ECG signal to learn a representation with high predictive power. The method achieves state-of-the-art performance and demonstrates the value of multiple instance learning for clinical risk stratification. We make two contributions in this paper: 1) reframing risk stratification for cardiovascular death (CVD) as a multiple instance learning problem and 2) introducing a risk score for which patients in the highest quartile are 15.9 times more likely to die of CVD within 90 days of hospital admission at any given time. We include evaluation on both real and simulated datasets to demonstrate the practical potential and hypothetical limits of the proposed method.

Introduction

Motivation Machine learning has led to improved risk stratification models for a number of outcomes, including stroke (Li et al. 2016), cancer (Heidari et al. 2018), and treatment resistance (Perlis 2013). In this paper, we present a new state-of-the-art model for predicting cardiovascular death (CVD) within 90 days of hospital admission using multiple instance learning.

The 90-day time scale is especially important because the majority of post-ACS patients who die of CVD suffer the outcome within three months. Equipped with informative 90-day risk metrics, doctors can make the necessary care decisions to reduce adverse outcomes. Furthermore, predicting CVD within 90 days is a common task in risk stratification literature (Morrow et al. 2007) (Liu et al. 2014), enabling direct comparison to existing methods.

Background The simplest means of predicting CVD is to construct a model based on easy-to-quantify patient characteristics, such as age, sex, or Left Ventricular Ejection Fraction (LVEF) (Cintron et al. 1993). However, there is strong evidence that leveraging electrocardiogram (ECG) signals, which measure electrical activity in a patient's heart, can add significant predictive power (Liu et al. 2014). Because these signals take the form of long time series, however, it is not obvious how best to incorporate them into a predictive model.

One approach is to treat the entire signal as an input to a model, either by extracting summary statistics or feeding it into a recurrent neural network (Myers, Scirica, and Stultz 2017). An alternative is to treat one ECG signal as a sequence of many examples of heartbeats. A particularly successful means of doing so is to extract pairs of consecutive heartbeats and represent each pair using a set of informative features (McCraty and Shaffer 2015), (Syed et al. 2009), (Liu et al. 2014). Such a pair is a special case of what we will refer to as a Consecutive Beat Series (CBS). This approach presents two challenges:

1. **Representation**. It is not obvious what features to extract from each beat pair. This is made especially difficult by the variable duration of heart beats.
2. **Nondeterminism**. Different levels of CVD risk at the patient level do not necessarily yield different characteristics at the level of individual beat pairs. I.e., a high risk patient may have *more* worrisome heartbeats, but will not have *exclusively* worrisome heartbeats.

We address these challenges by fusing ideas from two areas of machine learning: deep learning and Multi-Instance Learning (MIL). We address the representation challenge by learning predictive features directly from the data, with no domain-specific engineering, using a compact neural network. This is in sharp contrast to existing approaches, which extract handcrafted features.

To address the challenge of nondeterminism, we cast the task as a MIL problem. In MIL, labels are available for collections of examples, but not individual examples. In our case, the examples are consecutive beat series, collections are the sets of CBSs extracted from a given ECG signal, and the labels are the outcomes for the associated patients.

Contributions In this paper, we introduce:

- A new ECG risk score that outperforms existing risk metrics in terms of both AUC and hazard ratio, for both 60-day and 90-day risk of cardiovascular death.

- The framing of signal-based risk stratification as a Multi-Instance Learning problem.
- Evidence that, contrary to common wisdom, contiguous triplets of heartbeats are more informative than contiguous pairs for cardiovascular risk stratification.

We include a brief summary of related work in cardiac risk stratification and multiple instance learning in Section 2. We then outline our approach in Section 3, followed by our results on real data in Section 4. Section 5 demonstrates the models performance in simulated settings. The paper concludes with a discussion of future work in Section 6.

2 Related Work

The proposed work relates to two bodies of literature: cardiac risk stratification and multiple instance learning.

Cardiac Risk Stratification

Risk stratification for cardiovascular death has received considerable attention because of the life-saving benefits of early intervention (Keller, Carballo, and Carballo 2011). Approaches to stratification can be divided into two types: heuristic and learned. The former are based on handcrafted scoring functions or domain-specific measurements, while the latter are based on machine learning.

Heuristic Risk Metrics Prominent examples of heuristic risk metrics include the TIMI Risk Score (TRS), Deceleration Capacity (DC), and Morphological Variability (MV). TRS is a point-based system for indicating risk based on clinical variables such as age or ST-elevation in an ECG signal (Antman et al. 2000). DC operates on the theory that slower heart rate deceleration indicates a higher risk of death (Bauer et al. 2006). MV averages the dynamic time warping alignment cost between adjacent beats over the first 24 hours of a patient's ECG signal (Syed et al. 2009). Other methods make use of a patient's ECG signal in combination with other indicators, including the level of troponin (a cardiac-specific protein) (Jernberg, Lindahl, and Wallentin 2000) and left ventricular ejection fraction (LVEF) (Cintron et al. 1993).

