Classification of Cardiotocography Signals using Machine Learning

Sumedh Anand Sontakke
Department of Electrical Engineering
College of Engineering, Pune
sumedh.sontakke2@gmail.com

Reshul Dani
Department of Computer Science
University of California,
San Diego

Jay Lohokare
Department of Computer Science
SUNY Stony Brook

Pranav Shivagaje
Department of Electrical Engineering
College of Engineering, Pune

***Abstract*—Cardiotocography has been used to record and monitor fetal heartbeat and uterine contractions, both antepartum and intrapartum for several decades now, albeit not without considerable controversy. The International Federation of Obstetrics and Gynecology (FIGO) guidelines were the first set of universally accepted classification guidelines for CTG signals. During labor, changes in the CTG are useful as indicators of fetal conditions. This paper aims to utilize CTG signal parameters to classify fetuses into three fetal states: normal, suspect and pathological and into 10 morphological patterns.**

Keywords—Artificial neural networks; machine learning; bioinformatics

I. INTRODUCTION

Cardiotocography is performed prenatally during the third trimester of pregnancy and is a crucial step in determining the overall health of the fetus and the probable time of delivery [2]. Statistically, between one and seven in a thousand fetuses experience acute oxygen deprivation, primarily acidosis, severe enough to cause permanent brain damage and death [3]-[5], [20]. The clinical significance of most of the fetal heart rate signals are well understood. A value of about 140 beats per minute with fluctuations of 5-15 beats per minute is indicative of an adequate blood delivery and a responsive nervous system providing healthy modulation. Temporary decreases in fetal heart rate (less than 15 bpm for less than 15 minutes) are indicative of strain on heart muscle or compression of the umbilical cord whilst short term accelerations are typical of a healthy fetus [6]. While the clinical implications of CTG signals are well understood, inter and intra subject variability can be a significant caveat.

CTG signal are interpreted visually and are used to draw clinical inferences; however, their application has been rather inconsistent, subjective and prone to the obstetrician's discretion resulting in a significant false positive rate [7].

In this paper, we endeavor to produce a robust machine learning solution to the fetal classification problem, impervious to overfitting and with high generalizability. Unlike previous work, we endeavor to produce a machine learning system for prediction of both the fetal state and morphological state labels.

II. RELATED WORK

Cardiotocography analysis is a relatively old method to ascertain prenatal well-being. However, a computationally focused approach in its interpretation has only been explored at the turn of the millennium. Chen et al. [8] developed a LabView based system for analysis of the fetal heart rate (FHR) and uterine contractions (UC). The system showed promise with accuracy for FHR baseline at 100%, albeit the algorithm was tested on a rather small sample of 19 women.

Bernardes et al. [9]-[11] produced some of the most influential work in automated cardiotocography based analysis of fetal health based on the guidelines prescribed by the International Federation of Obstetrics and Gynecology (FIGO). Magenes et al. [12], [13] employed an artificial neural network based system for classification of fetal conditions. Chung et al developed an algorithm to predict the onset of acidosis [14]. Georgoulas et al. [15] used a support vector machine based system to predict the onset of metabolic acidosis. Salamalekis et al. used features extracted from the FHR and pulse oximetry to predict fetal hypoxia [16]. Alonso-Betanzos et al. created a computer aided fetal evaluator which integrated machine learning with traditional fetal health methodology [17]-[19].

III. DATA SET DESCRIPTION

The data set used for our research was obtained from the online Machine Learning Repository maintained by the University of California, Irvine [15]. The data was obtained from the Cardiotocography dataset made publicly available by Dr Bernardes at the University of Porto, Portugal. The given dataset included 2126 instances of fetal cardiotocographic parameters. The set contained no missing features.

The target labels for each of the data points were:

1) Morphological pattern: The data classified into 10 patterns namely calm sleep, REM sleep, calm and active vigilance, a constantly shifting pattern (in calm sleep),

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

accelerative or decelerative pattern due to both stress or vagal stimulation, a largely decelerative pattern, a flat sinusoidal pattern (pathological state) or a suspect pattern.

