Predicting and understanding bronchopulmonary dysplasia in premature infants

Laney Kuenzel

Bronchopulmonary dysplasia (BPD) is a serious but poorly understood lung disorder affecting premature infants. In this work, we use machine learning techniques to characterize the physiological signals (e.g., heart rate, breathing rate) of infants with BPD. We identify physiology that often precedes clinical interventions for BPD and laboratory measurements known to be predictive of BPD, and create a physiology-based classifier for BPD with strong performance on our data set. Such a classifier could easily be adopted by a hospital, allowing doctors to gauge an infant’s BPD risk using only noninvasive measurements already recorded for all infants. We also discuss several findings of medical interest concerning particular features—including signal frequency content and inter-signal correlation—with predictive value for BPD.

Bronchopulmonary dysplasia (BPD) is a lung disorder that affects infants, primarily those born prematurely. Defined as the requirement for oxygen therapy for at least 28 postnatal days, BPD occurs in nearly a third of infants with birth weight under 1000 grams [1]. BPD is associated with far-reaching negative consequences such as further respiratory problems, cerebral palsy, and cognitive impairment [2].

Unfortunately, BPD is one of the most poorly understood complications of prematurity. In particular, there is no consensus on the pathogenesis of the disease. Among the commonly hypothesized causes of BPD are ventilator-induced injury, lung immaturity, lung inflammation due to infection, and genetic predisposition [2,3].

Previous work on predicting BPD has focused primarily on correlating eventual BPD diagnosis with laboratory measurements, medicine administrations, and mechanical ventilator settings. Such studies have consistently shown BPD to be significantly associated with certain abnormal blood gas levels (e.g., low blood pH) as well as aggressive ventilation [4-6].

Despite the abundance of studies seeking to identify factors associated with BPD diagnosis, a large gap still exists in the body of literature on BPD prediction: very little is known about what characterizes the physiological signals (such as heart rate and respiratory rate) of infants eventually diagnosed with BPD. We believe that this area is under-explored largely because it has only recently become possible to obtain fine-grained physiological time series data for hospitalized infants. We were fortunate enough to have access to this type of data, enabling us to conduct novel research on the relationship between physiological signals and BPD. This endeavor was particularly exciting due to its potential to generate important new medical knowledge.

In our search for physiological signatures for BPD, we focused specifically on three signals: heart rate, respiratory rate, and oxygen saturation (a measure of the amount of oxygen carried by the blood). We had two main reasons for choosing these particular signals.

First, it makes sense from a biological perspective that lung problems would manifest in these three signals. The lungs serve to introduce oxygen from inhaled air into the bloodstream and to release carbon dioxide from the blood as exhaled air. Therefore, we would expect infants with poor lung function to exhibit low oxygen levels (motivating our use of the oxygen saturation signal) and high carbon dioxide levels. Furthermore, in infants with lung problems, we would expect to observe altered patterns in the breathing rate and heart rate (motivating our use of the respiratory rate and heart rate signals) as the body attempts to respond to the blood gas imbalance by adjusting the amount of air entering and exiting as well as the speed with which the blood is circulating.

Second, these three signals are recorded noninvasively for every infant in a standard neonatal intensive care unit (NICU). Consequently, a predictive model for BPD based only on features of these three signals could easily be adopted by any NICU. This is not the case for a model with features that rely on more invasive, expensive, or nonstandard measurements.

For these reasons, we formulated our overall objective: to understand whether and how BPD manifests in an infant’s heart rate, respiratory rate, and oxygen saturation level. More concretely, our goal was to construct a predictive model for BPD based on these three physiological signals.

I. Motivation and Real-World Applications

Before describing our efforts to create a predictive model, we first discuss how such a model would be useful in the real world. Suppose we had a model that could accurately predict BPD based on physiological signals alone. By examining how the model makes its predictions, we could identify the common characteristics of infants eventually diagnosed with BPD. This would constitute a valuable contribution to our currently limited understanding of the disease’s pathophysiology. Insight into the emergence and evolution of BPD, as well as its effect on an infant’s body systems, would help doctors and researchers find effective ways to treat and prevent the disease.