Learned Risk Metrics Two recent works have developed machine learning-based risk metrics as an alternative to the above heuristic approaches. Morphological variability in beat space (MVB) is an extension of MV. It learns the optimal frequency at which to measure variability in a patient's ECG signal (Liu et al. 2014). Myers et. al. combine a logistic regression over patient features with an RNN over a small sample of ECG data (Myers, Scirica, and Stultz 2017).

Our method is distinct from these approaches in that it frames the task differently, uses the raw signal instead of expert designed features, and does not require additional patient features. As we show in Section 4, it also provides better predictions.

Multiple Instance Learning

Multiple instance learning tasks involve three entities: collections, instances, and labels. In traditional supervised learning problems, instances are associated with labels. Here, instances belong to collections and each collection is associated with a label (Amores 2013).

A wealth of MIL research has occurred since the field's inception in 1998. Recently, significant effort has gone towards formalizing the problem space of MIL. Below, we position our work in relation to four MIL problem parameters discussed in a recent survey (Carbonneau et al. 2018).

Collection Composition MIL methods follow either the standard MI assumption or the collective assumption. The former states that positive collections have at least one positive instance while negative collections have none. The collective assumption states that positive and negative collections differ in *percentage* of instances that are positive (Frank and Xu 2003). These positive instances are termed witnesses. Our work makes the collective assumption.

Task There are two major tasks in multiple instance learning: instance label prediction and collection label prediction. Our goal is to predict labels for collections, which correspond to patients in our task.

Instance Origin Instances can represent different versions of a given object or distinct objects that are grouped together. We use the latter framing, since it better fits the case of many heartbeats from one patient.

Label Distribution Instances may originate from the same label space as collections, as in drug activity prediction (Zhao et al. 2013), or from distinct label spaces. Since CBSs are not explicitly associated with a patient's outcome, this paper deals with MIL in distinct label spaces.

Vocabulary-based MIL methods aim to classify collections based on an understanding of latent instance classes and have been shown to excel in problems governed by the collective assumption and distinct label spaces. Existing methods learn this vocabulary through clustering (Amores 2013) (Zhou and Zhang 2007) or using a kernel on the instance level (Gärtner et al. 2002) (Rubner, Tomasi, and Guibas 2000) or the bag level (Chen, Bi, and Wang 2006). State-of-the-art vocabulary-based methods include Distance-Based Bag of Words (DBBoW) and Histogram-Based Bag of Words (HBBoW). We explore their performance on the ECG risk stratification task in the following section.

Despite the analogies to patient risk stratification, MIL has, to our knowledge, not been previously applied to biometric signals.

3 Proposed Framework

We formulate the cardiovascular death risk stratification task as follows. We are given a collection of M ECG signals $T_1, \ldots, T_M$. Each signal is associated with a unique patient m, and consists of L_m scalar samples, $t_1, \ldots, t_{L_m}$, $t_i \in R$. Each patient is associated with a label $y_m \in \{0, 1\}$, where $y_m = 1$ indicates that the patient died of cardiac death within 90 days of hospital admission. Our task is to predict the label for held-out patients based on their ECG signals.

Our approach is to treat this as a multiple instance learning problem. This entails converting the raw ECG signals to collections of *instances*, classifying instances as belonging to patients of class 0 or 1, and then aggregating these predictions to produce an overall prediction for each patient. This can be formalized as the construction of three functions:

1. **Instance Extractor**. A map $H : \mathcal{T} \to \mathcal{X}$, from the space of ECG signals to the space of collections of instances.
2. **Instance Classifier**. A map $F : X \to [0, 1]$ from individual instances to probabilistic class predictions.
3. **Instance Aggregator**. A map $G : [0, 1]^N \to [0, 1]$ from predictions for each instance in a collection to an overall prediction for the collection.

We describe each of these functions below.

Instance Extractor

In the first step, H transforms a signal T_m into a set of N_m instances $X_m = \{\mathbf{x}^{mi}\}_{i=1}^{N_m}$. In the context of ECG signals, our instances are pairs (or, more generally, groups) of adjacent heartbeats. The sequence of all such groups taken from a given ECG signal is known as a Consecutive Beat Sequence (CBS). Because the beats must be consecutive, there are $B - G + 1$ groups of G beats in a signal containing B beats. To extract these groups, we identify the peak of each heartbeat using a waveform-based method (Martínez et al. 2004) and take one second of data, centered on each peak. This can result in significant overlap between instances when a patient's heart rate is faster or slower than 60 beats per minute. However, we opt for a simple instance segmentation procedure to demonstrate that our approach does not rely on carefully crafted instance segmentation protocols. We also limit the number of instances taken from each ECG signal to 1000, corresponding to roughly 15 minutes, in order to accelerate training. There is no *a priori* reason one could not use more instances; we simply found that 1000 was already sufficient to outperform existing methods.

