Risk Analysis, Vol. , No. , 0000 DOI: 000

Late-onset neonatal sepsis prediction using supervised learning techniques

Nadia Aly, Evan Dienstman, John Delos, Daniel McGibney, Anh Ninh

Neonatal sepsis is the leading cause of mortality in Very Low Birth Weight (VLBW) infants surviving past one week in the NICU.[1][2] If sepsis is diagnosed early then antibiotics can be administered. However, if the infection is missed, lasting neurological damage is possible and the risk of mortality increases. Currently, the methodology to diagnose sepsis is invasive and can result in complications. Consequently, a non-invasive process to predict early neonatal sepsis would have many benefits and potentially save premature infants' lives allowing timely administration of antibiotics. If the invading organism could also be predicted, antibiotics could be immediately administered, forgoing any invasive diagnostic procedure.

In this study, we derive features from the R-R intervals (distance between R peaks, Figure 1) and apply novel machine learning algorithms to predict if an infant will be diagnosed with a sepsis infection within the next twelve hours. The dataset used in this study consisted of the R-R intervals recorded by monitoring electrodes in the NICU for approximately 3,000 infants. For each sepsis diagnosis, the invading organism was also recorded; coagulase-negative Staphylococcal (CONS), gram-positive, gram-negative, fungal, and other. The encouraging results of this study imply potential clinical applications for the NICU to implement this algorithm on real-time heart rate data to influence decisions on when to proceed with diagnostic procedures. Further training on a more robust dataset is required to validate that these results are consistent and applicable to the general preterm infant population.

KEY WORDS: Neonatal Sepsis, Preterm Infant Mortality, R-R intervals, SVM

1. INTRODUCTION

Currently, the only non-invasive methodology marketed to predict neonatal sepsis is a monitoring system developed by the University of Virginia, the HeRO monitor[2]. The system displays a value calculated by an algorithm using several heart rate features termed the HeRO score[2]. In a randomized study across nine NICUs, when the HeRO score was displayed on a monitor for the clinicians, the rate of mortality was reduced by 20% compared to infants were the score was not displayed[2]. The HeRO score uses a multivariable logistic regression model to integrate the heart rate features of decreased Heart Rate Variability (HRV), and transient decelerations. These characteristics are believed to be indicative of early signs of neonatal sepsis[2][7]. In healthy infants, the heart rate is not constant, yet has many small, irregular accelerations and decelerations[7]. A deceleration is defined as a decrease in heart rate below the baseline followed by a return to base line[4]. Experiments suggest a decrease in HRV, associated with asphyxia, is a biomarker of physiological processes predictive of sepsis in neonates[3]. These characteristics are believed to be indicative of early signs of neonatal sepsis[2][7]. HRV is correlated to other vital signs including; respiration, temperature and blood pressure[3]. A few of the predictors used in our model are based off of the HeRO monitor study.

0272-4332//0100-0001$22.00/1

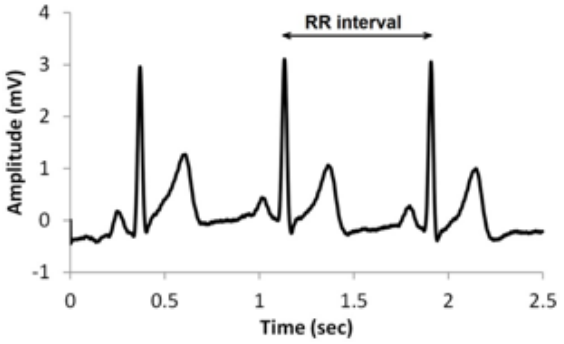

Fig. 1. Graph of Time vs. Amplitude of heart rate depicting the R-R interval, retrieved from Research Gate

In the NICU, determining if an infant will become ill in the near future is a critical question. Premature infants are at a high risk of infection due to a weak immune system. As high as 50% of the infant population in the NICU will experience a sepsis event during the duration of their stay [6]. Consequently, much work has been done to identify potential risk factors of neonatal sepsis. The risk of sepsis has demonstrated an inverse correlation with gestational age and birth weight in past studies [6]. Also independent risk factors have been identified such as ventilation support, in vitro diagnostic devices, and the administration of intravenous nutrition [6]. If data analysis and machine learning can be used to determine if an infant is at higher risk of becoming ill, then we can monitor these infants more closely and take precautionary measures to make sure the babies remain healthy. Recently, a great deal has been accomplished in this field, see [8,10,11].

