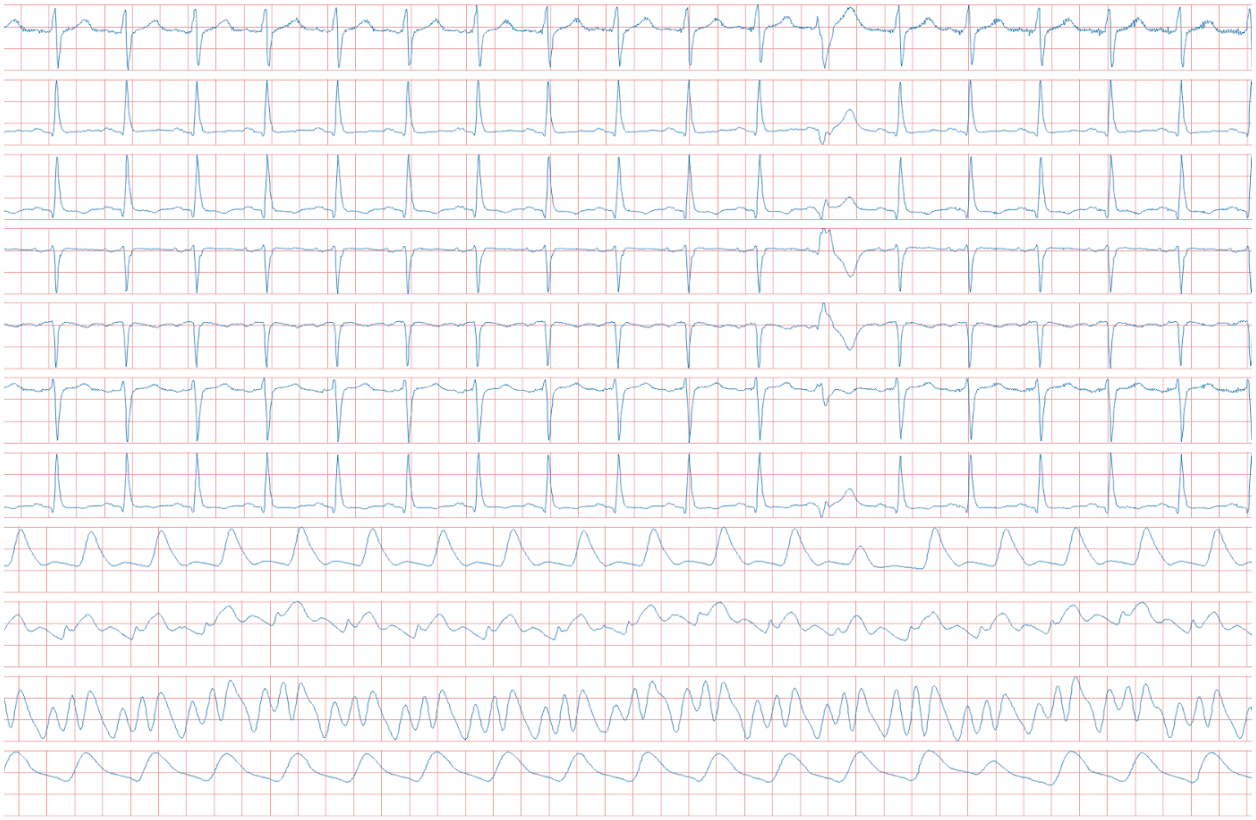

Cardiac Arrhythmia Classification for PVC Heartbeats

Relieving Alarm Fatigue in ICU Environments

UCSF Data Fusion Analytics M.Eng Capstone Team 2018

Final Report Deliverable • 4/30/2018

Alex Ackroyd • Adam Andrews • Siddhant Issar • Segev Malool • Umesh Thillaivasan • Yuntao Wang

Research in collaboration with Jacob Abba and Ran Xiao at UCSF; Advised by Xiao Hu, UCSF, and Gabriel Gomes, UC Berkeley.

Pledge of Academic Integrity

We affirm that we are the sole authors of this report and we give due credit (i.e., use correct citations) to all used sources.

Alex Ackroyd	April 30, 2018
Adam Andrews	April 30, 2018
Siddhant Issar	April 30, 2018
Segev Malool	April 30, 2018
Umesh Thillaivasan	April 30, 2018
Yuntao Wang	April 30, 2018

Table of Contents

Executive Summary

When the heart beats irregularly, it is known as an arrhythmia. A common heart arrhythmia, known as a premature ventricular contraction (PVC), accounts for the highest number of non-actionable and false-positive in-hospital patient monitoring alarms. Current in-hospital patient monitoring systems do not have the capabilities to discern true-PVC alarms from false-alarms, and therefore medical professionals experience alarm fatigue, a desensitization to alarms leading to lower quality of care. This presents an opportunity to apply sophisticated machine learning methods to improve the accuracy of these alarms so that only true-PVC alarms are generated to alert medical professionals. Our team explored several machine learning approaches to handle and classify electrocardiogram (ECG) signal data from two data sets: the famous MIT-BIH Arrhythmia labeled data set, and UC San Francisco's massive unlabeled data set. Our best performing model was a 2-layer neural network which achieved a test set accuracy of 93.64%. We also explored transfer learning and individual heartbeat classification using various approaches such as anomaly detection using K-means clustering. Based on our findings, future work should focus on using neural networks to classify individual heartbeats.

I. Motivation

Alarm fatigue in nursing staff compromises quality of care.

In modern intensive care units (ICUs), patients are connected to multiple electronic devices that monitor vital signs of health. While these bedside monitors make it possible to observe and quantify each moment of a patient's health status, a growing problem for clinical care staff is that the overwhelming majority of alarms are false-positives. This inaccuracy is problematic since responding to each alarm taxes the limited time and attention of available nurses (Hu et al., 2012, 913). The overload of alarms has been linked to avoidable patient deaths in cases where caregivers were so desensitized to monitor alarms that they failed to register and respond to actual life-threatening conditions (Wallis, 2010).

There are many reasons for the high number of false positive alarms. First, there are a multitude of monitors involved in each patient's care. It is standard for patients to be connected to 4 to 12 electrocardiogram (ECG) leads, blood pressure devices, blood oxygen saturation (SpO2) devices, and respiration monitors, producing 7 or more continuous streams of data (Drew et al., 2014). Compounding the large volume of data generated by each patient is the the high proportion of false-positive alarms. Reasons for the high false-positive rate include inappropriate tailoring of alarm thresholds and monitor settings to the individual patient, patient health conditions that are stable and non-actionable from a care perspective but abnormal from a monitoring perspective, algorithm deficiencies within the monitoring technology itself, hypersensitivity to typical patient motion, and poor exception handling, among others (Drew et al., 2014). For example, persistent false-alarms are triggered when a patient has a pacemaker and the ECG isn't configured to the

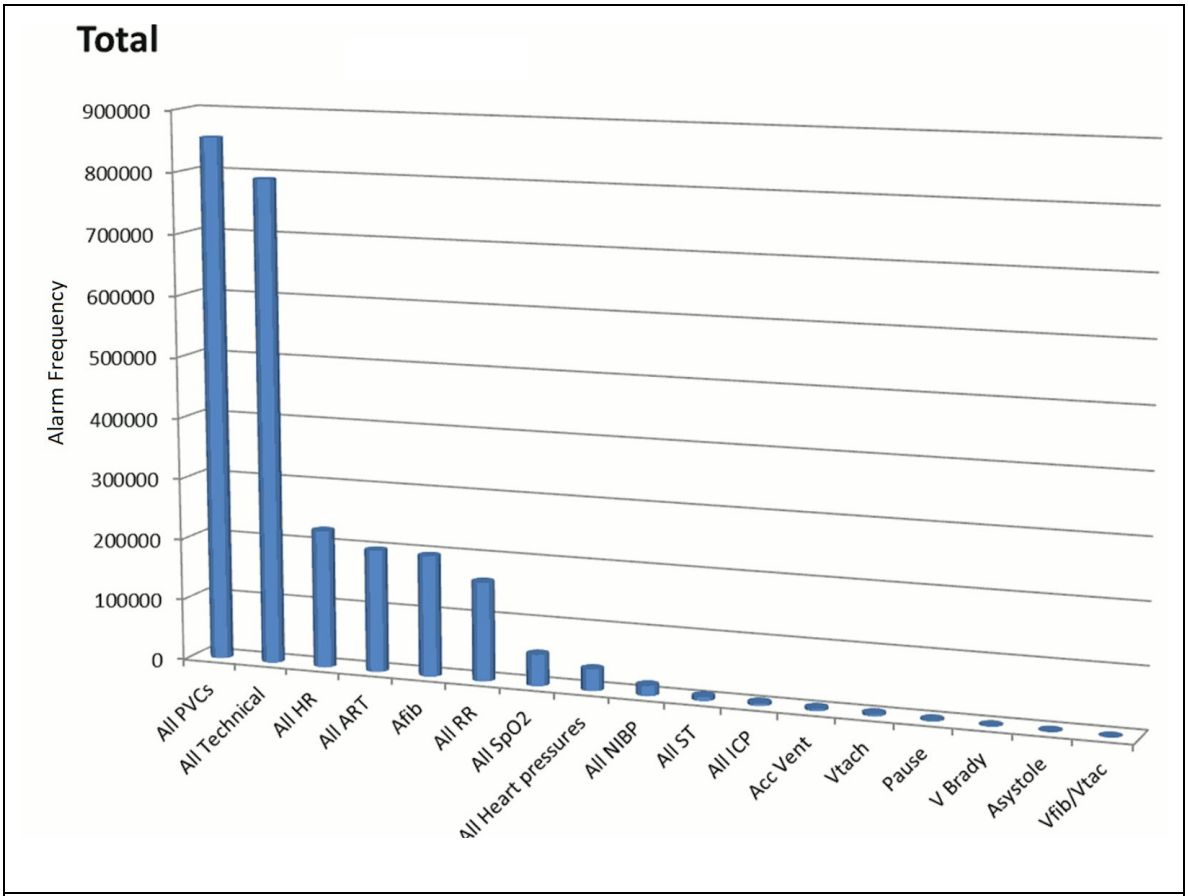

Figure 1: False alarm frequency by arrhythmia type showing Premature Ventricular Contractions (PVC) as most frequent alarm in comparison to several categorizations including Heart rate (HR) and Atrial fibrillation (Afib). Technical alarms refer to any silent alarm used for low-risk rhythms (Drew et al. 2014).