2) FIGO labels: The data were also labeled in strict accordance with the FIGO guidelines as normal, suspect, or pathological, postpartum and were assumed to be ground truths.

The distribution of the pathological labels is illustrated in Fig. 1 while the distribution of the morphological states is illustrated in Fig. 2. Table I describes the variables in the data set.

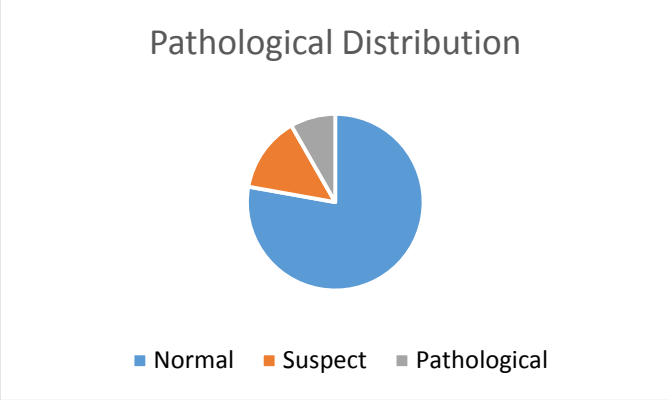

Fig. 1. Pathological distribution of the data: The data contained 3 kinds of fetal states – normal, suspect, and pathological.

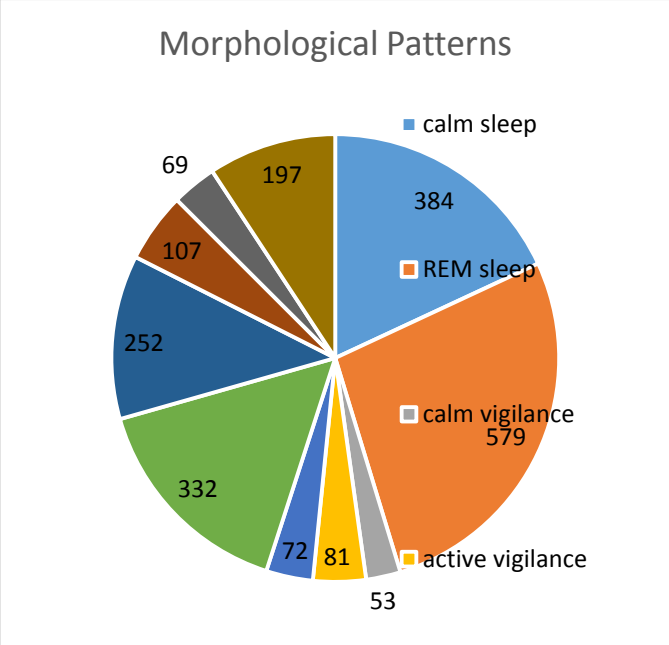

Fig. 2. Morphological pattern distribution in the data: The data contained 10 morphological patterns with the REM sleep containing the most in number (579). The flat sinusoidal pattern in the pathological state contained the least (53).

TABLE I. VARIABLE DESCRIPTION. THE DATA USED CONTAINED 21 VARIABLES AND 2 CLASS TARGET LABELS. THE FEATURES USED WERE ALL CONTINUOUS EXCEPT FOR THE HISTOGRAM TENDENCY WHICH WAS CATEGORICAL. THE CLASS VARIABLES WERE CODED AS CATEGORICAL DURING ANALYSIS. LBE WAS OBTAINED FROM A MEDICAL EXPERT WHILST THE REMAINING VARIABLES WERE OBTAINED FROM SISPORTO COMPUTERIZED ANALYSIS SYSTEM BUILT BY BERNARDES ET AL. [10]

LBE	baseline value (medical expert)
LB	baseline value
AC	accelerations
FM	fetal movement
UC	uterine contractions
ASTV	%age of time with abnormal short-term variability
mSTV	mean value of short term variability
ALTV	%age of time with abnormal long-term variability
mLTV	mean value of long term variability
DL	light decelerations
DS	severe decelerations
DP	prolonged decelerations
DR	repetitive decelerations
Width	histogram width
Min	low frequency of the histogram
Max	high freq. of the histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency

IV. METHODOLOGY

A. Preprocessing

The database collected needed to be pre-processed to remove possible errors and improve the quality of the machine learning models built from it. The first step in the pre-processing pipeline was data visualization. For each of the categorical variables, a 5-number summary (minimum, first quartile, second quartile, third quartile, maximum, and range) along with the standard deviation was obtained. The categorical variables were coded as numerical levels. Boxplots and histograms for continuous variables were plotted using the ggplot2 package in R. Univariate regression models were built variable for each of the continuous variables. Chi-squared tests were carried out for each of the categorical variables. Missing value imputation was not necessary as the data was complete. The data was subsequently split into training and test as illustrated in Fig. 3.

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

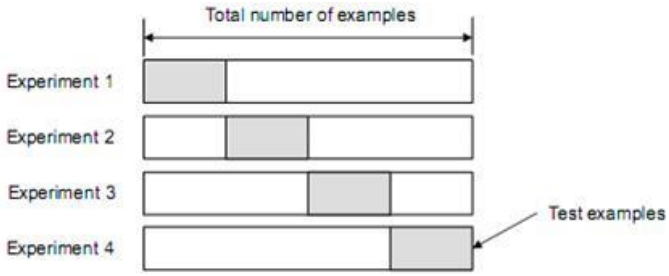

Fig. 3. Data split for training and test: The entire data consisting of 2126 instances was split into 4 parts. Three of the four parts were used as training data and the final portion was held out until the very end where it was used as test data.

After the data was cleaned, it was split 75:25 into training and test data respectively. Following the split, the training data was also subject to a near-zero-variance test. This resulted in the removal of variables with near zero variance in the data as they were not responsible for any predictive power in the models built subsequently, but only contributed in increasing the dimensionality of the data.

The training data was subject to a 10-fold Cross validation repeated 10 times (Fig. 4).

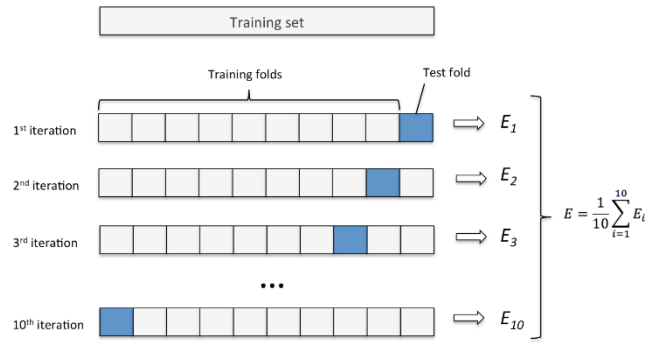

Fig. 4. Repeated K-Fold Cross Validation. A repeated 10-fold CV was applied. The 10-fold CV works by dividing the training data into 10 equal parts. These parts are iterated through 10 times. During each iteration, 9 of the 10 parts are treated as training data, and the remaining 10^{th} part as the validation set. The performance metrics are measured after each iteration. At the end, accuracy and kappa values are computed as measures of model performance. The above procedure is repeated 10 times.

B. *Spot Checking*

The data set was then subject to a series of spot checks. Several machine learning algorithms were applied to the data, with the goal of trying to find the highest intrinsic performance, i.e. the algorithm that generated the highest accuracy of prediction before its tuning parameters were manually adjusted.

This was achieved using the caret package in R. The package performed a grid search to find the optimum hyperparameters for each of the algorithms. The spot check was carried out using the 10-fold Cross validation repeated 10 times. The repeated CV works by splitting the training data into 10 folds, training the model on 9 of the folds and computing performance metrics on the hold-out fold (10^{th} fold). This process is subsequently repeated 9 times. The repeated cross validation scheme results in 10 instances of the performance metric allowing for the computation of a 95% CI (Fig. 5).