Instance Classifier

We employ a compact, fully connected neural network to map each instance to its patient outcome. The model consists of one fully connected layer connected to a sigmoid output. We assessed layer sizes of 1, 2, 5, and 10 units on validation data and discovered that performance plateaus at layer sizes greater than 2. Thus, the results presented here are for a fully connected 1-hidden-layer network of two ReLU-activated units and one sigmoid output. We implement the network in Keras and train it with an L2 regularization parameter of .0001.

This model has several advantages over existing approaches. First, it operates directly on the raw instances, with no feature extraction or costly alignment operations. Second, it allows the use of more than two beats as an instance; this is not the case for methods such as MV and MVB, which specifically compute distances between pairs of beats. As we show in Section 4, this can improve prediction accuracy.

Instance Aggregator

In MIL terms, we aggregate instances based on the collective assumption. That is, instead of assuming that only collections with $y = 1$ have *any* instances predicted as belonging to this class, we only assume that collections with $y = 1$ have *more* of these instances (Foulds and Frank 2010). For ease of exposition, we will refer to these positive-class instances as *witnesses* (Carbonneau et al. 2018), and their relative frequency within a collection as the *witness rate*.

Based on this assumption, we compute the probability of the collection having label $y_m = 1$ as the mean of the predictions for each of its instances. I.e.,

$$G(X_m) = \frac{1}{N_m} \sum_{i=1}^{N_m} F(\mathbf{x}^{mi}) \qquad (1)$$

This formalizes our hypothesis that patients likely to die within 90 days of hospital admission contain certain pathological beat sequences at a higher rate than low risk patients.

We will use the term *instance aggregation AUC* when referring to the AUC of G. Following a convention common in clinical literature, we designate patients that land in the upper quartile of the metric as high risk and the lower three quartiles as low risk.

4 Application

In this section, we describe our experiments on real data.

Data

We use a subset of 760 patients from the MERLIN-TIMI (Thrombolysis in Myocardial Infarction) dataset (Morrow et al. 2007) consisting of 666 non-CVD patients and 94 CVD patients. Each ECG signal was sampled at a rate of 128Hz, and spans approximately forty-eight consecutive hours of time.

Demographic factors for CVD and non-CVD patients are similar, as are their distributions of pre-existing conditions such as smoking. These steps are taken to ensure that the learned model's predictive power is not a result of its learning to merely predict a proxy for CVD risk, such as age.

Three pre-processing steps are typical when working with ECG data. The first is to remove abnormal beats from the ECG signal. Next is to remove baseline wander, or noise in the signal caused by patient motion or respiration (Gupta, Sharma, and Joshi 2015). Finally, one normalizes the entire signal based on each patient's R-wave amplitude in order to correct for inter-patient differences in measurement. This protocol is directly borrowed from (Liu et al. 2014), and we perform the relevant steps using the same Physionet SQI package (Li, Mark, and Clifford 2007).

Note that the incidence rate in our subset of data is 12.4% (94 CVD patients out of 760 total), far greater than the true incidence rate of 2.5%. This is an artifact of the subset. In order to produce a test set that mirrors this true incidence rate and matches the test cohort size of existing methods, we reserve 600 patients for the test set (586 non-CVD, 14 CVD). This leaves us with a training set much smaller than our test set. If we were to maintain the true incidence rate,

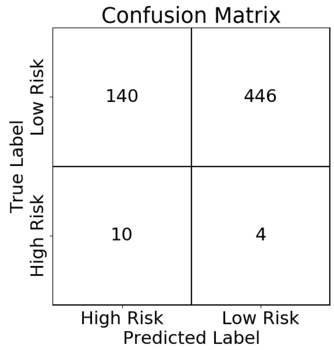

Figure 1: Confusion matrix of risk model. True high risk patients are those that die of cardiovascular death within 90 days of hospital admission, while low risk patients are those that do not.

this would include only 3 CVD patients. This makes learning the positive distribution infeasible. As a result, we opt for a balanced training set of 160 patients (80 non-CVD, 80 CVD). One could perhaps achieve better results with some smaller number of positive examples that better mirrors the true class balance, but we adopt this simple approach to ensure that our results err on the side of pessimism.

Availability Trial protocols for the clinical data were approved by local or central institutional review boards. Because of the sensitivity of the dataset, it is not publicly available. In the interest of reproducible research, however, the models and simulated data discussed are available.