In integrating the fields of physics, mathematics, and medicine; we can begin to discover the various predictors that lead up to events such as neonatal sepsis, necrotizing enterocolitis, and apnea of prematurity. Through the use of statistical measures, signal analysis, pattern recognition, and dynamic theories, in conjunction with the observations recorded by clinicians, we can determine whether certain signals will yield an early warning of these negative events for the infants. For example, one study determined the quantitative analysis of the non-invasive, electronically measured R-R intervals can provide an early warning of sepsis events [2] [7]. There still remains a number of important questions such as the cause of periodic decelerations, periodic apneas, and whether direct measures of decelerations provide a second early warning of sepsis. Our research will be building off these advances to answer a similar question.

Dunning et. al. [1] gathered data on children admitted to hospitals in northwest England with traumatic head injuries. The goal of the study was to build an algorithm to determine which children should undergo computed tomography scanning, a scan for intracranial pathology. Dunning et. al used the Cohen's Kappa Coefficient and univariate regression analysis to determine which variables were significantly associated with the dependent variable. They then employed a multivariate regression analysis with recursive partitioning to create a matrix of variables which, present in the admitted children, require the use of computed tomography scanning. [1]

In [11], Georgoulas et. al use: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), linear classifiers, and quadratic classifiers to classify fetuses that are at risk of developing metabolic acidosis based on fetal heart rate traces. They found, of these algorithms the SVM performed the best at predicting development of metabolic acidosis. [5] In our research, we will also use an SVM model while varying the kernel and margin penalty to classify which infants are at risk of sepsis.

2. THE DATASET

The dataset used in this research was from UVA's HeRO Monitor randomized clinical trial, consisting of approximately 3,000 VLBW infants collected from nine U.S. NICUs. From the 3,000 observations listed by ID number, a table of clinical events was provided that matched each infant with any sepsis event that occurred during his or her observation period. The time of diagnosis and the invading organism: CONS, gram-positive, gram-negative, fungal, and other were recorded for a total of 974 events. The dataset consisted of 49.3%, 20.8%, 21.7%, 6.7%, and 1.5% of each organism respectively as displayed in Figure 3. A few of the infants had sepsis diagnoses on more than one occasion as a result of the same type of invading organism which could have been the result of a flare up of the infection after discontinuation of antibiotics. There was an average of six weeks of data per infant. For each clinical event, we isolated the week prior to the event as the range of the observation.

Each subject had the following data recorded:

- time vector and the recorded R-R interval
- time of diagnosis

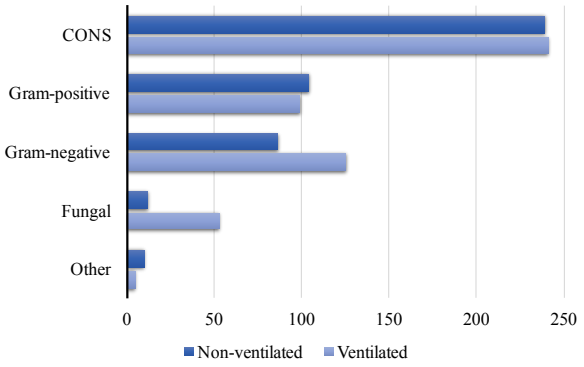

Fig. 2. Distribution of organisms in infants with documented sepsis events.

- invading organism
- ventilation status
- gestational age
- birth weight
- site ID

Many infants were missing time intervals of data as a result of being temporarily taken off the monitors or mechanical errors from the electrodes and transducers used for monitoring. These gaps and errors created outliers in the R-R measurements. We reviewed several methods to address these discrepancies. The two primary models evaluated were designed by Lake and Flower [8] [4]. Lake's model identified outliers and missing beats and removed these observations. Flower's model similarly identified outliers and missing beats, however, the model applied a smoothing technique which replaced these time intervals using a three-point running mean.