pacemaker mode. While it is easy for a nurse to correctly interpret the artificially-induced heart rhythms, the patient monitor will continuously alarm.

Vital sign monitoring is a critical component of enabling effective and timely care, and true positive alarms are essential in mobilizing a rapid response in life-threatening scenarios. Indispensable as monitors and alarms are to caregivers, the challenge has become a task of improving the alarm algorithms, such that the false-positive rate is reduced without impacting the false-negative rate.

In one study conducted at UCSF, premature ventricular contraction (PVC) alarms accounted for 33% of total alarms (Drew et al., 2014). Despite being so frequent, PVC

alarms are non-actionable due to the landmark 1989 Cardiac Arrhythmia Suppression Trial (CAST) that showed “antiarrhythmic therapy was associated with more deaths than placebo” (CAST Investigators 1989). However, even though PVC alarms are non-actionable and extremely frequent, under certain circumstances they can be an early warning sign of life-threatening arrhythmias such as torsade de pointes (Drew et al., 2014).

Since completely ignoring or disabling PVC alarms isn’t an option, if the false-positive rate for PVC alarms specifically can be reduced, then it would have an outsize impact in reducing the overall false-positive rate and alarm burden for ICU nurses.

II. Literature Review

Many machine learning and data mining methods have been applied to false alarms suppression.

Over the past five years, researchers have made steady progress, first in amassing datasets that can quantify the scope of false-positive alarms in clinical care settings, and then in using those datasets in developing data science algorithms to classify arrhythmias and correct for false-positives.

In 2012, Hu et al. published the first study to directly mine patient monitor alarm data (Hu et al., 2012, 914). In this study they mined alarm data to find frequent combinations of alarms that preceded code blue events (events where a patient requires immediate resuscitation), and evaluated their results using 4-way analysis of variance (ANOVA) on a test set of “223 adult code blue and 1768 control patients” (Hu et al., 2012). They were able to achieve a true-positive rate “between 66.7% and 90.9%” with their SuperAlarm, while reducing the number of false-positive alarms to “between 2.2% and 11.2% of regular monitor alarms” (Hu et al., 2012).

In 2014, Hu collaborated with another team of researchers to publish a model that “suppresses false positive ventricular tachycardia (VT) alarms without resulting in false-negative alarms”, using only the ECG waveforms from the MIMIC II dataset (Salas-Boni et al., 2014, 775). This L-1 regularized logistic regression classifier suppressed false alarms by 21% when evaluated against the MIMIC II dataset, and by 36% when evaluated against the UCSF & General Electric dataset; in both cases no true positive alarms were suppressed (Salas-Boni et al., 2014).

In the same year, two more papers were published. One including laboratory test results as a feature in the SuperAlarm model (Bai et al., 2014), and another publishing an observational study that is truly remarkable (Drew et al., 2014). In the Drew study, a team of nurses went through extensive training in order to review and label a dataset of 2,558,760 alarms as true or false positives. A dataset of this scale and quality was a milestone, enabling data science and machine learning models to train with much higher quality and validity.

Several additional papers have been published by the SuperAlarm group performing additional time-series and classification analysis (Bai et al. 2016) and linear discriminant analysis (LDA) (Shahriari et al., 2016).

Automatic arrhythmia classification has achieved cardiologist-level accuracy using convolutional neural networks.

In recent years, neural networks (CNN) have largely displaced the need to design convolution kernels manually. A neural network is a directed and acyclic computational graph, which may include convolution, matrix multiplication, function maps, and other transformations. It also includes a way to “reverse” each operator by differentiation with respect to internal parameters of the neural net, enabling a local search process to find a local optimum over these parameters. The depth, or composition of layers, reflects the idea that a basic set of abstract features may be combined to represent the signal.

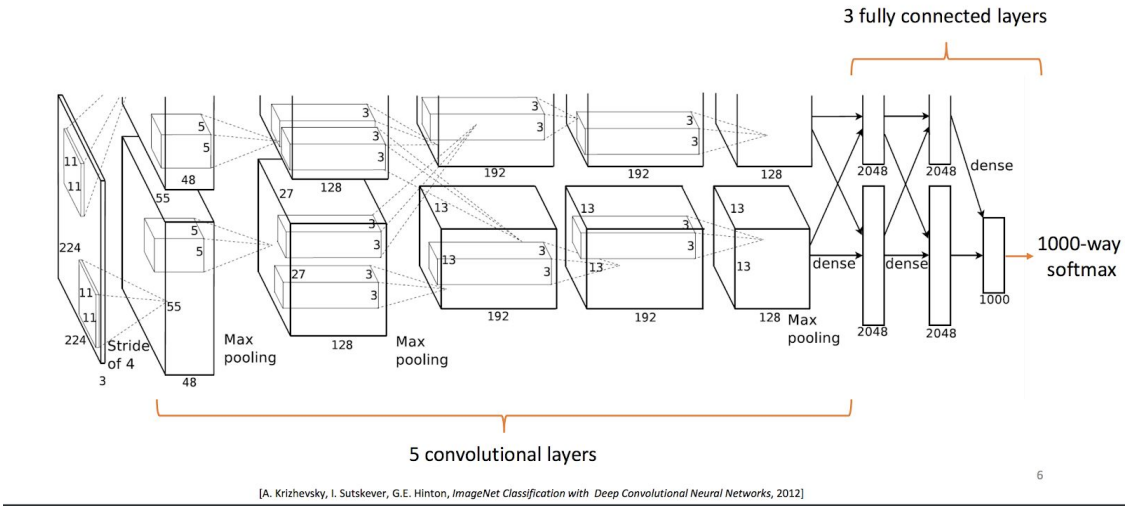

Figure 2: AlexNet architecture has been successfully applied to arrhythmia detection[Isin and Ozdalili, 2017]

One prominent application of CNNs to arrhythmia classification was carried out by the Stanford Machine Learning Group. They trained a deep CNN, and showed that it could be used to distinguish between many types of arrhythmias and beat patterns , and even determine if the signal is too noisy to be evaluated [Rajpurkar et al., 2017]. Their architecture used over 30 layers of convolution, as well as more sophisticated techniques (normalization, dropout, and residual connections) to transform the one dimensional signal. By comparing the F1 score (the harmonic mean of precision and recall) of the classifier with those of expert cardiologists, Rajpurkar and his team showed that their predictions outperformed cardiologists on most of the different arrhythmia types and beat patterns.

Other applications of CNNs to cardiac arrhythmias make very heterogeneous assumptions which are out of the scope of this paper. Interested readers should see (Isin and Ozdalili, 2017), (Yan Zhou et al., 2017), and (Jun et al, 2017).

Pre-trained convolutional neural networks are available for fine tuning.

Because training large CNN architectures is time consuming and risky, many researchers in industry have made their work available as pre-trained feature-extractors. These include AlexNet (figure 1) and GoogLeNet/Inception (Szegedy et al., 2015). Applying these pretrained networks to arrhythmia classification has been effective, and experimentation could yield results closer to the leading edge of this research.

Overall, cardiac arrhythmia classification has made dramatic gains from both a dataset and data science perspective in the past five years. Large, high-quality labelled alarm datasets have made it possible to train models with greater confidence, and the techniques used to yield more subtle analysis have resulted in models that continue to reduce false-positives overall with a variety of tolerances for false negatives. The CalCardiac team will continue to investigate advancements in this field by focusing on a specific type of alarm: premature ventricular contraction (PVC). PVC is the most frequent false positive alarm and source of fatigue in nurses and hospital staff (Drew et al., 2014).