Comparison Methods

We evaluate our approach against the following existing CVD risk metrics:

- **MV** Morphological variability (MV) measures risk for CVD by averaging the dynamic time warping distance between adjacent beats. MV operates on the first 24 hours of a patient's ECG signal.
- **MVB** Morphological variability in beat-space (MVB) improves upon MV by learning the best frequency at which to compute variability between segments in an ECG signal. Similar to MV, MVB operates on the first 24 hours of a patient's ECG signal.
- **TRS** The TIMI Risk Score (TRS) quantifies risk based on the presence or absence of seven risk factors. Patients with four or more risk factors are considered at high risk risk for CVD.
- **LR-RNN** The LR-RNN approach combines the output of an RNN trained the first minute of patient's ECG signal and the output of a logistic regression model over seven patient features, including HRV. The LR-RNN operates on the first usable minute of a patient's ECG signal.

Results for these methods are taken from their respective papers, with the exception of TRS results being taken from (Liu et al. 2014). Additional detail about these risk metrics may be found in the related work. In addition to these published methods, we also include the following MIL methods to evaluate our design choices:

Method	AUC	90-Day HR	60-Day HR	30-Day HR
Age	.68	1.06	1.06	3.60
MIL-LR	.72	5.52	5.18	3.62
STK	.56	2.40	2.66	1.21
MIL-NN1	.72	4.16	4.47	3.60
MIL-NN2	.73	9.66	11.13	6.35
MIL-NN3	**.79**	**15.90**	**16.79**	4.37
MIL-NN4	.75	10.74	12.54	6.03
MV	.72	8.45	-	**12.30**
MVB	.72	8.81	-	-
LR-RNN	.77	5.92	6.64	5.26
TRS	.67	-	3.82	3.31

Table 1: Model performance predicting cardiovascular risk. We report the AUC, 90-day, 60-day, and 30-day hazard ratios. We report the published performance of MV, MVB, LR-RNN and TRS.

- **STK** STK applies an SVM paired with a statistics kernel, which defines collection similarity as the dot product of collection-level statistics including mean, variance, and standard deviation (Gärtner et al. 2002).
- **MIL-LR** MIL-LR combines the MIL framework with logistic regression as the instance classifier. We include this as a baseline to evaluate the utility of a neural network.
- **MIL-NN[K]** MIL-NN[K] combines the proposed instance aggregator with an instance classifier trained on K consecutive heartbeats. We test over $K \in \{1, 2, 3, 4\}$.

Results

The MIL framework combined with a neural network instance classifier produces a competitive risk metric in terms of the hazard ratio and AUC. The confusion matrix relating the accuracy of the proposed risk score is shown in 1. We include the reported AUCs and HRs on the same dataset for each risk metric.

Cohort Results We can construct a high risk cohort and a low risk cohort of patients from the upper quartile and lower three quartiles of the resulting metric. Using this split, 150 of the 600 test patients are classified as high risk and the remaining 450 as low risk. As mentioned previously, there are 14 CVD patients in the test set. 10 of these patients are classified as high risk by our method, a number that far exceeds competing metrics. Thus, the CVD incidence rate for high risk patients is 6.67%, while the CVD incidence rate for low risk patients is .89%.

Hazard Ratio Hazard ratios are commonly reported in the literature for cardiac risk stratification; we report the HR for each model in Table 1. We rank patients based on our risk score and calculate the HR according to the Cox Proportional Hazards model (Lin and Wei 1989). This hazard ratio measures the time-average relative risk between the high risk and low risk cohort. Looking at Figure 3, we see the

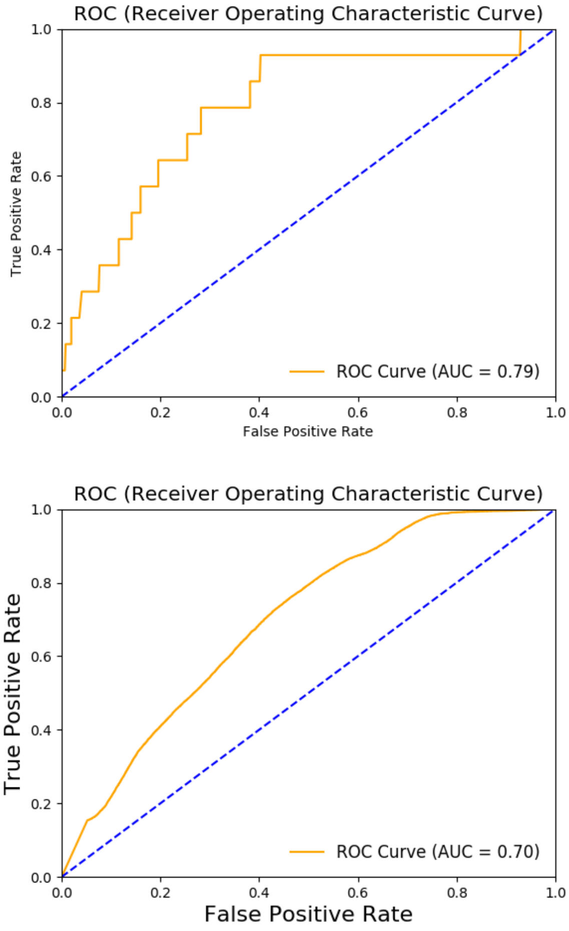

Figure 2: AUC of collection-level prediction (top) and AUC of instance-level prediction (bottom)