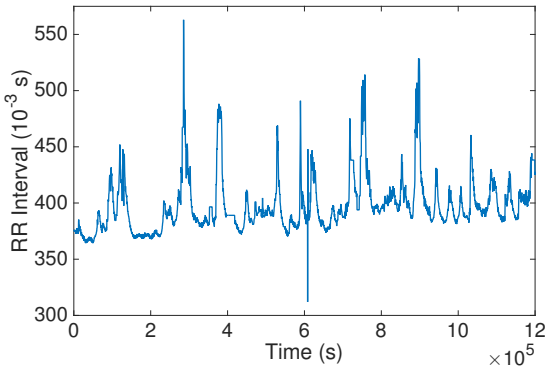

Fig. 3. Patient ID 2384: Graph of data after applying preprocessing algorithm.

3. METHODS

After the data set was preprocessed, it was divided into half-hour increments. The features derived from the R-R interval for each half-hour were: sample variance, sample entropy, sample asymmetry above the median, sample asymmetry below the median, mean, and number of decelerations. The signal features were then calculated using MATLAB programs developed from the HeRO Monitor study [2]. Python library Sklearn was used for all model simulations.

Consistent with previous literature [1], [3], [4], we expected to see signs of illness in the twelve hours leading up to a formal diagnosis. Therefore, the observations in this twelve-hour time frame were classified as sick and the four days at the beginning of the week prior to diagnosis were classified as healthy. Many observations were missing hours in the twelve hours preceding diagnosis, making a time series model impractical in use. After each feature was calculated for all observations, a panel-data problem is presented, in which there was both time-series and cross-section components. In order to address this, each half-hour interval was treated as a unique observation.

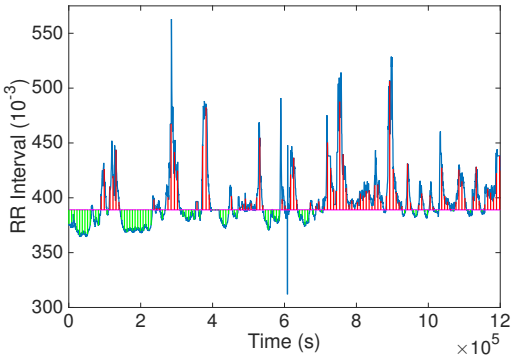

Fig. 4. Graph of R-R interval – Residual from mean shaded gray

3.1 Features: Sample Variance

The Sample Variance (see figure 4) for each half-hour interval was calculated using the following equation:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} \quad (1)$$

We then took a log-transform of the sample variance, $\ln(s^2)$.

3.2 Features: Sample Asymmetry

Next, sample asymmetry above the median and sample asymmetry below the median were calculated.

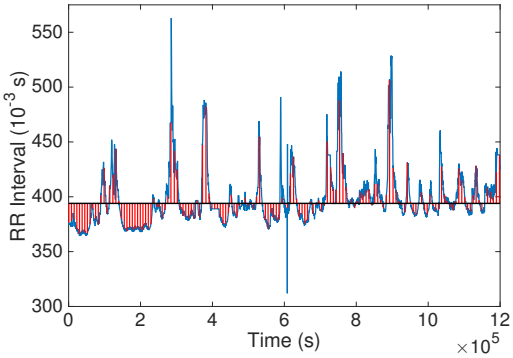

Fig. 5. Graph of R-R interval – Residual from median (above and below), shaded gray

3.3 Features: Sample Entropy

Sample entropy is a measure of sample randomness. The results of a Lake et al. study demonstrated a decreased measure of sample entropy prior to clinical diagnosis consistent with the results of Pincus et al.[8] To measure sample entropy, we selected a window size and tolerance, and then swept a pattern recognition tool developed by Lake et al.[8], through the signal identifying the reoccurrence of various identified patterns within the specified threshold. The parameters m and r were chosen as recommended by Lake et al.[8] In this study, sample entropy is derived as the negative natural logarithm of the conditional probability defined as:

- r = tolerance or threshold
- m = length of match
- CP = conditional probability
- s_b^2 = variance of the baseline process
- Δ = height of mean process with a square-topped spike
- $s_\mu^2 = \Delta^2 \epsilon(1-\epsilon)$
- SampEN = Sample Entropy

$$SampEn \approx -\log(\tfrac{r}{\sqrt{\pi}}) - \tfrac{s_\mu^2}{s_b^s} \tag{2}$$

3.4 Features: Declerations

Flower et. al[4], in "Periodic Heart Rate Decelerations in Premature Infants", investigate