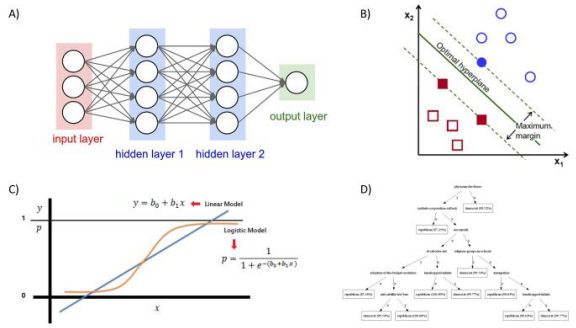

Fig. 5. Overview of the working of the various classification algorithms applied during spot-checking. A. *Neural Networks*. It consists of 'neurons' which are essentially non-linear activations functions which 'fire' when a condition is satisfied interspersed between the 'layers' of the model. These layers are high dimensional matrices which are learnt by finding the global minimum of the error function. Once trained, the neural network is used to make predictions about unknown vectors. B. *Support Vector Machines*. This algorithm separates the data points of each class using a high dimensional hyperplane called optimum hyperplane by maximizing the distance between the points of the class (color coded as red and blue). C. *Logistical Regression Classifier*. A logistical regression works by finding the label of a vector using an exponential function. The coefficients of the exponential function are learnt by minimizing the error function. D. *Tree Based methods*. Tree based methods work by dividing the training data at nodes based on values of the predictors. Several such trees are built. When a new data point is to be assigned a label, individual trees vote on the label and the label with the highest frequency is assigned to the data point.

The performance metrics for comparison were accuracy and kappa values. Accuracy is the proportion of correctly classified fetal states. However, this can be a misleading metric if the distribution of the output variable is unbalanced, i.e. if one of the classes contains far more instances in the training data than the other. Thus, the kappa value was included which denotes the square root of the proportion of variance in the output variable correctly explained by the model.

C. *Hyperparameter Tuning*

The last step involved tuning the hyperparameters of the best performing model identified from the previous step. The main aim of this procedure is to balance the bias and variance of the model. The model should perform with a high level of classification accuracy while still maintaining high generalizability to data points previously unseen by it. We also endeavored to ensure that the model is not susceptible to the nuances of the data and has not overfit to it.

Maintaining the balance between bias and variance was done by manually tuning the complete set of hyperparameters of the model using nested cross validation (CV). Nested CV is a two-step validation procedure. It consists of 2 loops used for validation. The outer loop divides the data into m folds. The inner loop then performs k fold cross validation on each of the m folds with different hyperparameters. The result consists of surrogate models trained using different hyperparameters allowing us to identify the highest performing

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

hyperparameters. This method allows one to reduce bias error by making the model as flexible as possible while simultaneously monitoring the variance error via the performance on the validation set. This main aim is to locate the minimum of the sum of the bias and the variance error. The process is illustrated in Fig. 6.

Thus, with the model parameters finalized, the model was then used for prediction of the regulatory success of unknown drugs in the test set.

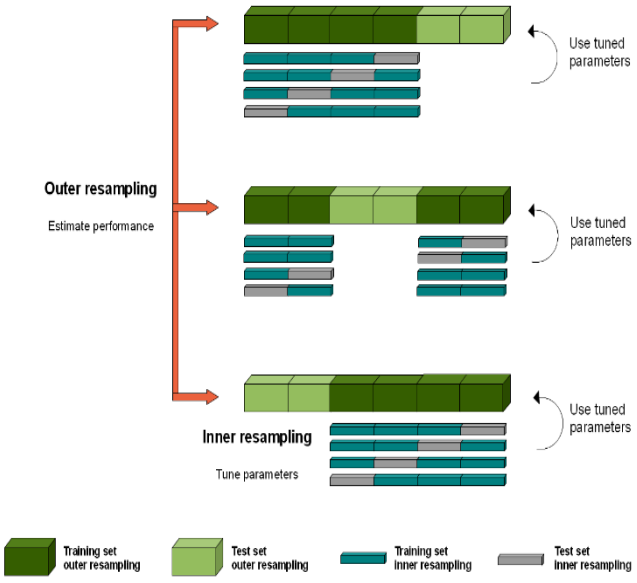

Fig. 6. Nested cross validation: The nested CV consists of 2 loops; the inner loop performs cross validation on the training data that is selected by the outer loop in 1 iteration. Subsequently, the parameters learnt are tested on the holdout set of that iteration. This process then repeats itself on different training data selected by the next iteration of the outer loop.