Kaplan-Meier mortality curve of the risk metric learned by our algorithm. Averaged across all days, a patient with a risk score in the top quartile is approximately 16 times more likely to die of cardiovascular death. A detailed comparison of the hazard ratios across different time scales can be found in Table 1. We test across three common forecasting intervals: death within 30, 60, and 90 days from hospital admission. Note that the hazard ratio differs from the odds ratio, also commonly reported in risk stratification research, in that the HR measures the relative risk averaged over time and the odds ratio measures the relative risk at the endpoint (past 90 days).

Although LR-RNN operates on the first minute of a patient's signal, as opposed to the first fifteen minutes for MIL-NN3, it is also informed by seven handcrafted patient features. Despite this added information, MIL-NN3 demonstrates higher predictive power.

It is worth noting that our metric excels in predicting adverse events at a longer time scale than competing methods, but performs slightly worse than both LR-RNN and MV in terms of 30-Day HR.

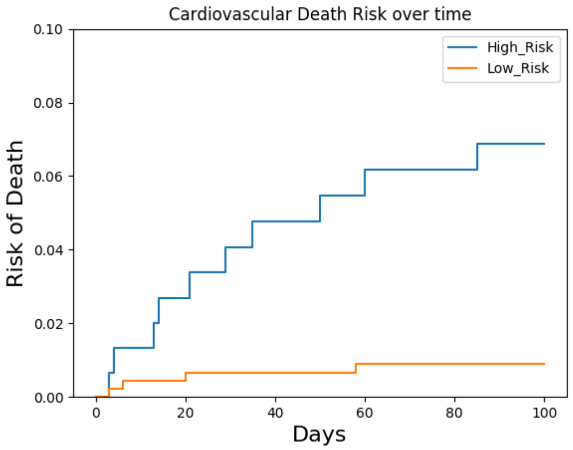

Figure 3: Mortality curve of patient population stratified by the learned risk metric. Patients classified as high risk are 16 times more likely to suffer from CVD than patients classified as low risk.

AUC As shown in Table 1, the AUC of our MIL-NN3 is higher than that of any competing method. To better understand its performance, we plot the ROC curves for both its collection-level and instance-level classification in Figure 2. Because the collection classifier averages many instance predictions, it is unsurprising that the AUC at the collection level is higher than the AUC at the instance level. Interestingly, however, there is a significant amount of discriminative power in the instances themselves—the instance classifier achieves an AUC of .70.

This fact, along with the fact that averaging 1000 such predictors gives the collection classifier an AUC of .79, indicates that errors within a collection are highly correlated. As we show in Section 5, bagging this many independent estimators would yield nearly perfect classification.

Performance of MIL-NN3 Focusing on the comparison between the MIL-NN* models, we see that MIL-NN3 outperforms its counterparts that operate on differing numbers of consecutive beats. Existing CVD metrics, including MVB and MV, focus on adjacent pairs of beats. Our results suggest that contiguous triplets of heartbeats are more predictive of cardiac risk than consecutive pairs of heartbeats.

Interestingly, groups of four heartbeats are worse than three. We hypothesize that this is because the larger input space of four contiguous beats makes it harder for the model to identify important patterns and avoid overfitting.

Comparison to Other MIL Methods We also compare our method to other MIL approaches. Unfortunately, because the dataset used here far exceeds the typical scale of MIL datasets (Cheplygina and Tax 2015), most MIL methods are inapplicable. In particular, many of the most well known MIL algorithms are kernel-based methods that rely on the pairwise distances between all training instances. This requires a prohibitive amount of memory as the number of instances increases. Even limiting the training data to 1000 instances for each ECG signal results in a pairwise distance matrix requiring over 200 gigabytes of storage.

As a result, we evaluate three memory-efficient MIL approaches. MIL-NN and MIL-LR are identical save for the instance classifier; MIL-NN uses a neural network to map instances to collection labels, while MIL-LR uses logistic regression. STK applies an SVM to vectors of collection-level statistics and defines bag similarity as the distance between these statistics vectors. Because STK operates exclusively on summary statistics, it is not surprising that MIL-NN and MIL-LR outperform it.

Label Imbalance We also see that the proposed MIL model is resilient to strong differences between the training label distribution and the test label distribution. Despite being trained on a population with a 50% CVD incidence rate, the resulting risk metric successfully stratifies a population with a 2.5% CVD incidence rate.