Yoder et al. [6] point out another important application of predictive models for BPD: study design. Many therapies have been proposed to prevent BPD, and some of them may have dangerous side effects. If researchers had access to a model that could predict the likelihood of BPD, they could include only those infants with high BPD risk in their studies, thus sparing low-risk infants from unnecessary and potentially dangerous interventions.

Most importantly, a predictive model for BPD could be used by doctors in an NICU setting. As described in the introduction, the heart rate, respiratory rate, and oxygen saturation of each infant in an NICU is continuously monitored. Thus, an NICU could adopt a computational model using this data with no changes to their existing monitoring procedure. Doctors could then use the output of the model to make informed decisions about treatments. As a concrete example, corticosteroids administered shortly after birth can reduce the risk of BPD, but they can also have serious negative effects, such as gastrointestinal bleeding and
dangerous thickening of the heart muscle [7]. Thus, the decision of whether to administer corticosteroids to a given infant is a very difficult one. With more accurate information about an infant’s BPD risk level, doctors would be better able to make the best choice for the infant’s long-term health.

These examples demonstrate how a predictive model for BPD could substantially improve the treatment of premature infants. With this motivation in mind, we move on to describe the toolset we used in creating our model.

II. Machine Learning Approach
In our efforts to develop a predictive model for BPD, we used techniques from machine learning, a subfield of artificial intelligence that focuses on computer algorithms that use experience to improve at a specific task. In our case, the task was predicting BPD diagnosis in infants based on physiological signals observed in the first few days of life. We hoped to develop an algorithm which, given data on BPD and non-BPD infants, would learn general rules for predicting BPD in other infants.

We will now define the machine learning concepts and terminology used in the remainder of this paper. Readers already familiar with machine learning can proceed to Section III.

A. Basic Definitions
We refer to data points as examples. Positive examples refer to those taken from infants with BPD, and negative examples refer to those taken from infants without BPD. We say that positive examples have label 1 and negative examples have label 0. A classifier is an algorithm that takes an example as input and outputs a binary prediction. During the training phase, the algorithm uses labeled data, or training examples, to learn prediction rules. During the testing phase, the algorithm makes predictions for new examples (testing examples) without labels. A “good” classifier is one that makes accurate predictions on the testing examples (i.e., predictions which agree with the true labels).

B. ROC Curve and AUC
To quantify the accuracy of predictions made by a classifier, we use the receiver operating characteristic (ROC) curve. Suppose that we have a classifier which, rather than simply outputting a binary prediction for each example, assigns each example a score indicating how likely it is to be positive. To obtain actual predictions from such a classifier, we specify a cutoff score such that all examples with a score above the cutoff are classified as positive and all others are classified as negative. The ROC curve for a score-assigning classifier is a plot of the true positive rate versus the false positive rate for different possible cutoff scores. Figure I shows the ROC curves for two example classifiers.

C. Logistic Regression
In this work, we used logistic regression, a common and relatively simple classification algorithm. Before explaining logistic regression, let us first define a hyperplane in n-dimensional space as an (n-1)-dimensional subspace, possibly translated by some distance from the origin. A hyperplane in two-dimensional space is just a line, and a hyperplane in three-dimensional space is an ordinary plane. Informally, logistic regression attempts to find a hyperplane in the example space that separates the positive examples from the negative examples as fully as possible. During the training phase, the algorithm learns a weight corresponding to each input variable. Together, these weights specify a hyperplane.

D. Overfitting
One large challenge in machine learning
is a phenomenon called overfitting.
Overfitting occurs when a classifier—or more generally, any learning algorithm—learns a prediction rule based not only on the data’s underlying patterns but also on the noise present in the training data. The resulting classifier performs poorly on new testing examples because its rules do not generalize well. Figure II illustrates the concept of overfitting.

Overfitting is more likely to occur when the number of training examples is small, since a learning algorithm has less information to use in distinguishing between true patterns and noise in the data. Similarly, high-dimensional data increases the risk of overfitting since there are more ways in which noise can coincidentally look like a pattern to the classifier.

We can often avoid overfitting by transforming raw data into a lower-dimensional representation prior to running it through a learning algorithm. Consider a scenario in which we have a day’s worth of minute-interval heart rate and respiratory rate data for each infant in our training set. We could simply provide our algorithm with the raw data values, but overfitting would then likely occur due to the high dimensionality of the data. A better approach would be to calculate the mean heart rate and mean respiratory rate and present these aggregate quantities to our algorithm.