III. Industry Analysis

United States hospitals and medical device manufacturers would be interested in the CalCardiac PVC classifier

The CalCardiac capstone project operates in a space where modern techniques are being deployed in the more traditional healthcare industry. IBISWorld analysts forecast that over the next five years, the US hospitals industry will experience an "annualized rate of 3.3% to $1.2 trillion dollars" [IBISWorld, 2018a], while responding to healthcare reform, and reimbursement trends (IBISWorld, 2018a). Revenue growth is forecasted to be supported by the continually aging population, healthcare reform such as Medicaid and other government or similarly related insurance (38.2% of 2017's industry revenue), and a private health insurance (45.2% of 2017's industry revenue). According to IBISWorld Industry Report 62211, "There are no major players in this industry" (IBISWorld, 2018a) and therefore hospitals are continually seeking to differentiate themselves amongst their competitors as the premiere healthcare provider, while striving to minimize operating costs.

Similarly, IBISWorld analysts forecast the medical device manufacturing industry to experience an annualized growth of 2.9% to $49.3 billion dollars in revenue by 2022. The cardiovascular (CV) device segment of the medical device manufacturing market is saturated, and it accounts for 27.7% of the industry revenue (IBISWorld, 2018b). Competitors such as Medtronics (38.9% market share) and General Electric Company (19%) currently produce similar products with low differentiation and room for innovation.

Both the hospitals industry and the medical device manufacturing industry benefit from differentiating technologies to increase their competitive edge. This creates a

very profitable market to enter as there are motivated buyers and technology providers that would be willing to incorporate this valuable service to gain higher profitability.

Porter's Five Forces analysis illustrates both low and high competitive forces in the industry for data science improvements to patient monitoring.

Porter's Five Forces framework consists of five main business-related considerations: bargaining power of suppliers and customers, threats of new competitors and substitutes, and established rivals with the power of precedent. Improvements to bedside monitors are most likely to be adopted by firms invested in bedside monitor design and manufacturing, who feel that anticipating user requirements is an important aspect of their product.

Because of the novelty of their technology-based suppliers, these companies would depend on a large degree of testing, validation, and robustness analysis on the part of their underlying infrastructure. In addition, regulatory constraints about patient data may make it difficult for users to send crash reports to manufacturers over a public network.

Since the demands of customers reflect a high degree of institutional inertia and risk aversion, buyers of hospital equipment may require certification by oversight agencies or other forms of liability awareness, before purchasing bedside monitors with hype-driven components. Once the initial skepticism is overcome, however, the inertia can work in favor of the companies offering differentiated and novel bedside monitors.

Currently, human operators must frequently attend to alarms generated by patient monitors, which detracts from their ability to relate with patients on a more personal level. This situation, which serves as a substitute to a learning-integrated monitor, is not preferable to a more sophisticated monitor. Other substitutes have not emerged.

Competition is posed by many companies invested in machine learning and artificial intelligence research, as well as academic institutions working on the same thing. Additionally, there are more specialized companies and organizations working specifically in health care. As a result, it is likely that new competition will emerge.

Finally, currently established bedside monitors do not offer integrated machine learning, and so the threat of established competitors is less pronounced than that of new entrants.

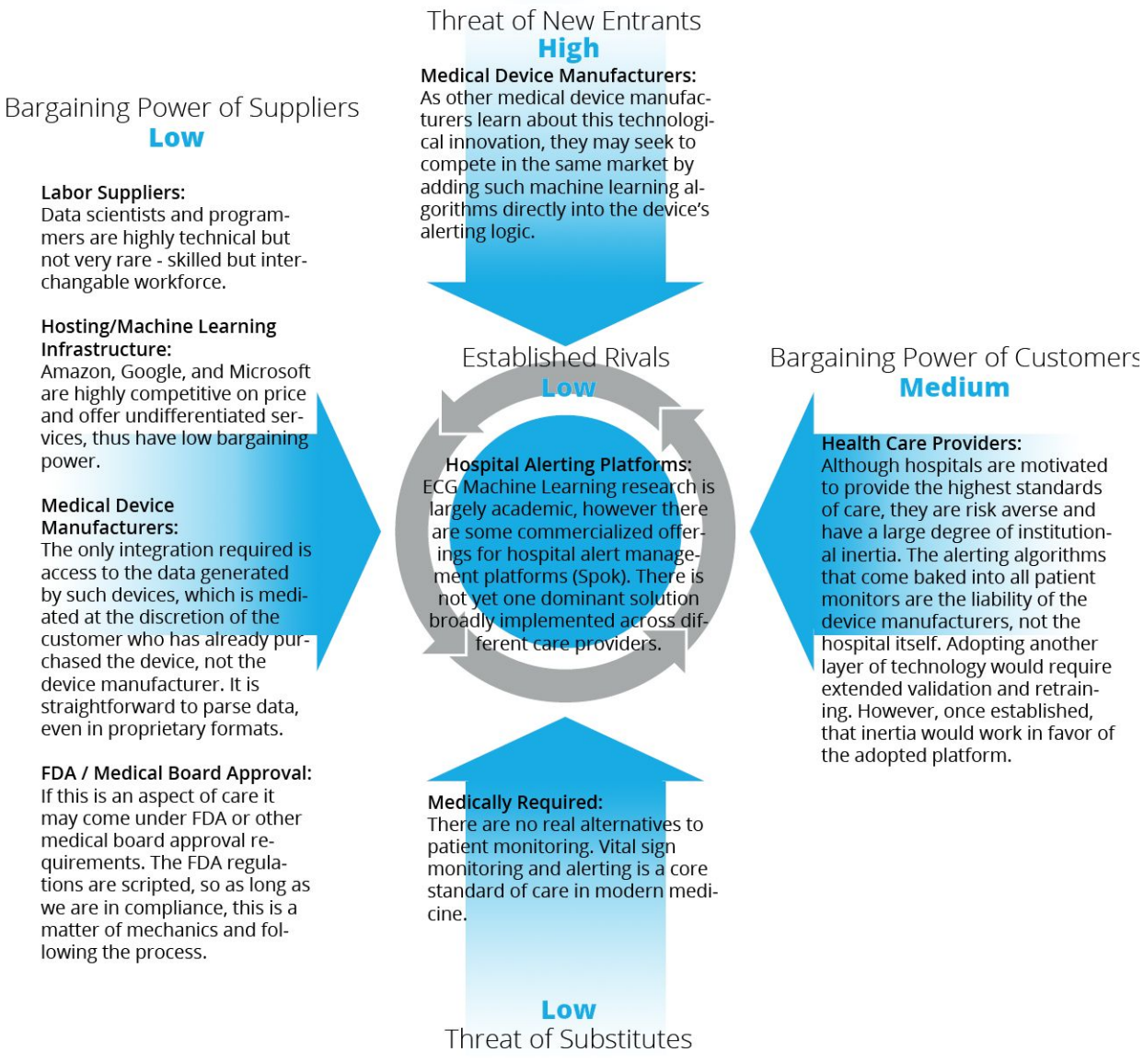

Figure 3: Porter's five forces diagram demonstrates how CalCardiac can add value to the hospital alerting platform industry.

Intelligent functionality should be adopted by patient monitor manufacturers with the power and economic incentive to improve their product offerings.

The end-users for this product are the millions of caregivers and patients who could benefit from automatic interpretation of electrocardiogram (ECG) signals. Medical device manufacturers would be motivated to incorporate this technology into their product offering to differentiate amongst a competitive and saturated industry. Even if device manufacturers lag to incorporate the technology into the devices themselves, hospitals can implement this technology unilaterally in their Electronic Medical Record Systems or in a custom software platform to increase quality of care and reduce medical errors.

IV. Technical Contributions

ECG arrhythmia labels are based on physiological events.

Supervised learning models require a reliable method of generating a true label for premature ventricular contraction (PVC). The methods for identifying an arrhythmia with a ventricular origin (such as PVC) are well-established by the medical community (Malmivuo & Plonsey ,1995.), and rely on several conditions.