5 Simulation

The ECG risk stratification task is one instance of the broader problem of binary classification of collections with different witness rates. We include simulations of the proposed architecture's behavior to show: 1) failure modes, 2) broader relationships between witness rate and collection size, and 3) method performance in conditions outside of the presented example.

We focus on three parameters: the witness rate in the positive collections (p^+), the witness rate in the negative collections (p^-), and the number of samples per collection (N_m).

Experimental Setup

Instances from the witness class are drawn from a 2D isotropic Gaussian centered on (1, 1). Non-witness instances are drawn from a similar Gaussian centered on (1, -1).

Similar to the ECG setting, we train the network on 200 collections evenly split over the positive and negative classes. We measure parameter effects in two ways: instance AUC (corresponding to the instance classifier F) and collection AUC (corresponding to the instance aggregator G). AUC values are averaged over 10 trials.

Results

Relative Witness Rates We measure the model's ability to distinguish positive collections from negative collections by its instance AUC and collection AUC. As before, the instance AUC is associated with the instance prediction task and the collection AUC corresponds to the instance aggregation task. In Figure 3, we demonstrate the effect of varying witness rates on the resulting instance AUC (above) and collection AUC (below). The x axis represents p^+ and the y axis represents p^- and the corresponding boxs color represents the achieved AUC. Darker boxes indicate lower AUCs than lighter boxes, as demonstrated by the colorbar to the right of each graph.

Let us consider the cell at $p^+ = .85$ and $p^- = .6$. This box represents the scenario where on average, 85% of instances in positive collections are witnesses and 60% of instances in negative collections are witnesses. Applying the same model, with the same instance aggregator and instance predictor, remains an effective way to differentiate collections despite the fact that 27.5% of the training labels are incorrect. This generalizes to other risk prediction settings in which a patient may be treated as a collection of concepts.

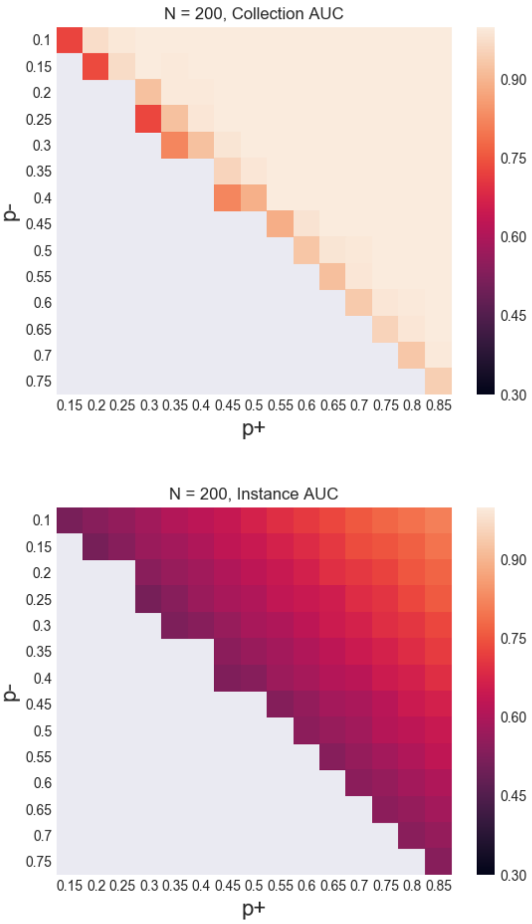

Figure 4: Relationship of witness rates with 1) AUC of instance aggregation (top) and 2) AUC of instance prediction (bottom)

Also of note is that the relationship between AUCs is nearly upper triangular, indicating that the difference between witness rates matters more than their magnitudes to the achievable AUC.

Collection Size In Figure 5, we show that a minor increase in collection size results in a drastic improvement in collection-level AUC. The figure demonstrates the relationship between AUC and collection size for three settings in which the positive witness rate is close to the negative witness rate, where δ represents the difference in witness rates. Common criticism of the multiple instance learning framework is its difficulty in learning a separator when δ is small. In the top figure, only 150 instances are necessary per collection to learn a classifier with an AUC of .9. As δ increases, the number of instances required per collection plummets.

With a witness rate difference of .15, each collection need only have 30 samples to achieve an AUC of .9.

Failure Modes This method struggles to discriminate between collections when the positive witness rate is less than 5% - 10% greater than the negative witness rate, at a collection size of 50 instances. In clinical applications, this means that the proposed framework may fail to identify signal differences that do not accumulate over time. Tasks in which the occurrence of an instance, rather than the recurrence, differentiate classes might pose a greater challenge to this framework.