E. Features
In the example above, we would refer to the mean heart rate and mean respiratory rate as features. More generally, features are the new variables that result when we transform our data from raw form into another representation. We say that a feature is informative or predictive with respect to a particular classification task if it helps us distinguish between positive and negative examples. In the case of BPD prediction, we would expect that mean heart rate is probably an informative feature, whereas the number of letters in the infant’s last name is most likely not informative.

In logistic regression, the weight learned for a given feature is indicative of how heavily the algorithm relied on that feature to separate the positive and negative examples (with weights high in absolute value corresponding to the most useful features and weights close to zero corresponding to the least useful).

Now that we have provided the necessary background information, we move on to describing our attempt to apply machine learning techniques to predict BPD.

### III. Selection Criteria
We had access to minute-interval time series data collected from monitoring devices attached to premature infants during their entire stay in Stanford Hospital's NICU. For these infants, we also had data on all of the clinical events—e.g., laboratory tests and ventilator setting changes—that occurred during their hospitalization.

Of the infants admitted to Stanford’s NICU between March 2008 and March 2009, we considered those satisfying the following criteria: gestational age ≤ 34 weeks, birth weight ≤ 2000 grams, length of life ≥ 28 days (thus allowing for BPD diagnosis), and availability of ≥ 5000 minutes of monitor data. Of these infants, 30 with a positive BPD diagnosis were included in the study. As negative examples, an additional 37 infants were chosen at random from those diagnosed with respiratory distress syndrome, an indicator of breathing problems at the time.

![Fig. 2 An illustration of the concept of overfitting.](image)

While the boundary in (a) successfully separates all positive examples from negative examples, it likely would not generalize to further data as its shape depends highly on the particularities of these data points. The boundary in (b) does not perfectly separate the examples, but it seems to give a much better fit to the underlying pattern in the data without capturing the noise.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Full Set</th>
<th>Match Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dextrose administrations</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>Number of ISTAT blood gas measurements</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td>Number of ventilator setting increases</td>
<td>0.82</td>
<td>0.62</td>
</tr>
<tr>
<td>Number of ventilator FiO&lt;sub&gt;2&lt;/sub&gt; setting increases</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>Number of ventilator rate setting increases</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>Number of 10% dextrose administrations</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>Maximum airway resistance measurement</td>
<td>0.84</td>
<td>0.74</td>
</tr>
<tr>
<td>Mean blood pH measurement</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>Minimum blood pH measurement</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>Maximum blood carbon dioxide measurement</td>
<td>0.79</td>
<td>0.84</td>
</tr>
<tr>
<td>Range of blood carbon dioxide measurements</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>Range of blood pH measurements</td>
<td>0.74</td>
<td>0.87</td>
</tr>
<tr>
<td>Minimum blood oxygen measurement</td>
<td>0.74</td>
<td>0.77</td>
</tr>
<tr>
<td>Median RR range over 5-minute windows</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td>Mean abs. diff. between consecutive RR values</td>
<td>0.77</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1 Conservative mean AUCs for top features.
of birth, but not of BPD.

In this set of infants, low gestational age and low birth weight were found to be highly predictive of BPD, with areas under the ROC curve (AUCs) of 0.93 and 0.91, respectively. For this reason, we also created an age/weight-matched set of 20 infants (10 with BPD and 10 controls), for which gestational age and birth weight had much lower AUCs of 0.62 and 0.59, respectively.

### IV. Initial Feature Exploration

For each infant, we had a large quantity of data, including several long physiological time series and information on tens of thousands of diverse clinical events. As a first step, we wanted to determine which of the many available pieces of data would be most useful to us for predicting BPD. We drew from the BPD literature, our discussions with Stanford NICU clinicians, and our observations from visualizing the data to identify 160 potentially interesting features of the first 5000 minutes of data available for each infant.

Around forty of these features were functions of the physiological signals, such as mean, range, standard deviation, and amount of time below or above threshold values. Sixty more features were related to laboratory measurements (e.g., platelet count and blood oxygen level) identified in previous studies as predictive of BPD. For each type of measurement, we included as features the first, mean, minimum, and maximum values, the range of values, and the number of measurements taken. We also included about thirty features related to ventilator type and settings. Rounding out the set of features we considered were the numbers of administrations of over twenty types of medicine.