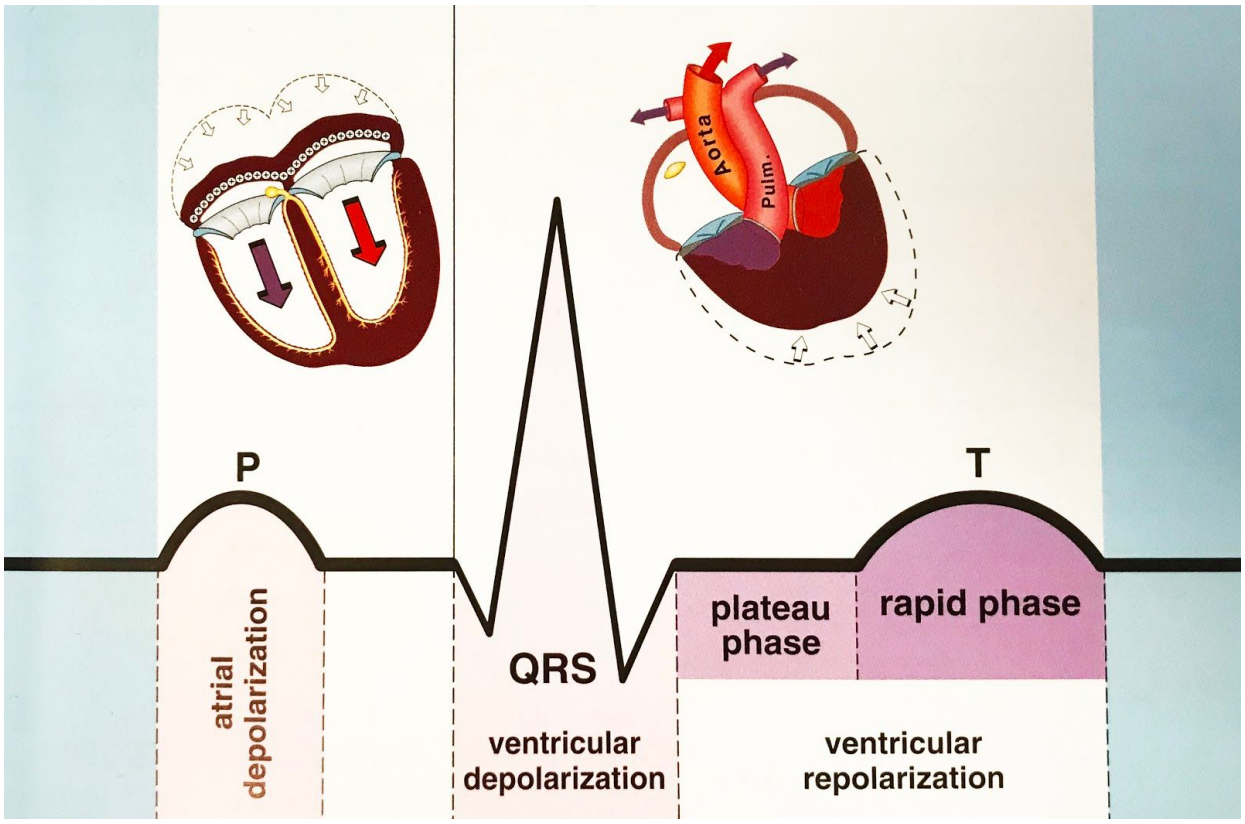

Figure 4: Diagram of the P,QRS, and T wave archetypes for a single healthy heartbeat (Dubin, 2000, pg.29). These correspond to atrial depolarization, ventricular depolarization, and ventricular repolarization, respectively. The atria and ventricles are the upper and lower chambers of the heart, respectively. The atrioventricular node and bundle of His conduction pathway is shown in yellow.

First, there will generally be a widened QRS complex, typically lasting longer than .1 seconds, on account of the electrical impulse that must propagate through bulk cardiac tissue and not along the typical bundle branch pathways. Secondly, the QRS complex will occur earlier than would be expected based on the frequency of the preceding beats. The previous beats are driven impulses originating at the sinoatrial node and thus occur with a well-defined frequency. A PVC impulse has an ectopic electrical origin, and so often will have a premature QRS complex. Third, there will usually not be a P wave associated with the QRS complex of the PVC contraction. This is because the P wave is characteristic of the depolarization of the atria of the heart following signal initiation by the sinoatrial node. Due to the ectopic origin of a PVC impulse, a P wave is not observed in many cases. The final distinguishing feature of a PVC event is a discordant T wave. The T-wave in a normal beat is associated with the repolarization of the ventricles. Observed discordance in PVC is a result of an abnormal repolarization pathway that often results after an atypical-pathway bulk-tissue depolarization of the ventricles.

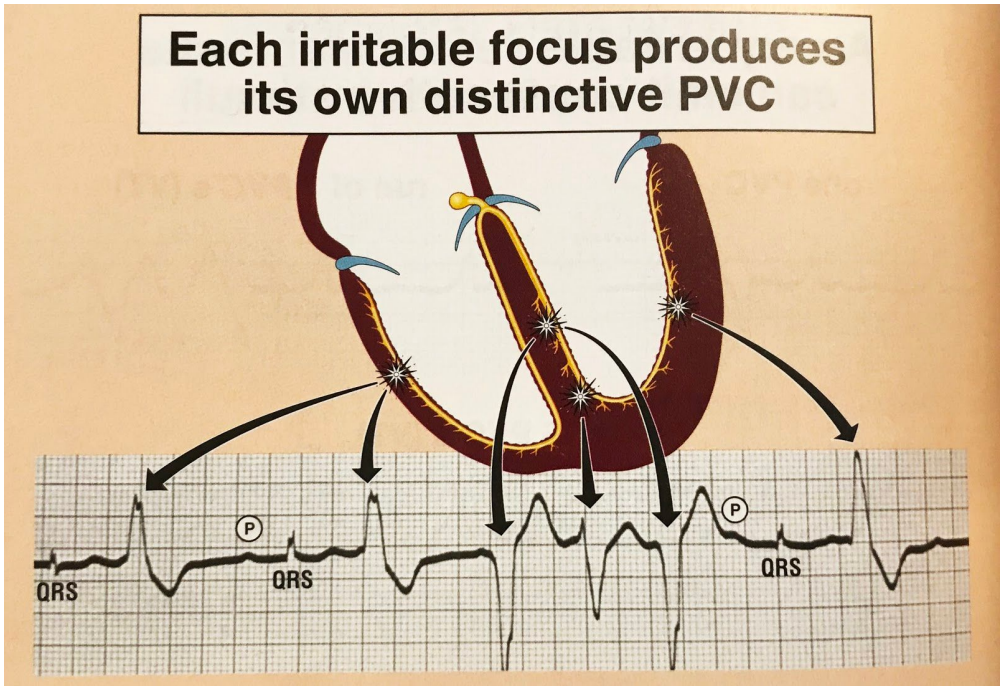

Figure 5: Example waveforms of PVC heartbeats originating from different ectopic nodes (Dubin, 2000, pg.142). Ectopic nodes originate electrical signals like the atrioventricular nodes do, but their placement in the ventricles causes a slower bulk myocardial conduction pathway to be followed rather than the bundles of His, leading to interruption in the normal beat pattern with a wide QRS complex and other abnormal indicators on the ECG.

Our annotation procedure will rely on these indicators. Expert and non-expert annotators are trained according on the same examples and precepts, and studies have shown that non-expert annotation based on simple rule-based systems can be

Positive Indicators	Negative Indicators
Wide QRS Complex	Reference beats are irregular
Early R-wave	Absence of 2 or more positive indicators
No P-wave	
Discordant T-wave	
Compensatory Pause	
Present in at least 4 bipolar leads	

Table 1 : Simple rules are used to identify PVC heartbeats. Beat patterns that are believed to have resulted from lead irregularities should to be labelled as artifacts (Dubin, 2000).

very effective for machine learning (Snow et al., 2008). Events that do not exhibit at least three of these criteria will not be considered PVC events. ECG records where no normal beat pattern can be discerned also cannot be considered to be PVC events.

Building an annotation utility will help create and expand labelled datasets for arrhythmias.

Data for this project comes from two sources: MIT-BIH arrhythmia dataset, and UCSF 2013-2014 dataset. The MIT-BIH database was annotated by professional cardiologists at points along each ECG strip. The annotations fall roughly at each R-peak, and indicate one of several categories (Goldberger et al., 2000).

The UCSF 2013-2014 dataset does not have annotations, only ECG strips. The data was extracted from UCSF's database by querying only PVC-labelled alarms, resulting in over 10 million alarms. Each record has 10 seconds of ECG recordings across

several leads, but they are not professionally annotated to confirm if the recording contains a PVC arrhythmia or does not.

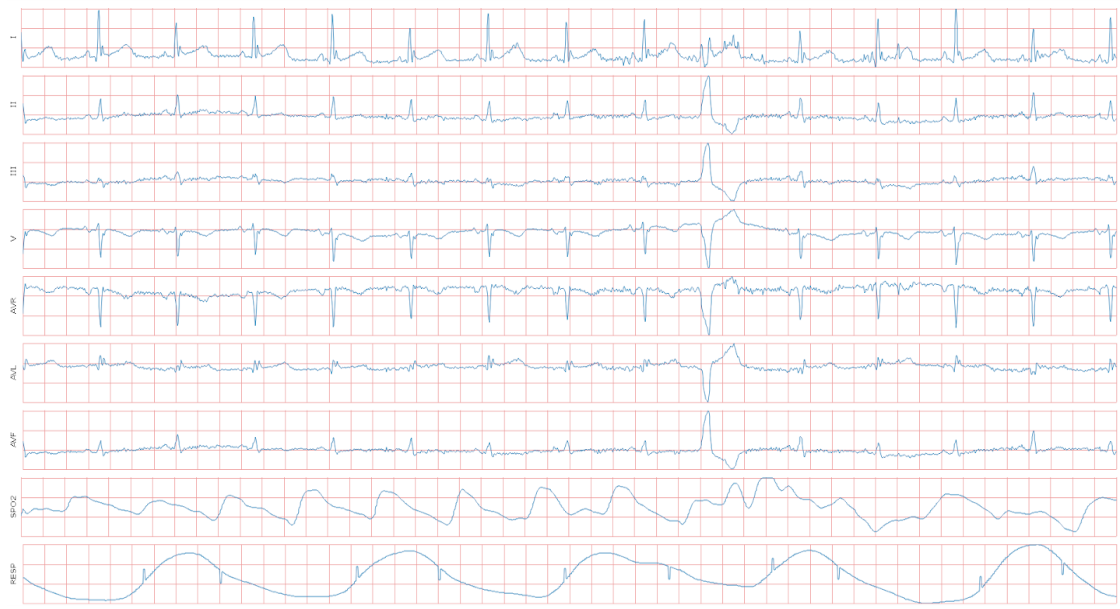

Figure 6: Example 10-second ECG strip of signal data with a PVC arrhythmia occuring after the fifth complex. The complex is premature, the QRS complex is broad, there is no visible P-wave, there is a discordant S-T wave, the event can be seen across multiple leads, and there is a compensatory pause seen at the sixth normal complex.