V. Results

A. Spot Checking Results

The spot checking was carried out on the training data using the 10-fold repeated Cross validation. This process yielded 10 metrics of comparison accuracy. To better understand algorithmic performance, the 95% confidence intervals were calculated for each (Table II).

TABLE II. 95% Confidence Intervals for Algorithmic Accuracy on the Morphological State Label. The Table Displays the Results of the 10-Fold Repeated Cross Validation. The 95% Confidence Intervals Assume a Normal Distribution of Accuracy

Sr. No	Model Name	Lower Bound	Upper bound
1	Random Forest	0.852	0.884
2	GLM	0.805	0.849
3	K Nearest neighbors	0.73	0.771
4	Artificial Neural Nets	0.777	0.823
5	Support Vector Machines	0.824	0.858

TABLE III. 95% Confidence Intervals for Algorithmic Accuracy on the FIGO Fetal Labels. The Table Displays the Results of the 10-Fold Repeated Cross Validation. The 95% Confidence Intervals Assume a Normal Distribution of Accuracy

Sr. No	Model Name	Lower Bound	Upper bound
1	Random Forest	0.9242053	0.9454714
2	GLM	0.8803536	0.9028288
3	K Nearest neighbors	0.8804376	0.906448
4	Artificial Neural Nets	0.8888815	0.9269203
5	Support Vector Machines	0.8971566	0.9261896

The Cross-validation accuracy is indicative of the algorithmic performance on test data since in each iteration, one-fold is left out as a hold-out set. Repeated CV is a good method to estimate test error. However, as rightly pointed out by Tibshirani and Hastie [22], the training data is already preprocessed. This induces some bias into the model. The results of the repeated CV are thus interpreted with caution. The results of the test data shown in Table III are presented in Fig. 8.

B. Nested Cross Validation and Test Results

Nested Cross-Validation was performed to obtain the highest algorithmic accuracy whilst ensuring that the model is generalizable.

Tables IV and V present algorithmic accuracy post hyperparameter tuning. In addition to accuracy, kappa values are also represented. Kappa values are an essential tool to understand algorithmic performance as they signify the proportion of the variance in the output target variable explained by the model. Although varying schools of thought express interpretations of the kappa values differently, a value above 0.8 is accepted to be indicative of high accuracy. The results of the test data shown in Table IV are presented in Fig. 7.

TABLE IV. Accuracy and Kappa for Algorithmic Accuracy on the Morphological State Label. The Table Displays the Results of the Test Data. A Kappa Value of Greater than 0.75 is Acceptable

Sr. No	Model Name	Accuracy	Kappa Value
1	Random Forest	0.868	0.842
2	GLM	0.827	0.793
3	K Nearest neighbors	0.751	0.700
4	Artificial Neural Nets	0.800	0.759
5	Support Vector Machines	0.841	0.810

TABLE V. Accuracy and Kappa for Algorithmic Accuracy on the Fetal Pathological State Label. The Table Displays the Results of the Test Data. A Kappa Value of Greater than 0.75 is Acceptable

Sr. No	Model Name	Accuracy	Kappa Value
1	Random Forest	0.934	0.817
2	GLM	0.891	0.700
3	K Nearest neighbors	0.893	0.690
4	Artificial Neural Nets	0.907	0.747
5	Support Vector Machines	0.911	0.810