To determine the predictive value of each feature, we computed the conservative mean AUC (with \( k = 10 \) folds), as proposed by Khosla et al., in order to penalize features sensitive to variations in sampling.

The fifteen top-performing features across both the full set and the age/weight-matched set are displayed in Table I along with their conservative mean AUCs. We observed that these top features fell into three categories:

(i) Six of the features measured the frequency of various clinical interventions (namely medicine administrations, blood draws, and ventilator adjustments) ordered by the NICU doctors. The strong performance of these intervention frequencies as features suggests that Stanford clinicians have a good sense of which infants are at highest risk for BPD or other complications and therefore require the most treatment and surveillance.

(ii) Seven of the features were functions of measurements other than the physiological signals. More specifically, six were related to blood gas values and the last was a function of the airway resistance measurement, which is taken only for infants on a particular type of ventilator. Interestingly, the frequency of blood draws was generally more predictive of BPD than the actual measurement values resulting from those draws. Assuming that the NICU doctors order more blood draws for the infants that they deem sicker, this result suggests that the doctors’ assessment of an infant’s health status is richer in predictive information than the blood gas levels for that infant.

(iii) The remaining two features were functions of the respiratory rate. We were surprised to find that, of the over forty physiological features we considered, only these two were among the fifteen most informative overall.

As described in the introduction, our aim in this project was to shed light on the relationship between physiological signals and BPD. We found in this initial exploration that most of the highly predictive features for BPD were directly related not to physiology but instead to intervention frequencies and laboratory measurements.

### V. Motivation for “Proxy” Approach

Essentially, we decided to examine the relationship between our three physiological signals and the interventions and measurements corresponding to the features in categories (i) and (ii) above. More specifically, we hoped to identify physiological signatures characterizing the time when a given intervention is ordered, in the case of category (i), or the time when an infant has a certain measurement value, in the case of category (ii). Our rationale was that because the intervention frequencies and measurement values were so predictive of BPD, good physiological proxies for them would likely be predictive of BPD as well.

In addition to their potential value in BPD prediction, physiological proxies for the informative features would be useful in several other ways. In the case of category (i), we could use physiological signatures for a given intervention to create a tool for reporting whether an infant is exhibiting physiological signatures that typically precede that intervention. Such a tool would help NICU doctors decide whether and when a certain intervention is necessary.

Physiological proxies for the blood gas and resistance measurements in category (ii) would be valuable for two reasons beyond BPD prediction. First, we could incorporate such proxies into a tool that would let doctors noninvasively obtain a rough estimate of blood gas levels or resistance for an infant. This type of tool would reduce the number of invasive blood draws performed, benefiting both the infant’s health and the hospital’s budget. Second, independent of any prediction application, an understanding of how abnormal measurements manifest in the physiological signals of premature infants would constitute valuable medical knowledge.

Why did we believe that physiological proxies for our predictive features would exist at all? In terms of interventions, Stanford NICU clinicians informed us that they decide what interventions to order based partly on an infant’s physiological signals. For example, doctors often consider low oxygen saturation and high respiratory rate as a sign that the ventilator is not
working effectively and requires setting increases. In terms of measurements, we expected that abnormal blood gas values would be reflected in the physiological signals since the body sets the heart and respiratory rates based on blood levels of oxygen and carbon dioxide.

For the above reasons, we believed that physiological proxies for our informative features were likely to exist. The goal of our subsequent experiments was to identify such proxies so that we could ultimately use them to predict BPD.

VI. Experimental Setup

A. Data

In our first set of experiments, we considered five interventions with predictive frequencies: dextrose administrations, ISTAT blood gas measurements (note that “ISTAT” refers to a type of handheld blood gas meter), and increases in three types of ventilator settings. We examined the 60-minute intervals of our physiological signals preceding these interventions. The intervals were taken from the first 5000 minutes of monitor data available for each of the infants in our set. As negative examples, we wanted to find intervals during which the doctor considered the infant’s state and could have ordered the intervention but did not. For this purpose, we chose 60-minute intervals preceding times when the doctor entered electronic comments about the infant but did not order the intervention in question during the hour before or after comment entry. This procedure resulted in sets of 1862, 966, 930, 806, and 146 intervals for the five interventions.