In order to train any model, we require labelled training and test data. By building a small utility function using Python and Node.js (as seen in the figure below), our team can divide up the UCSF dataset into batches of images and attempt to recreate the same annotation process that the MIT-BIH arrhythmia dataset used, creating a labelled UCSF dataset.

For simplification, the UCSF data will have 3 possible annotations: True-PVC, False-PVC, Artifact. This allows our team to quickly iterate through records and annotate them for model training. Since we are not subject matter experts, records will be annotated by multiple members independently, and consensus on images will be formed.

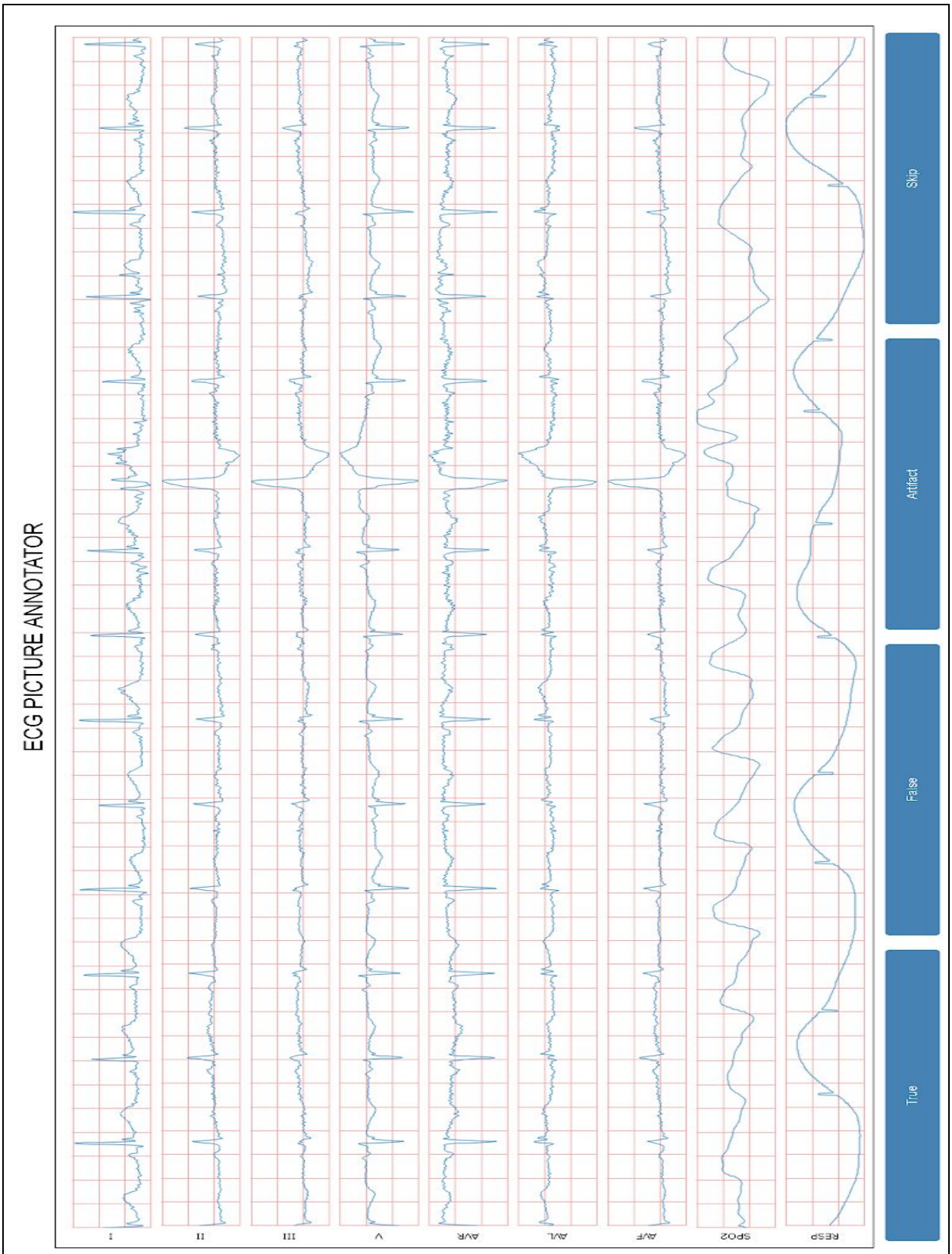

Figure 7: A browser-based ECG annotator, built by CalCardiac, helped generate labeled training data.

The benefit of this annotation utility is that we can expand the MIT-BIH dataset with the UCSF data and leverage them both to strengthen our model approaches.

Monitor alarms with ECG strips can be simulated by statistical processing of community arrhythmia datasets, although caution is necessary because of subtle differences.

The MIT-BIH Arrhythmia Database is one of the earliest raw ECG databases available to the public. It contains half-hour excerpts of two-channel ECG recordings from 48 patients that have been studied by the BIH Arrhythmia Laboratory in the period of 1975 to 1979 . One of the channels is a modified limb lead (MLII) mostly and other is V1. The ECG signals are sampled at 360 samples per second per channel (Moody & Mark, 2001).

All 48 records of half-hour ECG signals were sliced into 0.8-second ECG signal strips, 0.4-second before and 0.4-second after each annotated beat. Here 0.8 second was chosen because through exploratory analysis of the ECG signals we found that the median time between R-R peaks is about 0.8 second. The resulting dataset consist of 112,552 samples of 0.8-seconds ECG signals (each has 288 signal values). Of the 112,552 samples, 105,423 samples were annotated 0 meaning these 0.8-seconds ECG signal strips are not PVC beats while 7,129 samples were annotated 1 meaning these 0.8-seconds ECG signal strips are PVC beats.

We started with the modeling efforts by first considering the dataset containing the entire 111,985 records. The dataset was split into training and test dataset

Model	Accuracy	True Positive Rate	False Positive Rate
Baseline	0.9366	0.0000	0.0000
Logistic - training set	0.9510	0.2949	0.0046
Logistic - test set	0.9472	0.2500	0.0056
Logistic (Fourier transform) - training set	0.9442	0.1804	0.0042
Logistic (Fourier transform) - test set	0.9389	0.1467	0.0070

Table 2a: prediction results of MIT-BIH dataset using different models

ratio of 0.2. Apart from building the logistic model, we also used fourier transformation and then compared the accuracy results.

If all 112,552 observations are to be used for training machine learning models, then the baseline model (which predicts every observation to be 0) will have a quite high accuracy of 93.67% (as can be seen from the above table). Then it's hard to tell how much a machine learning model can improve the prediction accuracy. So from the 105,423 observations which were annotated non-PVC, 7,129 observations were randomly selected and joined with the 7,129 observations which were annotated PVC to form a new smaller dataset. For the smaller dataset, as there are the same amount of observations being 0 and 1, the baseline model will have prediction accuracy of 50.00% (the prediction accuracy for test set is 50.14% due to splitting of the dataset into training and testing subset).

Based on the smaller dataset consisting of 7,129 non-PVC and 7,129 PVC observations, logistic regression model, random forest model and a 2-layer fully connected neural network were built and the model prediction results are shown in Table 2 below. The dataset was splitted into training and testing subset using 0.3 ratio.

The 10-seconds ECG records were also analyzed as part of exploratory data analysis. The median time duration between two consecutive R peaks for patients was found by exploring the non PVC 10-second ECG records. For example, for one patient the R-R peak time was found to be 0.8 seconds (which is in conformation to the normal range of 0.6s - 1). On examining the PVC 10-second ECG signal strip for the same patient, it was found that the RR peak difference was more than the median difference.

Model	Accuracy	True Positive rate	False Positive rate
Baseline	0.5014	0.0	0.0
Logistic regression (training set)	0.7415	0.7272	0.2442
Logistic regression (testing set)	0.7415	0.7342	0.2513
Random forest (training set)	0.8810	0.7622	0.0
Random forest (testing set)	0.8167	0.6418	0.0093
Neural network (training set)	1.0	1.0	0.0
Neural network (testing set)	0.9364	0.9147	0.0420

Table 2b: prediction results of small MIT-BIH dataset using different models

The difference between the maximum and minimum peak was also analyzed . This corresponds to the difference between the R peak and the S peak. For non PVC 10 second records the median difference was found to be 1.37 mV whereas for the PVC record the median difference was found to be 3.4 mV.

These two basic data exploratory techniques would eventually help us select features for the model and further improve the accuracy of the model.