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

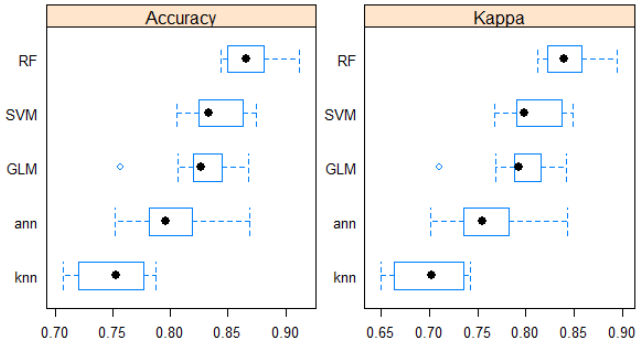

Fig. 7. Box plots for Accuracy and Kappa for Algorithmic Accuracy on the Morphological State Label. The figure displays the results of the test data in Table IV.

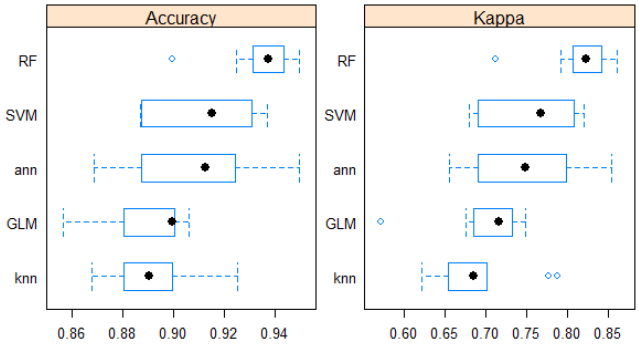

Fig. 8. Box plots for Accuracy and Kappa for Algorithmic Accuracy on the Fetal State Label. The figure displays the results of the test data in Table III.

In both cases, the random forest fits well to the data. The variable importance levels for the random forest algorithm are illustrated in Fig. 9 and 10.

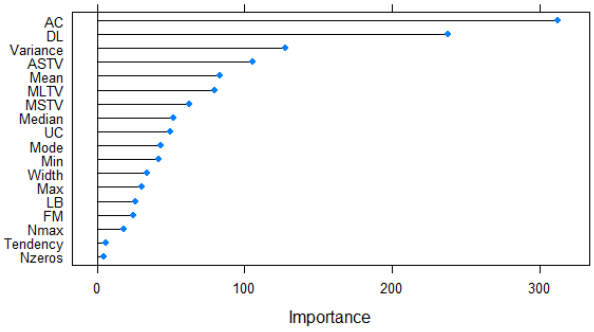

Fig. 9. Variable importance plots for the Random Forest on the Morphological State Label. The table displays the degrees of importance for each of the variables in predicting morphological states.

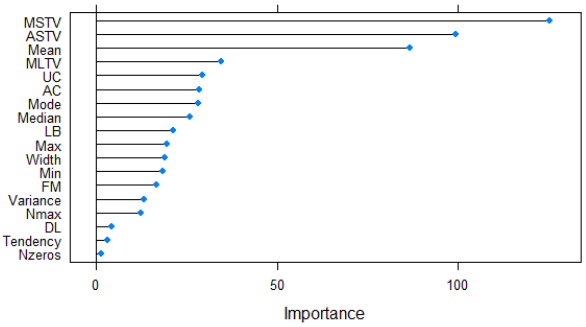

Fig. 10. Variable importance plots for the Random Forest on the Fetal State Label. The table displays the degrees of importance for each of the variables in predicting the fetal pathological states.

VI. Discussion

Interpretation of cardiotocography has been an incessantly debated topic, and a litany of malpractice litigations and controversies mar its short history [1]. Inter and intra observer variability has been the biggest barrier in the effective interpretation of cardiotocographic data. The FIGO (International Federation of Obstetrics and Gynecology) in the 1980s provided the first set of internationally accepted guidelines and consequently several regional variations of these guidelines were adopted, around the world. However, these guidelines have proven to be very complex and difficult to follow [23] resulting in the high variability of interpretation.