In our second set of experiments, we took the 60-minute intervals prior to measurements of blood pH, blood oxygen, blood carbon dioxide, and airway resistance that occurred during the first 5000 minutes of monitor data for the infants in our set. We had 827, 827, 771, and 189 intervals, respectively, corresponding to these four measurements.

B. Feature Extraction

Based on both visual analysis of the data and the literature on time series feature extraction, we chose five approaches to obtaining features from the physiological signals, resulting in 448 total features. We now describe each of the feature extraction approaches:

(i) We computed simple functions of the signals, such as mean, range, and variability. We believed these functions had promise as features because Saria et al. demonstrated them to be predictive of general morbidity in premature infants [9].

(ii) We calculated time-lagged correlation between each pair of signals. We believed that correlation might be informative based on our visualization of the data. Moreover, research on other complications of prematurity shows that sick infants often have impaired autoregulation [10], leading to synchronization of physiological signals.

(iii) We used the discrete Fourier transform (DFT) to obtain features capturing the frequency content of each signal. In essence, the Fourier transform allows us to characterize the variability in the signals at different time scales, and previous studies suggest that heart rate variability is a meaningful indicator of infant health [11].

(iv) We applied the Time Series Topic Model (TSTM) developed by Saria et al. [12]. The TSTM segments the physiological signal into regions (“words”) generated by the same autoregressive process, indicating similar short-term dynamics. The TSTM also learns higher-level “topics” corresponding to different distributions over words. As features, we used both word and topic frequencies obtained from the TSTM.

(v) We learned a two-layer belief network using the sparse Restricted Boltzmann Machine (RBM) algorithm proposed by Lee et al. [13]. A single example consisted of the sixty values each of heart rate, respiratory rate, and oxygen saturation that occurred during one 60-minute interval. We pre-processed the data by applying PCA whitening and then learned a sparse RBM model with 400 hidden units. With the resulting hidden unit probabilities, we trained a second sparse RBM layer with 400 hidden units. As features for a given interval, we used the inferred values of the second-layer hidden units resulting from feeding the interval’s signals forward through the trained model.

We normalized each feature to have mean zero and unit standard deviation so that the weights learned by classifiers would be meaningful in comparison to one another.

C. Predicting Interventions

For each of the five interventions, we trained a penalized logistic regression (PLR) classifier using the algorithm proposed by Zhu and Hastie, which introduces a quadratic regularization term to penalize high weights [14]. On a training set consisting of 70% of the intervals, we performed five-fold cross-validation to select the optimal regularization parameter. More specifically, for each of a range of possible parameter values and for each fold, we computed the AUC of the ROC curve for the classifier’s outputted probabilities.

We repeated this procedure three times: once with all 448 of our features, once with only the 400 RBM features (i.e., those of type (v) in Section VI-B), and once with only the 48 non-RBM features. We separated the features in this way because we were especially interested in comparing the belief networks with our other feature extraction methods, in terms of how effectively they captured the information in the physiological signals. The resulting AUCs are reported in Table II and will be discussed in Section VII.

D. Predicting Measurement Values

We first tried linear regression to predict measurement values but found that the amount of error was unacceptably high for each measurement type. We instead decided to attempt the simpler task of predicting whether the values fell above or below a given threshold. We observed that NICU doctors often mentally represent an infant’s blood gas values with ternary (“high”, “normal”, or “low”) or binary

<table>
<thead>
<tr>
<th>Measurement Type</th>
<th>Non-RBM</th>
<th>RBM</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pH</td>
<td>0.86</td>
<td>0.76</td>
<td>0.87</td>
</tr>
<tr>
<td>Blood oxygen</td>
<td>0.74</td>
<td>0.58</td>
<td>0.68</td>
</tr>
<tr>
<td>Blood carbon dioxide</td>
<td>0.69</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Airway resistance</td>
<td>0.72</td>
<td>0.56</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 3 AUCs for PLR Classifiers to predict measurements.
As thresholds separating “low” from “high” for blood pH, oxygen, and carbon dioxide, we used values indicated as clinically meaningful by Stanford physicians. For airway resistance, since such a value was not available, we took the ROC curve generated for the “maximum resistance” feature from Section IV and selected the threshold with the best sensitivity and specificity (i.e., the one leading to the point on the ROC curve closest to the upper left corner (0,1)).