Consistent input data are essential for intelligent algorithms.

The MIT-BIH and UCSF datasets are differ by lead placement, sampling rate, equipment, and channel typology. Generalization of machine learning algorithms away from the training set requires that input data represent samples from a consistent distribution, so it is necessary to create an analogy, as well as perform standardization or segmentation, when inputting data from new sources. In order to maintain this data fidelity, for example, it is useful to downsample high-frequency data. For this downsampling, we used the Fourier resampling method implemented (Pedregosa et al., 2012). The method lends itself to this application due to the smoothness of the ECG signal.

Modern and classical techniques can capture high dimensional structure in spatial and temporal data.

Statistical data modelling techniques can be grouped into two major categories (Hinton, 2007). The first type is characterized by low dimension (less than 100 dimensions), noisiness, lack of structure, and parsimonious modelling. In these datasets, simple models such as linear separation or linear fit are sufficient, and the main problem is to distinguish signal from noise. The study of such methods falls in the domain of classical statistics. The second type is characterized by high dimension (more than 100 dimensions), high degrees of complex structure, and high capacity modelling (using, for example, neural networks). The main problem of this second type is finding a good way to capture the complex high dimensional structure without overfitting artifacts of the dataset. The study of these methods falls in the domain of machine learning (and intersects computer vision as well).

Under the statistical paradigm, the PVC classification problem may require a small number of data features such as the distance between R peaks (R-R interval), QRS width, and R-S slope.

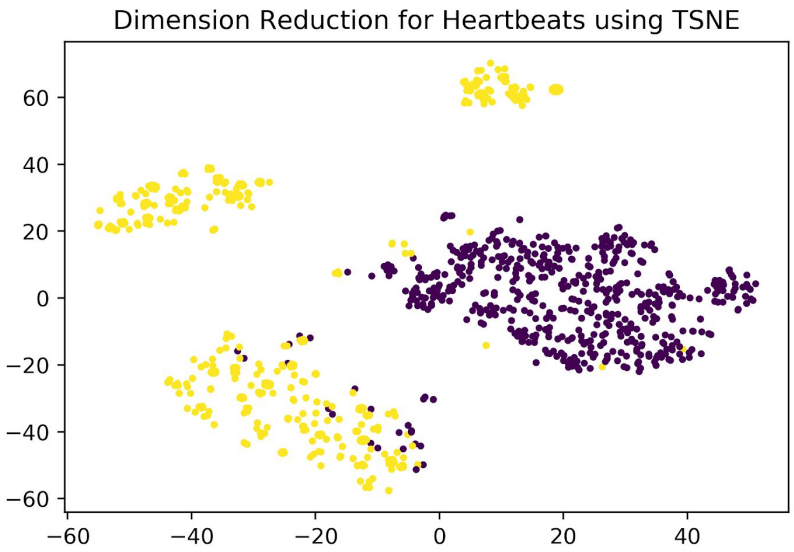

Figure 8: Bootstrapped heartbeats from a clean ECG display complex, high-dimensional structure which can be visualized using dimension reduction techniques (Maaten and Hinton, 2008). The color indicates a whether a heartbeat is marked as PVC. Note that the binning process of the dataset doesn't represent the variability in non-pvc heartbeats.

Under the machine learning paradigm, raw ECG samples (anywhere from 200-500 per second) are treated as independent dimensions. These very high-dimensional points may then be fed directly as input to a well-designed neural computer, mechanistic model (wavelet transform with a human-selected basis), or dimension reduction technique (such as multidimensional scaling, stochastic neighbor embedding, or principal components analysis). The specific procedure would depend on the modelling choices pursued in a particular context.

As a first attempt to understand the structure of the ECG signal, we extracted a single channel (channel I) from a single patient (number 213) in the MIT-BIH-Arrhythmia database, sampled at 360 hertz. From the annotations, we generated a random subset of windows (width=400 samples), which are assumed to be centered on r-peaks of the heartbeat. These windows were then sampled with repetition from class-label buckets, and transformed into two dimensions using t-distributed stochastic neighbor embedding. TSNE attempts to preserve high dimensional neighborhoods using a probabilistic interpretation of transformed data

	Total Accuracy	Precision (PVC)	Recall (PVC)
In-sample, Same patient, bootstrapped	99.5%	99%	100%
Out-of-sample, Same patient, bootstrapped	96%	95%	97%
Out-of-sample, Same patient, All heartbeats	17%	7%	90%

Table 3: performance metrics for single patient heartbeat classifier with clean ECG signal. Trained on n = 200 randomly sampled heartbeat windows from a single patient. The low total accuracy for all heartbeats shows that non-pvc heartbeats (mostly normal) were incorrectly classed as pvc. This effectively recreates the problem of false alarms.

(Maaten and Hinton, 2008). The resulting plot (see figure 3) shows very good separation properties for heartbeats generated this way from the signal.

Based on the degree of separability visible in the above plot, we fit a radial basis function support vector machine [Pedregosa et al. 2012] to the raw points. Unfortunately, the bootstrapped dataset doesn't accurately capture the structure of heartbeats, even for the same patient! The scores in figure 4 show that the limiting class of PVC examples makes it hard to prevent overfitting to data generated according to the bootstrap procedure, even if the data don't intersect. In addition, it is crucial that patient we selected had an extremely clean ECG reading with little noise. These possibilities make the current model interesting as a baseline because it detects a large number of PVCs, but useless in practice due to lack of robustness. In essence, this recreates the problem of a high rate of false alarms.

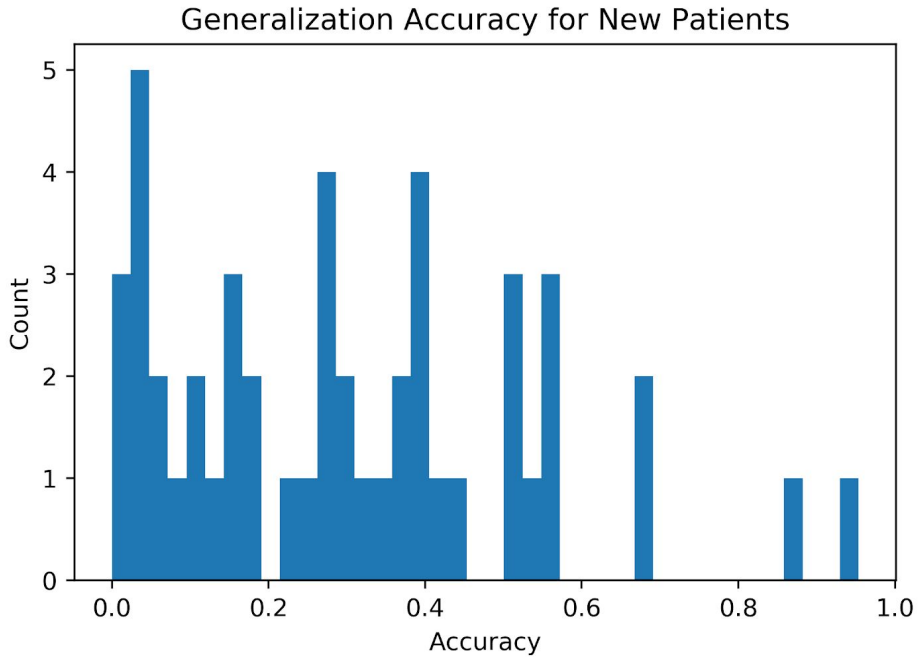

Figure 9: Accuracy of single-patient-based classifier on other patients in the BIH dataset. This plot shows the classifier's failure to generalize, and the large number of very low accuracy patients suggests that the decision boundary learned is actually misleading.

Generalization to new patients presents a major challenge.

Figure 5 shows the shapes of lead-I heartbeat windows for a variety of patients. These examples demonstrate the variability in the MIT-BIH dataset. Firstly, although the labels are frequently related to R-peaks, they are not well-centered enough to provide a true point of reference. This observation may be related to the inherent ambiguity of choosing a particular point for the R-peak, and some appear to adopt the convention of placing the annotation immediately before QRS complex. Secondly, the patient we selected for training captures only a tiny fraction of the possible shapes of various arrhythmias, normal heartbeats, and noisy signals, which can unexpectedly thwart induction by the classifier. This observation is called the curse of dimensionality, and appears when the training examples do not reflect a larger structure which would enable generalization. Thirdly, the bootstrapping procedure we used limited the number non-PVC examples dramatically. As a result,

the classifier captured a large number of PVCs, but did not distinguish them well from non-PVCs. Again, this effort effectively recreated the problem of false-alarms.