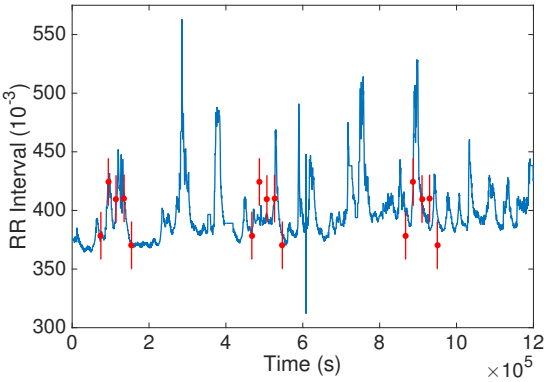

Fig. 6.

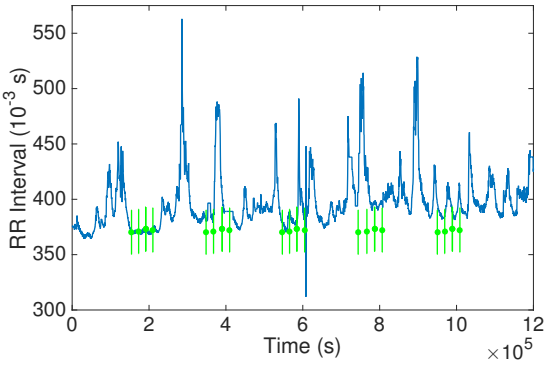

Fig. 7.

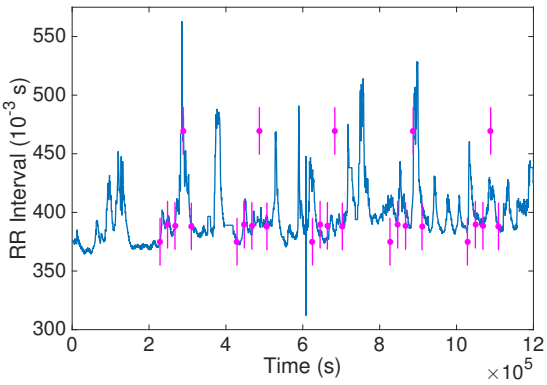

Fig. 8. Figures 6,7,8 - Demonstrates the process of sweeping a pattern through a signal to determine Sample Entropy

the decelerations in infant heart rates in a state of already minimal variability, which have been demonstrated to precede acute neonatal illness. The focus of this paper is on observation of decelerations and simulation of the same decelerations, while our research will use the frequency of observed

decelerations to make predictions regarding neonatal sepsis risk. The authors used a pattern-matching algorithm for heart rate deceleration detection. We passed our R-R interval waves through this program for each observation to identify points at which the heart rate signal matched the template within a provided threshold to locate and quantify decelerations. Flower et al[4] hypothesized infants with frequent decelerations were more likely to be diagnosed with neonatal sepsis in the next 24 hours. However, their work showed additional predictors would be necessary, because such observations also occurred periodically in the healthy population.

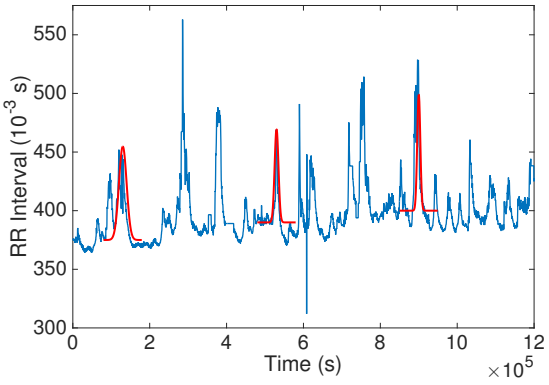

Fig. 9. Demonstrates the process of sweeping a deceleration pattern with threshold through a signal to calculate the number of decelerations

4. METHODS: SUPERVISED LEARNING TECHNIQUES

Initially, we ran a logistic regression model on the dataset to derive a baseline for improvement. To predict the binary class value of sick or healthy the following predictors were used: sample variance, sample entropy, sample asymmetry above the median, sample asymmetry below the median, mean, number of decelerations, ventilation status, gestational age, and birth weight. A validation set consisting of 20% of the data, randomly selected was set aside and the model was fit with the remaining 80%. We then ran a logistic regression with *l1* penalty (Lasso). Lasso performs variable selection in that it can shrink coefficients towards zero. Next, we ran logistic regression with *l2* penalty), and logistic regression with elastic net penalty (utilizes both *l1* and *l2 penalties*). We performed 5-fold cross validation on each model to select the tuning parameter λ.