We used the same procedure as described in Section VI-C to train PLR classifiers. Table III shows the resulting AUC values.

VII. Results and Discussion

We found that in each case but one, the non-RBM classifier outperformed its RBM counterpart. Furthermore, when we learned classifiers using the non-RBM and RBM features together, we obtained AUCs that were not substantially higher than those for the non-RBM classifiers. For the remainder of this discussion, then, we focus on the non-RBM classifiers.

A. Predicting Interventions

We found that, given a 60-minute interval, we were unable to accurately predict whether the interval preceded a dextrose administration, blood gas measurement, or ventilator rate increase. To better understand the problems with our models, we ran the classifiers for these three interventions on their respective training sets to compute training AUCs.

For dextrose administrations and blood gas measurements, the training AUCs were low (0.64 and 0.68, respectively), indicating that the classifiers failed to separate even the training examples well. We believe that the most likely explanation for this poor performance is that NICU doctors rely primarily on factors other than the physiological signals in deciding when to order these two interventions. Indeed, one Stanford clinician informed us that he mainly orders ISTAT blood draws after adjusting the ventilator, and thus his choice of whether to order the intervention at a given time does not depend strongly on the infant’s physiology. In terms of dextrose administrations, we observed that they generally occurred at regularly spaced intervals that varied in length for different infants, suggesting that doctors set a dextrose administration schedule in advance, rather than based on physiological signals from the past hour. Given that the frequency of dextrose administrations was predictive of BPD, it would be fruitful to investigate how NICU doctors set this administration schedule.

On the other hand, the ventilator rate increase classifier had a training AUC (0.82) much higher than its testing AUC. Thus, it appears that the classifier was overfitting, especially given the relatively small number of examples (146) for this intervention.

We were fairly successful in predicting whether an interval preceded a ventilator setting increase or a ventilator FiO₂ (fraction of inspired oxygen) increase. We note that the two classification tasks were similar, since the majority of setting increases were FiO₂ increases. One of our main goals in creating these classifiers was to identify physiological signatures useful for BPD prediction. To do so, we examined the weights that the two classifiers assigned to the features. Recall that we had normalized feature values so that the weights would be comparable.

We found that all of the highly weighted features for both classifiers were functions of the oxygen saturation (OS) signal. This result was not surprising,
given that NICU clinicians told us that they often adjust the ventilator in response to desaturation events (i.e., periods of low OS). Interestingly, the top features for both classifiers included not only simple functions of the OS signal like mean and variability, but also several features obtained from the signal’s discrete Fourier transform (DFT). This observation suggests that the frequency content of the OS signal captures useful information about desaturation events and therefore has potential for BPD prediction.

B. Predicting Measurement Values

The classifier for blood pH performed very well. The other three classifiers were also reasonably successful in differentiating high values from low. As discussed in Section V, these classifiers could be extremely useful in an NICU, giving doctors noninvasive real-time estimates of an infant’s blood gas values.

To gain insight into BPD prediction, we examined the feature weights that our classifiers learned. One general observation was that, of our nine total classifiers, eight assigned very high weight to the mean OS feature. This finding raises the question of why the mean OS feature was not among the most predictive of BPD in our initial feature exploration. We hypothesized that the answer was our segmentation of the signal into intervals, and that the distribution of mean OS values over intervals captures more useful information than the signal’s overall mean OS. Indeed, we found that the minimum of mean OS values over 60-minute intervals was more predictive of BPD than either the mean or the minimum OS value over the first 5000 minutes. This result demonstrates the benefit of focusing on short intervals of the physiological signals.

Besides mean OS, the measurement classifiers learned high weights for other simple OS features such as variability. Additionally, DFT-based features of the signals appeared frequently among the top features for all four classifiers, again suggesting that the frequency contents of the physiological signals have promise for predicting BPD.

We made two observations that were particularly interesting from a medical perspective. First, we found that the most highly weighted feature for the resistance classifier was a function of the heart rate (HR) signal’s DFT. Based on the sign of the weight, periods of high airway resistance are characterized by HR signals with more high-frequency components. This result sheds light on the effect of lung dynamics on heart rate and, more broadly, the interrelation between the respiratory and circulatory systems of premature infants. Our second interesting observation was that both the blood pH and blood oxygen classifiers learned high weight for the correlation between the HR and OS signals. Beyond suggesting a feature for BPD prediction, this link between high HR/OS alignment and low blood pH and blood oxygen levels provides insight into an unhealthy infant’s inability to autoregulate body systems.