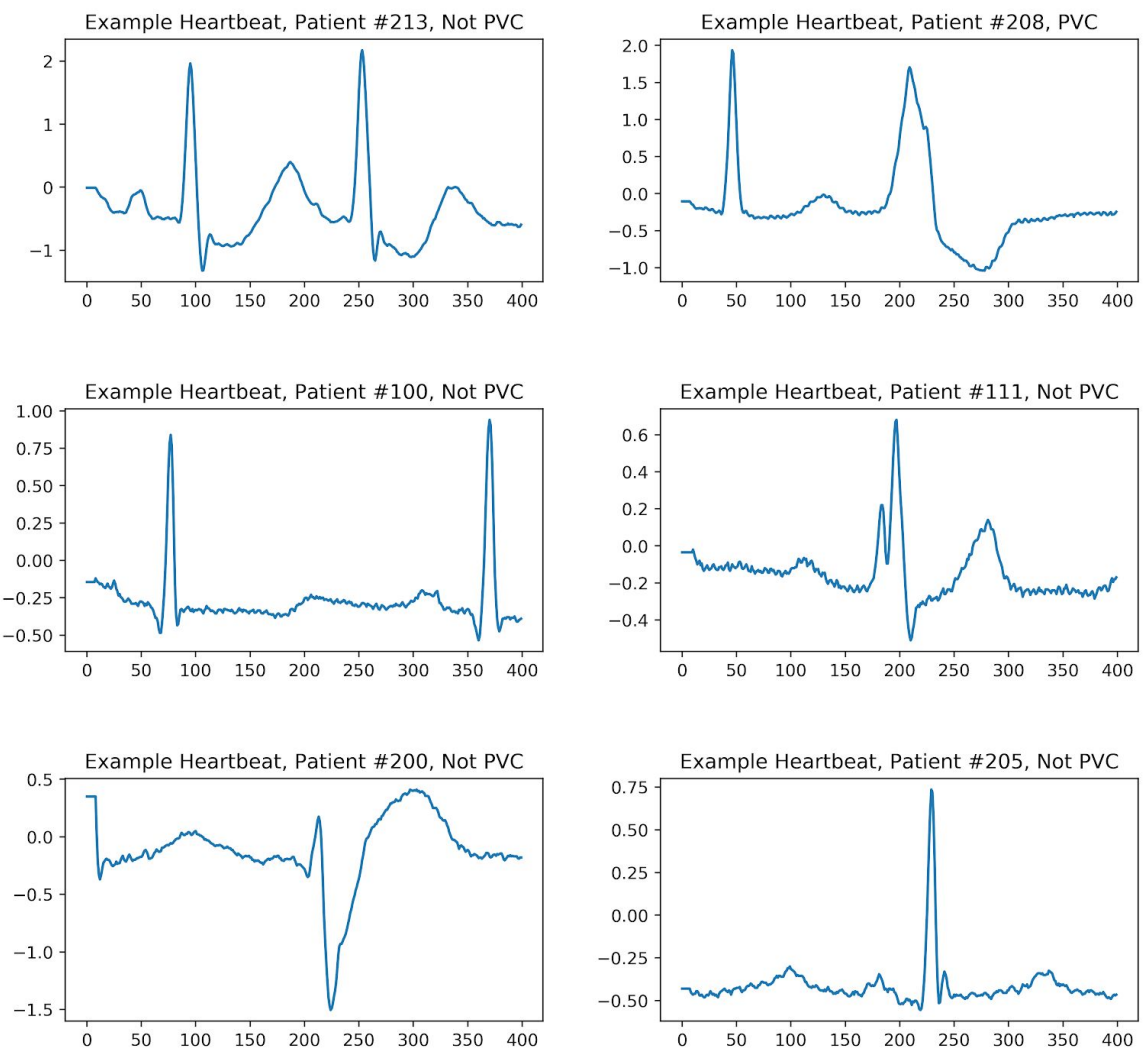

Figure 10: heartbeat windows from a variety of patients in the MIT-BIH dataset demonstrate the differences in scale/intensity of the signal, position, and shape of PVC and non-PVC measurements. Normalization and segmentation could create more comparable data (Goldberger et al., 2000).

Anomaly detection with K-Means clustering algorithms can be used for QRS complex detection in time series ECG signal data.

In 1985, Jiapu Pan and Willis J. Tompkins published a paper in IEEE in March 1985 titled “A Real-Time QRS Detection Algorithm” (Pan and Tompkins, 1985) . They developed an algorithm to detect QRS complexes in ECG signals in real-time by performing various analyses and filtrations on the signals. This approach showed that ECG signal analysis could be done in real-time and illustrated how it could be done.

Clustering algorithms are another approach for anomaly detection in time-series data. After studying the MIT-BIH dataset, as well as the UCSF dataset, we were looking for an anomaly, in this case it was PVC, in a stream of ECG signals behaving normally. Similar to the approach that Pan and Tompkins took, we tried a multiple step approach to see if we could teach a model to identify normal behaviour heartbeats or anomalies, and then to determine if the anomaly behaved like a PVC arrhythmia or another arrhythmia. In order to test out this method, the first step would be to use K-means clustering to learn the various waveform shapes of normal heart beats, and then the algorithm would try to recreate newly presented data with these cluster centroids and monitor the reconstruction error rate. If the newly presented data exceeded a threshold of reconstruction error it would indicate that the new signal data has a different waveform shape than what was learned to be normal, and therefore would have a high probability of being an anomaly. The second step would be to use K-means clustering on known PVC waveforms and then re-test the new data to see if it could be reconstructed with PVC learned waveform shapes. If it could be, then there would be a high probability that the new anomalous data would be a PVC-type arrhythmia as opposed to another anomalous arrhythmia like atrial fibrillation.

The approach first read in the ECG signal data, and split the signal into time windows that took snapshots of the signal of a predetermined length. The window would then “slide” a finite amount of time to reveal a new, overlapping window segment of the waveform. In order to normalize the waveform in each window segment so that they would not create artificial errors during reconstruction, each waveform captured in a window would be multiplied by a sine wave which forces each window to start and end at 0, eliminating the possibility that the algorithm learns fictitious waveforms due to windows not starting and ending at a normalized point, and instead only learn the core of the window waveform shape in the center of each window.

The next step once the windows are created are to use K-Means clustering to teach the model what waveform shapes are to be considered normal, creating a learned library of waveforms. This is where fine-tuning is required as the number of clusters affects end performance and training time. Using the clusters, we can split up new data to reconstruct them with the waveform library, and monitor the reconstruction error rate. If the error rate surpasses a certain threshold (to be determined by testing), then the probability of the data having an anomaly is high.

We are exploring this approach using the MIT-BIH dataset of PVC and normal heartbeat waveforms. If it is possible to construct an anomaly detector using just normal ECG data, then it will be tested on PVC data to see if the system can discern PVC anomalies versus other anomalous waveforms. Finally, it will be tested to see if the algorithm can generalize to the UCSF data from a similar channel.

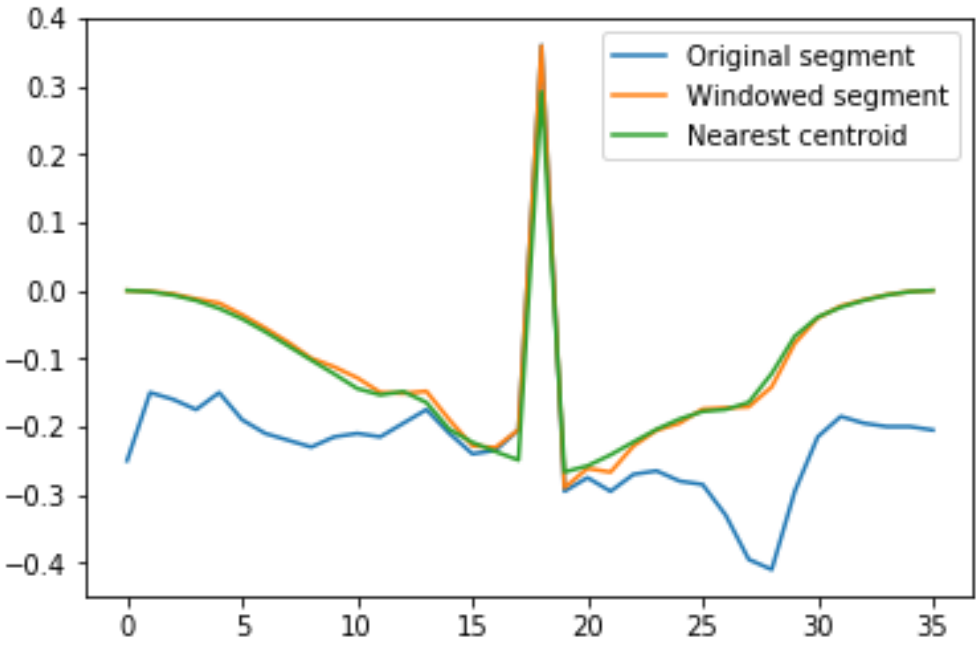

Figure 11: Anomaly detection approach using K-Means clustering. Here is the original waveform of a 'Normal' heartbeat pulled from the MIT-BIH database in blue, with a normalized version of the heartbeat using windowed segments of a learned shapes library of ECG wavelets in orange, and finally a reconstruction of the waveform using the cluster centroids in green. The reconstruction error (difference between orange and green plots) is then computed and can then be used with a threshold detector to raise an alarm if the anomaly exceeds a prescribed threshold.