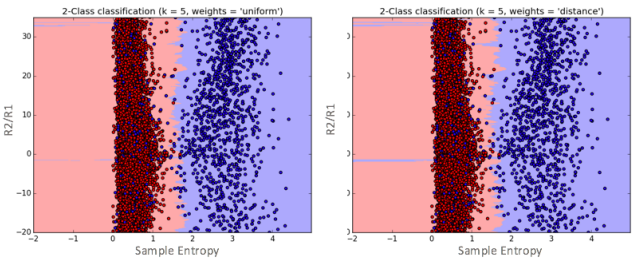

Fig. 10. Plot of KNN decision boundary; sample entropy plotted on the x-axis,R2/R1 (interaction term of asymmetry above the median/asymmetry below the median) plotted on the y-axis. Left: Uniform weights, Right: Minkowski distance. Limited to 2 feature-dimensions for purposes of interpretability of visualization. Healthy observations–red, unhealthy-blue.

We then ran a K-Nearest Neighbors (KNN) model, for the distance metric using both Euclidean and Minkowski distance formulas (see figure 10). Again, to predict healthy or sick the above predictors were used: sample variance, sample entropy, sample asymmetry above the median, sample asymmetry below the median, mean, number of decelerations, gestational age, and birth weight. However, the dummy variable: ventilation status was omitted as these distance metrics cannot be used on discrete variables. We were careful not to set the number of neighbors too low and risk over-fitting or too high and risk under-fitting. The range of neighbors evaluated was 4-15. Again, 5-fold cross validation was used, and the best performing parameter, n=6 was selected. Also, the model was run with various permutations of the above variables and variables that did not improve or decreased the overall accuracy were omitted from the final model for which the results were recorded. A secondary goal of this study was to predict the invading organism. The dataset was then partitioned into only the observations labeled sick. From here, using the multiclass KNN setting, the above process was repeated with the invading organism as the class to be predicted.

A soft-margin support vector machine (SVM) model was then used, varying the kernel and the value of the soft-margin penalty for misclassification, *c*. To use a SVM model, we initially assume the dataset is linearly separable. However, as indicated in figures 11 and 12, the dataset was not perfectly linearly separable. To address this, we introduce slack variables that allow the margin to make a few misclassifications when outliers are inside of the

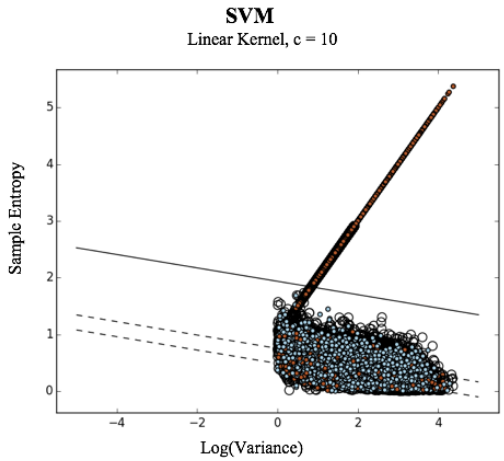

Fig. 11. Plot of SVM, linear kernal, margin cost, c = 10. Limited to 2-Dimensions for interpretability of visual.

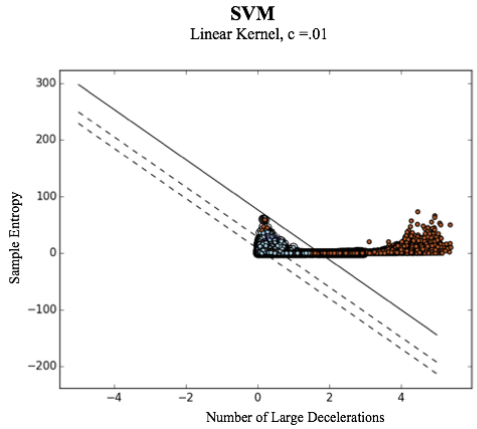

Fig. 12. Plot of SVM, linear kernal, margin cost, c = .01. Limited to 2-Dimensions for interpretability of visual.