To lower the inter-intra-observer variability in the interpretation of fetal data, a computationally focused approach is necessary. Previous approaches to classification of fetal cardiotocographic data may have been hampered by small sample sizes. Small data sets and high dimensionality are patent problems in computationally focused approaches to fetal classification and often result in overfitting of the models to data [8]. This results in low out-of-sample performance and a lack of generalizability of the model.

This study endeavored to overcome these caveats. The data used in the model was labeled in strict adherence to FIGO guidelines. This automatically eliminated the conundrum of inter-intra-observer variability. The sample size for the study was adequate for deep learning methods as well (N = 2126), a rarity in bioinformatics. The data was pre-processed to remove possible confounding variables. The machine learning models were also trained using several robust training techniques like repeated cross validation. This ensured that the model built was impervious to overfitting.

However, the most significant advantage of the system developed through this study is its ability to out-perform human obstetricians in identifying pathological fetuses. The accuracy of diagnosis of pathological fetuses by human

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

obstetricians ranges between 50-66% [24], [25]. This system outperforms human obstetricians significantly with an excess of 25 percentage points in accuracy.

The next challenge in cardiotocographic based prediction systems lies in real-time diagnosis. Using continuous time recurrent neural networks and other time series based classifiers may be the next incremental innovation in predicting fetal health.

VII. Conclusion

Although the figure in the UK still stands at 8 perinatal deaths in 1000, the global figure has been on the decline from 4.5 million perinatal deaths annually in 1990 to about 2.6 million in 2013. Infant respiratory distress syndrome is the leading cause of these deaths and affects 1% of all infants [21]. And thus, it is imperative that advanced diagnostic procedures be used to prevent the permanent and often fatal effects of respiratory distress. In this paper we attempted to address this problem through a computationally focused approach.

The paper explores solutions to two classification problems based on the cardiotocography data obtained from 2126 fetuses. The data was preprocessed, and fit to several high performing machine learning algorithms with high generalizability as indicated by a low estimated test error. The best of these models was then fine-tuned to obtain the final models.

References

[1] Sartwelle TP. Electronic fetal monitoring: a bridge too far. J Leg Med 2012;33(3):313–79.

[2] Ayres de Campos et al. (2000) SisPorto 2.0, A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318 "

[3] ACOG, Neonatal Encephalopathy and Cerebral Palsy: Defining the Pathogenesis and Pathophysiology. ACOG Task force on Neonatal Encephalopathy and Cerebral Palsy, Jan. 2003.Kun-Hong Liu and De-Shuang Huang. "Cancer classification using Rotation forest"

[4] N. Badawi, J. Kurinczuk, J. Keogh, L. Alessandri, F. O'Sullivan, P. Burton, P. Pemberton, and F. Stanley, "Antepartum risk factors for newborn encephalopathy: The Western Australian case-control study," Brit. Med. J., vol. 317, pp. 1549–1553, 1998.

[5] E. Draper, J. Kurinczuk, C. Lamming, M. Clarke, D. James, and D. Field, "A confidential enquiry into cases of neonatal encephalopathy," Arch. Dis. Child. Fetal Neonatal Ed., vol. 87, pp. F176–F180, 2002.V.N.

[6] Philip A. Warrick∗, Emily F. Hamilton, Doina Precup, and Robert E. Kearney, "Classification of Normal and Hypoxic Fetuses From Systems Modeling of Intrapartum Cardiotocography," IEEE Transacations on Biomedical Engineering, Vol. 57, No. 4, April 2010

[7] J. T. Parer, T. King, S. Flanders, M. Fox, and S. J. Kilpatrick, "Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association?" J. Matern.-Fetal Neonatal Med., vol. 19, no. 5, pp. 289–294, May 2006.

[8] Chen CY, Chen JC, Yu C, Lin CW.2009. A comparative study of a new Cardiotocography analysis program. Conf Proc IEEE Eng Med Biol Soc. :2567-70.