Relation to ECG Experiment Through these simulations, we see that given enough instances, we should observe a near perfect AUC at the collection level. In the ECG setting, however, the instance classification AUC is .70 while the instance aggregation AUC is .79. This suggests two possible explanations: either the errors are dependent on one another or the difference in witness rates between high risk and low risk patients is very small. The experiments above and the large difference between the average witness rate in high risk patients (74.4%) and low risk patients (44.1%) lend credence to the former explanation.

6 Discussion

We describe a new risk stratification method for cardiovascular death that significantly outperforms the current state of the art on 60-day and 90-day risk prediction, despite requiring less information about patients. This is made possible by framing the task as a multiple instance learning problem, which offers a natural framework for learning from long and noisy electrocardiogram data. Moreover, in contrast to traditional approaches based on handcrafted risk scores or patient features, we use a simple and efficient neural network to learn a useful representation directly from the data.

The success of our approach suggests at least three fruitful directions for future work. First, our ability to learn from longitudinal ECG recordings suggests that models for long-term risk stratification—on the order of years instead of months—may be effective. This has remained a relatively understudied problem (Bueno and Asenjo 2016), but may be tractable using extensions of the ideas presented herein. Second, because there is nothing in our approach that is specific to ECG signals, it could be applied to risk stratification using other biometric signals as well. Third, the objective of this task is binary classification but multiple instance regression could be of value to forecasting patient outcomes at a more granular level.

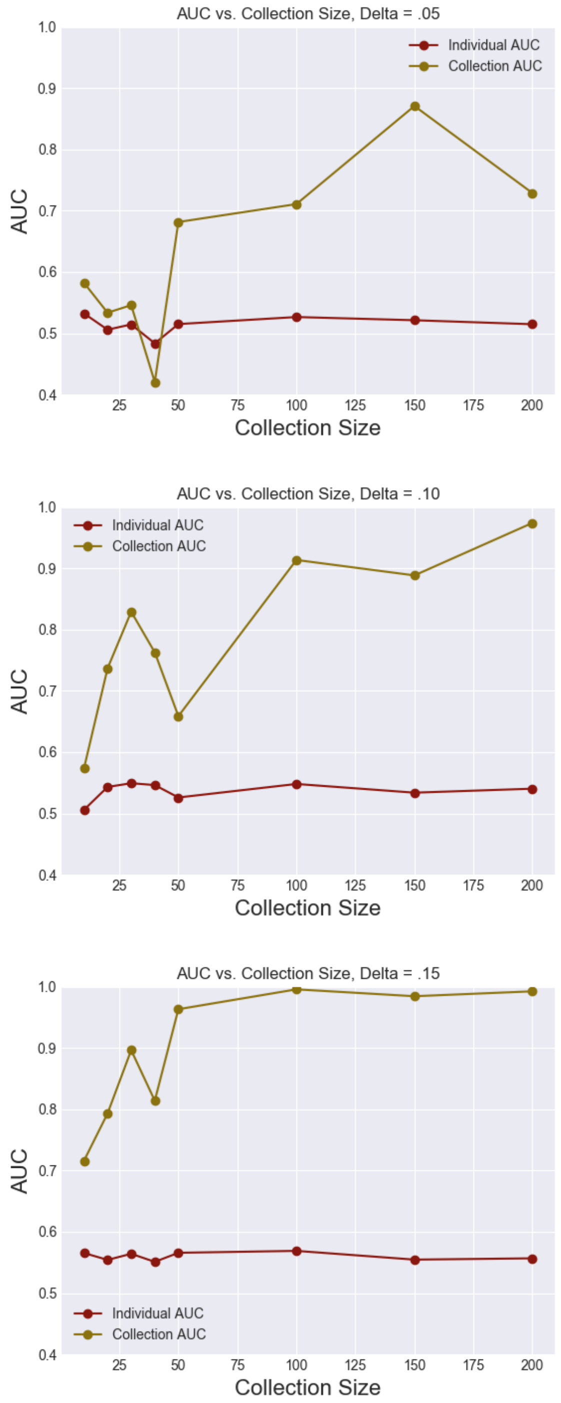

Figure 5: Relationships between instance AUC, collection AUC, and collection size. Each plot represents a certain difference in witness rate. As the collection size increases, there is a slight increase in instance AUC but a drastic increase in collection AUC.

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201:81–105.

Antman, E. M.; Cohen, M.; Bernink, P. J.; McCabe, C. H.; Horacek, T.; Papuchis, G.; Mautner, B.; Corbalan, R.; Radley, D.; and Braunwald, E. 2000. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama* 284(7):835–842.

Bauer, A.; Kantelhardt, J. W.; Barthel, P.; Schneider, R.; Mäkikallio, T.; Ulm, K.; Hnatkova, K.; Schömig, A.; Huikuri, H.; Bunde, A.; et al. 2006. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. *The lancet* 367(9523):1674–1681.