VIII. Creating an Integrated Model for BPD Prediction

By examining the weights learned by our various classifiers, we identified sixteen features as potentially useful for BPD prediction: four simple OS features, ten DFT features, HR/OS correlation, and mean RR. With the addition of the two physiology-based features found to be predictive of BPD in our initial exploration, we had a set of eighteen features.

Rather than classifying infants as BPD or control, we decided to create a model that would classify 60-minute intervals as BPD or not. We had several reasons for making this choice. First, we saw an example in Section VII-B in which focusing on 60-minute intervals led to a more predictive feature for BPD than working with the entire 5000-minute signal at once. Second, our previous experiments were performed using 60-minute intervals, so some of our features (namely, those based on the DFT) applied specifically to intervals of that length. Finally, we believed that an interval classifier would be more valuable to NICU doctors, as it would let them obtain predictions every hour and thereby gauge changes in an infant’s BPD risk over time and with different treatments.

Ideally, we would train and test our BPD interval classifier on data taken from a new set of infants, since we already used data from our current set to create the intervention and measurement classifiers. Unfortunately, as obtaining data for new infants requires significant effort by multiple people to label and pre-process the data into usable form, we were unable to create a new data set within the time frame of the current work.

Instead, we learned a BPD classifier using the same data set from our previous experiments. We started by segmenting each infant’s first 5000 minutes of data into 60-minute intervals. We then split the set of infants into a training set (consisting of 10 positive and 10 negative examples) and a testing set. With the eighteen features described above, we trained a PLR classifier on 500 randomly chosen intervals from the infants in the training set, using 10-fold cross validation to select the optimal value of the regularization parameter. We ran the resulting classifier on all of the intervals from the infants in the testing set. To obtain an overall BPD risk score for a given infant in the testing set, we computed the proportion of that infant’s intervals that were classified as BPD. These scores achieved an AUC of 0.87. Figure III shows example outputs from the interval classifier.

IX. Conclusions and Future Work

The final product of our project was a physiology-based BPD interval classifier that achieved strong results on our data set. We would like to train and test this classifier on a new and larger set of infants to verify whether accurate BPD prediction is indeed possible with our chosen features.

In addition to developing a BPD classifier, we created nine intervention and measurement classifiers, several of which (most notably the blood pH classifier) performed well and could be very useful in an NICU. We also made a number of interesting discoveries in the course of the project. For example, we found evidence that segmenting a signal into intervals and examining them individually leads to better features for BPD prediction than treating a long signal as a whole. Furthermore, we identified informative features related to signal frequency content and HR/OS correlation, a valuable finding not just for predicting BPD but also for understanding its biological basis.

There are a number of directions in which we would like to extend this work. Given our observation that segmenting a signal into intervals can yield more predictive features, we would like to repeat our experiments using different interval lengths to determine which one is optimal. We are also interested in further exploring the relationship between cross-signal correlation and BPD by developing more sophisticated measures of correlation tailored to our specific application. In terms of learning algorithms, we would like to check whether support vector machines (another widely-used type of classification
Overall, this project represents an important step towards developing an accurate predictive model for BPD based on physiological signals alone. If the strong performance of our BPD classifier generalizes to larger data sets, the classifier could be adopted by NICUs to help doctors make more informed treatment decisions. In this way, our work has the potential to substantially improve the care of premature infants.

X. Acknowledgments
I would like to thank Suchi Saria and Professor Daphne Koller for their guidance through the course of the project, Dr. Anand Rajani and Dr. Valerie Chock for answering our questions about the NICU, and Jiquan Ngiam for letting us use and helping us understand his sparse RBM code. I would like to thank Professor Andrew Ng for teaching CS 229, the machine learning course for which I completed this project. Finally, I would like to thank Joel Jean and the SURJ reviewers for their helpful advice about making this paper more clear and thorough.

References

Laney Kuenzel is a junior from Cleveland Heights, Ohio. She is double majoring in computer science and math. Besides doing research, she enjoys tutoring, running, and working for the Stanford Flipside.