margin or on the wrong side of the margin. The value of c determines the tradeoff between classification accuracy and an increasing margin size or the bias-variance tradeoff where increasing the value of c increases bias and lowers the model's variance. We used a validation set to evaluate the results using a linear, radial, and degree-three polynomial kernel. The data is in dimension >2, therefore the best fitting kernel cannot be visualized and most be chosen from validation accuracy. We used the same predictors defined in the logistic regression model to predict if an infant will be diagnosed with sepsis in the next twelve hours. A secondary goal of this study was to predict the invading organism. The dataset was then partitioned into only the observations labeled sick. From here, using the multiclass SVM setting, the above process was repeated with the invading organism as the class to be predicted.

We used a validation set to evaluate the results using a linear, radial, and degree-three polynomial kernel. The data is in dimension >2, therefore the best fitting kernel cannot be visualized and most be chosen from validation accuracy. We used the same predictors defined in the logistic regression model to predict if an infant will be diagnosed with sepsis in the next twelve hours. The multiclass SVM setting was also used to predict the invading organism.

5. RESULTS

In this study, a false alarm is when the model predicted sepsis when the infant did not experience a sepsis diagnosis in the next twelve hours. This is equivalent to a Type 1 error, a result that indicates a condition is present when it is not present. The false alarm rate (FP=False Positives, TP = True Positives) is calculated as,

$$\text{False Alarm Rate} = \frac{FP}{TP+FP}. \tag{3}$$

The following table illustrates the results of the logistic regression models with various regularization terms. All of these models performed within a small margin of each other. Logistic regression with $L1$ and elastic net penalty performed almost as well in terms of classification within .01% , $L1$ model's false alarm rate was 1.2% higher than elastic net's false alarm rate. The following features were used: log(variance), vent, sample entropy, sample asymmetry above and below the median, and decelerations.

Table I .

Summary of Logistic Regression Results

Model	False Alarm Rate	Classification Accuracy
Logistic (b)	54.0%	91.36%
Logistic (b), L1 penalty	55.9%	91.45%
Logistic (b), L2 penalty	54.8%	91.43%
Logistic (b), Elastic Net	54.7%	91.44%

The coefficient values selected by the logistic regression with $L1$ penalty model, for equation 5 are: $w^T x = 0.16 + .002x_{log(var)} + 0.23x_{SampEN} + 0.011x_{R1} + 0.001x_{R2} + 0.062x_{vent} + 0.0001x_{decels}$

$$P(y = 1|x) = \frac{1}{1+e^{-w^T x}} \tag{4}$$

The best performing combination of predictors

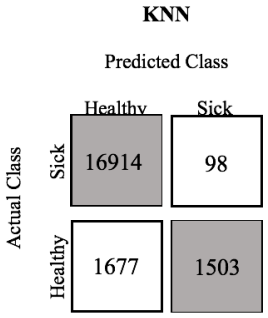

Fig. 13. KNN Confusion Matrix

for the KNN model was: log(variance), vent, sample entropy, and decelerations.

The following table illustrates the cross-validation results of the SVM model, using maximum margin parameter, c. The radial kernel performed the best in terms of classification accuracy by 0.30%, however, the false alarm rate of the linear kernel was 0.33% lower that that of the radial.

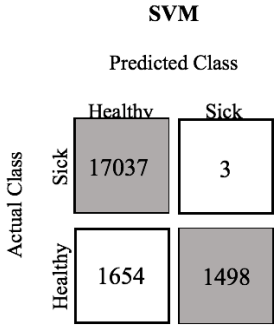

Fig. 14. SVM Confusion Matrix

False alarm rate $= \frac{3}{3+1498} * 100 = 0.07\%$ Classification Accuracy $= \frac{17037+1498}{17037+1498+3+1654} * 100 = 91.8\%$

The best performing combination of predictors for the SVM model was: log(variance), vent, sample entropy, and decelerations. In both KNN and the SVM models, adding sample asymmetry above and below the median actually increased the misclassification rate.

6. RESULTS: INVADING ORGANISM PREDICTION

The invading organism was predicted with the KNN and SVM models. The SVM models, despite attempting a large range of parameters, classified all results as Organism 1: CONS. Therefore, the results with the SVM model were not further reported in detail.