Bueno, H., and Asenjo, R. M. 2016. Long-term cardiovascular risk after acute coronary syndrome, an ongoing challenge. *Revista Española de Cardiología* 69(01):1–2.

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77:329–353.

Chen, Y.; Bi, J.; and Wang, J. Z. 2006. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12):1931–1947.

Cheplygina, V., and Tax, D. M. 2015. Characterizing multiple instance datasets. In *International Workshop on Similarity-Based Pattern Recognition*, 15–27. Springer.

Cintron, G.; Johnson, G.; Francis, G.; Cobb, F.; and Cohn, J. N. 1993. Prognostic significance of serial changes in left ventricular ejection fraction in patients with congestive heart failure. the v-heft va cooperative studies group. *Circulation* 87(6 Suppl):VI17–23.

Foulds, J., and Frank, E. 2010. A review of multi-instance learning assumptions. *The Knowledge Engineering Review* 25(1):1–25.

Frank, E., and Xu, X. 2003. Applying propositional learning algorithms to multi-instance data.

Gärtner, T.; Flach, P. A.; Kowalczyk, A.; and Smola, A. J. 2002. Multi-instance kernels. In *ICML*, volume 2, 179–186.

Gupta, P.; Sharma, K. K.; and Joshi, S. D. 2015. Baseline wander removal of electrocardiogram signals using multivariate empirical mode decomposition. *Healthcare technology letters* 2(6):164.

Heidari, M.; Khuzani, A. Z.; Hollingsworth, A. B.; Danala, G.; Mirniaharikandehei, S.; Qiu, Y.; Liu, H.; and Zheng, B. 2018. Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. *Physics in Medicine & Biology* 63(3):035020.

Jernberg, T.; Lindahl, B.; and Wallentin, L. 2000. The combination of a continuous 12-lead ecg and troponin t. a valuable tool for risk stratification during the first 6 hours in patients with chest pain and a non-diagnostic ecg. *European heart journal* 21(17):1464–1472.

Keller, P.-F.; Carballo, S.; and Carballo, D. 2011. Present and future of secondary prevention after an acute coronary syndrome. *EPMA Journal* 2(4):371–379.

Li, X.; Liu, H.; Du, X.; Zhang, P.; Hu, G.; Xie, G.; Guo, S.; Xu, M.; and Xie, X. 2016. Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. In *AMIA Annual Symposium Proceedings*, volume 2016, 799. American Medical Informatics Association.

Li, Q.; Mark, R. G.; and Clifford, G. D. 2007. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological measurement* 29(1):15.

Lin, D. Y., and Wei, L.-J. 1989. The robust inference for the cox proportional hazards model. *Journal of the American statistical Association* 84(408):1074–1078.

Liu, Y.; Syed, Z.; Scirica, B. M.; Morrow, D. A.; Guttag, J. V.; and Stultz, C. M. 2014. Ecg morphological variability in beat space for risk stratification after acute coronary syndrome. *Journal of the American Heart Association* 3(3):e000981.

Martínez, J. P.; Almeida, R.; Olmos, S.; Rocha, A. P.; and Laguna, P. 2004. A wavelet-based ecg delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering* 51(4):570–581.

McCraty, R., and Shaffer, F. 2015. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. *Global Advances in Health and Medicine* 4(1):46–61.

Morrow, D. A.; Scirica, B. M.; Karwatowska-Prokopczuk, E.; Murphy, S. A.; Budaj, A.; Varshavsky, S.; Wolff, A. A.; Skene, A.; McCabe, C. H.; Braunwald, E.; et al. 2007. Effects of ranolazine on recurrent cardiovascular events in patients with non–st-elevation acute coronary syndromes: the merlin-timi 36 randomized trial. *Jama* 297(16):1775–1783.

Myers, P. D.; Scirica, B. M.; and Stultz, C. M. 2017. Machine learning improves risk stratification after acute coronary syndrome. *Scientific reports* 7(1):12692.

Perlis, R. H. 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological psychiatry* 74(1):7–14.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40(2):99–121.

Syed, Z.; Scirica, B. M.; Mohanavelu, S.; Sung, P.; Michelson, E. L.; Cannon, C. P.; Stone, P. H.; Stultz, C. M.; and Guttag, J. V. 2009. Relation of death within 90 days of non-st-elevation acute coronary syndromes to variability in electrocardiographic morphology. *The American journal of cardiology* 103(3):307–311.

Zhao, Z.; Fu, G.; Liu, S.; Elokely, K. M.; Doerksen, R. J.; Chen, Y.; and Wilkins, D. E. 2013. Drug activity prediction using multiple-instance learning via joint instance and feature selection. In *BMC bioinformatics*, volume 14, S16. BioMed Central.

Zhou, Z.-H., and Zhang, M.-L. 2007. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems* 11(2):155–170.