[9] J. Bernardes, C. Moura, J. P. M. de Sa, and L. Pereira-Leite, "The Porto system for automated cardiotocographic signal analysis," J. Perinat. Med., vol. 19, pp. 61–65, 1991

[10] J. Bernardes, C. Moura, J. P. M. de Sa, L. Pereira-Leite, and H. P. van Geijn, "The Porto system," in A Critical Appraisal of Fetal Surveillance, H. P. van Geijn and F. J. A. Copray, Eds. New York: Elsevier Science, 1994, pp. 315–324

[11] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. P. M. de Sa, and L. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," J. Matern.. Fetal Med., vol. 9, pp. 311–318, 2000.

[12] G. Magenes, M. G. Signorini, and D. Arduini, "Classification of cardiotocographic records by neural networks Neural Networks," in Proc. IEEE-INNS-ENNS Int. Joint Conf. on Neural Networks (IJCNN'00), 2000, vol. 3, pp. 637–641.

[13] G. Magenes, M. G. Signorini, and D. Arduini, "Multiparametric analysis of fetal heart rate: comparison of neural and statistical methods," in Proc. Medicon 2001, pp. 360–363.

[14] T. K. H. Chung, M. P. Mohajer, X. J. Yang, A. M. Z. Chang, and D. S. Sahota, "The prediction of fetal acidosis at birth by computerized analysis of intrapartum cardiotocography," Br. J. Obstet. Gynaecol., vol. 102, pp. 454–460, Jun. 1995.

[15] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[16] E. Salamalekis, P. Thomopoulos, D. Giannaris, I. Salloum, G. Vasios, A. Prentza, and D. Koutsouris, "Computerised intrapartum diagnosis of fetal hypoxia based on fetal heart rate monitoring and fetal pulse oximetry recordings utilising wavelet analysis and neural networks," Br. J. Obstet. Gynaecol., vol. 109, no. 10, pp. 1137–1142, Oct. 2002.

[17] A. Alonso-Betanzos, V. Moret-Bonillo, L. D. Devoe, J. R. Searle, B. Banias, and E. Ramos, "Computerized antenatal assessment: the NSTEXPERT project," Automedica, vol. 14, pp. 3–22, 1992.

[18] A. Alonso-Betanzos, B. Guijarro-Berdinas, V. Moret-Bonillo, and S. Lopez-Gonzalez, "The NST-EXPERT project: the need to evolve," Artif. Intell. Med., vol. 7, no. 4, pp. 297–313, 1995.

[19] B. Guijarro-Berdinas, A. Alonso-Betanzos, and O. Fontenla-Romero, "Intelligent analysis and pattern recognition in cardiotocographic signals using a tightly coupled hybrid system," Artif. Intell., vol. 136, pp. 1–27, 2002.

[20] GBD 2013 Mortality and Causes of Death, Collaborators (17 December 2014). "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013". *Lancet*. 385 (9963): 117–71. doi:10.1016/S0140-6736(14)61682-2. PMC 4340604. PMID 25530442

[21] Rodriguez RJ, Martin RJ, and Fanaroff, AA. *Respiratory distress syndrome and its management.* Fanaroff and Martin (eds.) Neonatal-perinatal medicine: Diseases of the fetus and infant; 7th ed. (2002):1001-1011. St. Louis: Mosby.

[22] T Hastie, R Tibshirani, J Friedman. Elements of Statistical Learning, Springer Series in Statistics, 2001

[23] de Campos DA, Bernardes J. Twenty-five years after the FIGO guidelines for the use of fetal monitoring: time for a simplified approach? Int J Gynecol Obstet 2010;110(1):1–6.

[24] Intra- and inter-observer variability in the assessment of intrapartum cardiotocograms. P. V. Nielsen, B. Stigsby, C. Nickelsen & J. Nim. Acta Obstetricia et Gynecologica Scandinavica Vol 66 1987.

[25] Intrapartum cardiotocography – the dilemma of interpretational variation. Outi Palomäki, Tiina Luukkaala, Riikka Luoto, Risto Tuimala. Journal of Perinatal Medicine, Volume 34, Issue 4 2006.

978-1-5386-4185-9/18/$31.00 ©2018 IEEE