Using the TensorFlow Object Detection API to build an ECG classifier

Since the objective of this project can be interpreted as identifying a PVC waveform in a ECG data, we attempted to build an ECG waveform classifier to recognize the PVC waveform shapes from other waveform shapes. Using the MIT-BIH labelled dataset, images of PVC arrhythmias and non-PVC heartbeats, an image classifier was built using limited training data. Images were created by creating a picture of 1 seconds worth of ECG data centered on either a normal or PVC heartbeat or another annotated beat. The images were then formatted to be used to train a pretrained CNN using open source architecture by leveraging a

pretrained model from Google's Object Detection API for transfer learning and training on Google Cloud Computing.

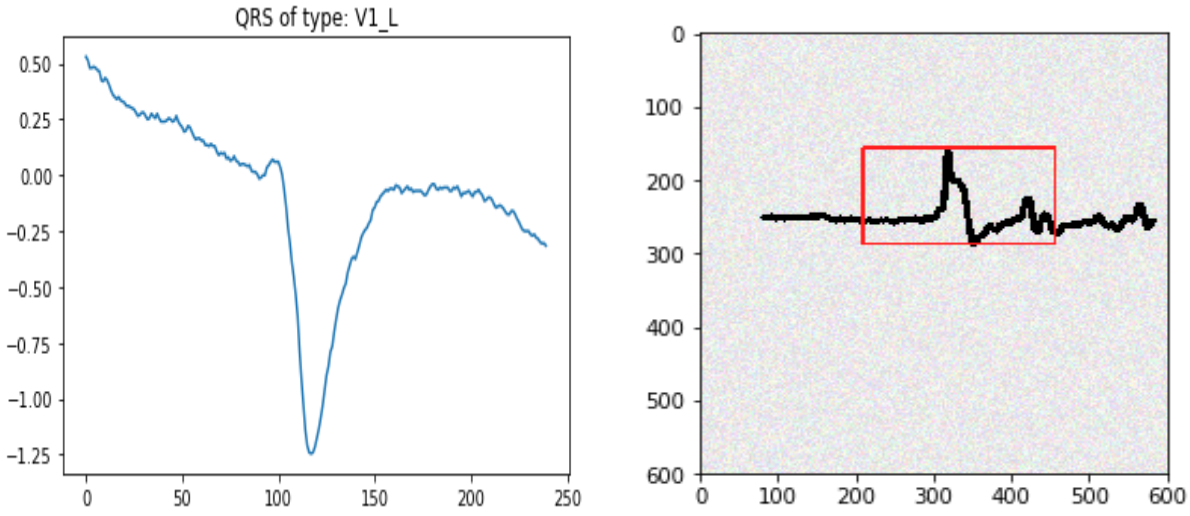

Figure 12: Using the TensorFlow Object Detection API, Google Cloud Computing, and MIT-BIH dataset to use a pretrained model for transfer learning to analyze and detect ECG heartbeats. Left image: plot of isolated ECG data from a V1 channel of left bundle branch block beat. Right image is preparing files to run on TensorFlow and Google Cloud with Google's Object Detection API to learn ECG signals and classify them bounded in the red box.

V. Conclusion

This project originally started out as an effort to identify and suppress PVC alarms in ECG signal data; however, as the project unfolded, it evolved to reveal the intricacies and challenges across multiple components of this common goal. The dataset provided to our group was unlabelled and massive, so without subject matter expertise, a proven model or method to perform unsupervised learning on the data, nor the time to label the data to perform supervised learning on the data, we had to reevaluate our approach.

While we annotated 3000 UCSF alarms, this was more of an exercise in understanding our data, and building up a small, new labelled data set that we could perhaps leverage in the future. Without a clear and promising method of leveraging the annotated data, the team fell back to a more famous and widely used dataset called the MIT-BIH arrhythmia dataset.

In an effort to maximize concept generation, our team explored numerous models and approaches to generate baselines and models to compare and possibly combine in a future ensemble model. Efforts on the MIT-BIH dataset included classical signal processing, regression, random forest, t-SNE, CNN, K-Means, and transfer learning. Not all have been successful, but they are important efforts nonetheless.

The results from our small experiments have been discrete and with varied success, but the proof of concept of being able to identify PVC and reliably classify a PVC-type arrhythmia versus other ECG signals is there.

Our next steps are to generalize our models to new patients and new types of data streams. For example, the MIT-BIH dataset is different in form in comparison to the UCSF dataset. Being able to generalize our model to various patients over time and datastreams is where we look to focus our future efforts.

Appendix A: Tables and Figures

References

Note: Original bibtex file available on request.

Bai, Y., Do, D., Ding, Q., Palacios, J. A., Shahriari, Y., Pel- ter, M. M., Boyle, N., Fidler, R., and Hu, X. (2017). Is the Sequence of SuperAlarm Triggers More Predictive Than Sequence of the Currently Utilized Patient Monitor Alarms? IEEE Transactions on Biomedical Engineering, 64(5):1023–1032.

CAST Investigators (1989). Effect of Encainide and Flecainide on Mortality in a Randomised Trial of Arrhythmia Suppression After Myocardial Infarction. The New England Journal of Medicine, 321(6):406– 412.

De Oliveira, L., Andre˜ ao, R., and Sarcinelli-Filho, M. (2008). Classification of premature ventricular beat using Bayesian networks. In HEALTHINF 2008 - 1st International Conference on Health Informatics, Proceedings.

de Oliveira, L. S. C., Andreao, R. V., and Sarcinelli- Filho, M. (2011). Premature Ventricular beat classification using a dynamic Bayesian Network. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2011:4984–4987.

Drew, B. J., Harris, P., Z`egre-Hemsey, J. K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., and Hu, X. (2014). Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients. PLoS ONE, 9(10).

Dubin, D. (2000). *Rapid interpretation of EKG's: An interactive course* (6th ed.). Tampa, Fla.: Cover Pub. Co..

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals. Circulation, 101(23):e215–e220.

Goodfellow, Ian, Bengio, Yoshua, Courville, A. (2016). Deep Learning. MIT Press.

Graham, R., Mccoy, M. A., and Schultz, A. M. (2015). Strategies to Improve Cardiac Arrest Survival : A Time to Act.

Hinton, G. (2007). 2007 NIPS Tutorial on: Deep Belief Nets. In 2007 NIPS Tutorial on: Deep Belief Nets.

Hu, X., Sapo, M., Nenov, V., Barry, T., Kim, S., Do, D. H., Boyle, N., and Martin, N. (2012). Predictive combinations of monitor alarms preceding in-hospital code blue events. Journal of Biomedical Informatics, 45(5):913–921.

IBISWorld. (2018). US INDUSTRY REPORTS (NAICS) – Hospitals. Retrieved March 15, 2017 from IBISWorld database.

IBISWorld. (2018). US INDUSTRY REPORTS (NAICS) – Medical Device Manufacturing. Retrieved March 15, 2017 from IBISWorld database.

Isin, A. and Ozdalili, S. (2017). Cardiac arrhythmia detection using deep learning. In Procedia Computer Science, volume 120, pages 268–275.

Jun, T. J., Park, H. J., Minh, N. H., Kim, D., and Kim, Y. H. (2017). Premature ventricular contraction beat detection with deep neural networks. In Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016, pages 859–864.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Alexnet. Advances In Neural Information Processing Systems, pages 1–9.

Maaten, L. V. D. and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research 1, 620(1):267–84.

Malmivuo, J. (2004). Bioelectromagnetism. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, page 5217.

Moody, G. B. and Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database.

Pan, J. and Tompkins, W. J. (1985). A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering, BME-32(3):230–236.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E´. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., and Ng, A. Y. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. stanfordmlgroup.

Salas-Boni, R., Bai, Y., Harris, P. R. E., Drew, B. J., and Hu, X. (2014). False ventricular tachycardia alarm suppression in the ICU based on the discrete wavelet transform in the ECG signal. Journal of Electrocardiology, 47(6):775–780.

Salas-Boni, R., Bai, Y., and Hu, X. (2015). Cumulative Time Series Representation for Code Blue prediction in the Intensive Care Unit. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2015:162–7.

Service, C., Us, A., and Locations, G. (2013). Insights & Publications How big data can revolutionize pharmaceutical R&D. McKinsey Global Institute, pages 1–5.

Shahriari, Y., Fidler, R., Pelter, M., Bai, Y., Villaro- man, A., and Hu, X. (2017). Electrocardiogram Signal Quality Assessment Based on Structural Image Similarity Metric. IEEE Transactions on Biomedical Engineering.

Snow, R., O'Connor, B., Jurafsky, D., and a.Y. Ng (2008). Cheap and fast but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (October):254–263.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 07-12-June, pages 1–9.

Wallis, L. (2010). Alarm Fatigue Linked to Patient's Death. . American Journal of Nursing, 110(7):16.

Winters, B. D., Cvach, M. M., Bonafide, C. P., Hu, X., Konkani, A., O'Connor, M. F., Rothschild, J. M., Selby, N. M., Pelter, M. M., McLean, B., and Kane-Gill, S. L. (2018). Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms with Intent to Reduce Alarm Fatigue.

yan Zhou, F., peng Jin, L., and Dong, J. (2017). Premature ventricular contraction detection combining deep neural networks and rules inference. Artificial Intelligence in Medicine, 79:42–51.