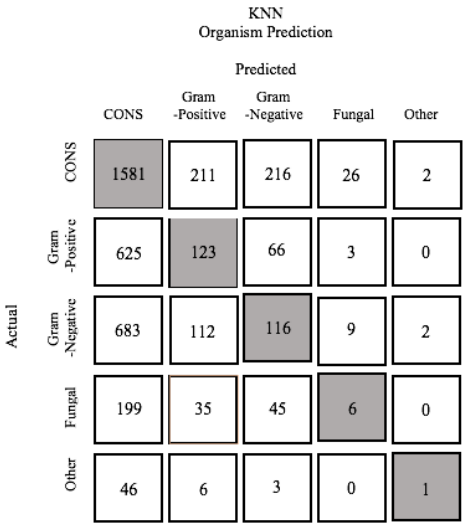

Fig. 15. KNN, Organism Prediction Confusion Matrix

KNN, n=6. Predictors: log-variance, sample asymmetry above the median, sample entropy. The classification accuracy is 44.4%. Decelerations were not yet calculated for this partition of the dataset, adding this predictor could potentially increase classification accuracy.

7. CONCLUSION

A low false alarm rate is important in this study when it comes to planning to apply the results in a clinical setting. If the false alarm rate is low enough, then diagnostic tests can be immediately performed without concern of undergoing the procedure when in fact the infant is not ill. Considering false alarm rate, the best performing result was the SVM with a linear kernel. The logistic models had similar classification accuracy, however, their false alarm rate was high (averaging $\approx$ 55%), making the model impractical for potential NICU monitoring applications.

It is interesting to note that the logistic, $L1$ penalty model shrunk both features: decelerations and sample asymmetry towards zero and the ventilation status and sample entropy explained the most variation in the response variable. We hypothesized that the decelerations would have been a more significant predictor in this model. Also, the value

Table II .

Summary of SVM Model Results

SVM Model Kernel	Margin Parameter; c	False Alarm Rate	Classification Accuracy
Linear	10	0.07%	91.70%
Radial	10	0.40%	91.80%
Polynomial-degree 3	1	0.07%	91.60%

of sample entropy was expected to have a negative correlation with the probability of sepsis. A lower sample entropy should represent a decrease in heart rate variability indicative of illness.

The SVM model, with a linear kernel, outperformed all other models with a 0.07% false alarm rate and a 91.7% classification accuracy. These encouraging results imply potential clinical applications for the NICU to implement this algorithm on real-time heart rate data to influence decisions on when to proceed with diagnostic procedures. If more clinicians record neonatal heart rate data and an increased number of NICU's contribute to the dataset to train the model in the future, it would allow for a more robust classifier.

The secondary goal of this study was to predict the invading organism so that an antibiotic could be immediately administered. The SVM model failed to produce any interpretable results. The most accurate results, were with a KNN model, n=6, with an overall classification accuracy of 44.4%. This low classification accuracy could be a result of the unequal distribution of the training dataset provided of the 974 diagnosed event, 49.3% the invading organism were CONS. To improve this accuracy in future studies, more observations of the remaining four classes are necessary to produce a more robust classifier prior to attempting to make any generalizations.

ACKNOWLEDGMENT

Ask Delos to write

REFERENCES

1. J Dunning, J Patrick Daly, JP Lomas, F Lecky, J Batchelor, and K Mackway-Jones. Derivation of the children's head injury algorithm for the prediction of important clinical events decision rule for head injury in children. *Archives of disease in childhood*, 91(11):885–891, 2006.
2. Karen D Fairchild and Judy L Aschner. Hero monitoring to reduce mortality in nicu patients. *Research and Reports in Neonatology*, 2:65–76, 2012.
3. Karen D Fairchild and T Michael O'Shea. Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis. *Clinics in perinatology*, 37(3):581–598, 2010.
4. Abigail A Flower, J Randall Moorman, Douglas E Lake, and John B Delos. Periodic heart rate decelerations in premature infants. *Experimental Biology and Medicine*, 235(4):531–538, 2010.
5. George Georgoulas, Chrysostomos D Stylios, and Petros P Groumpos. Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING BME*, 53(5):875, 2006.
6. James W Gray. Surveillance of infection in neonatal intensive care units. *Early human development*, 83(3):157–163, 2007.
7. J.F. Hicks and Karen D Fairchild. Hero monitoring in the nicu: sepsis detection and beyond.
8. Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.