Siddharth Chaphekar

DSC 540 – Advanced Machine Learning

Autumn 2019

Application of Machine Learning Methods in Fetal Abnormality Detection Using Cardiotocography Data

Abstract

The growing focus in the medical community on early detection of disease has led to significant improvements in patient outcomes, both in terms of health and overall associated costs. In the case of fetal abnormalities, studies have shown that early discoveries of conditions such as fetal hypoxia, which can be inferred from an abnormal heart rate, have led to early interventions that improved long-term outcomes for both mothers and children. Cardiotocography is an electronic monitoring system that facilitates these discoveries by recording data on fetal heart rate and uterine contractions. A 2018 study by Tang et al. devised a neural network implementation that eliminates the need for manual feature acquisition. However, the lack of explainability or "black box" nature of neural networks, particularly in the context of deep learning, presents a challenging outlook from a compliance perspective. Thus, the goal of this paper is to explore several, more simple approaches of classification in an attempt to replicate the diagnoses of domain experts. The results of the paper show that while there are indeed simpler approaches that are viable to solve the problem, the comparisons in performance are complicated by a variety of factors and thus more extensive research is needed in this domain.

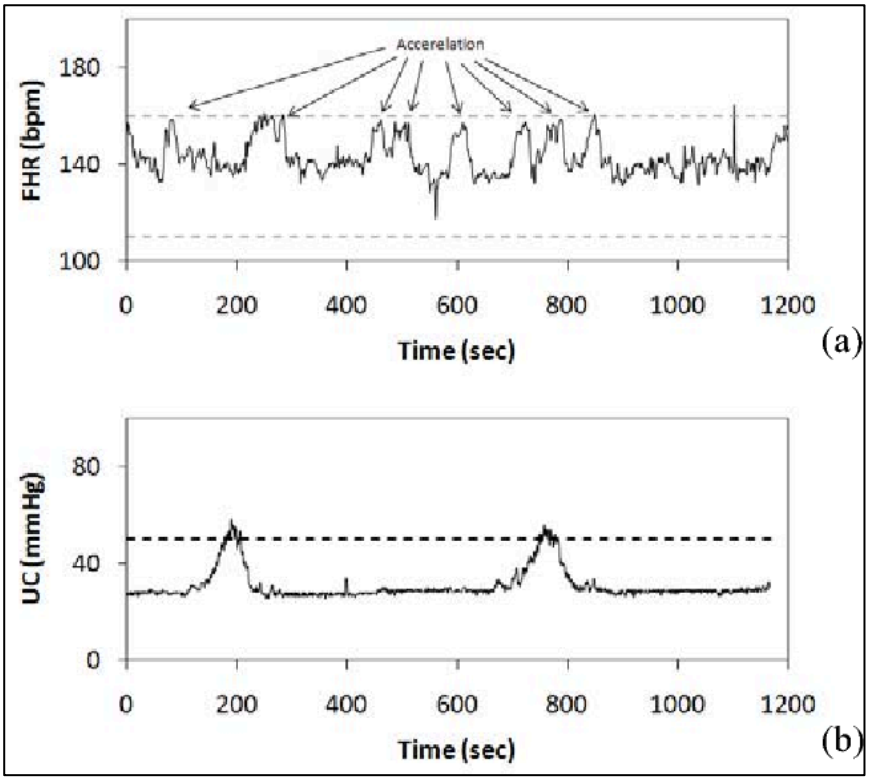
Introduction

From the advent of personalized medicine to the rise of digital health, the healthcare industry has experienced remarkable transformations across several domains in recent times. In particular the application of new technologies for health monitoring has expanded the domain of diagnostics, which now not only aims to detect disease, but also detect precursors to disease. Expanding this "window" of early diagnosis has extremely promising implications in terms of health outcomes and overall cost of care. For instance, a 2017 study published by Elder et al. showed that an earlier diagnosis of autism can facilitate better interventions in the natural surroundings of the child, leading to improved long-term outcomes.

But the benefits of early diagnosis are not confined to certain age demographics. Abnormalities such as hypoxia in fetuses can also be diagnosed to facilitate potential interventions. In such cases, cardiotocograms are used to detect irregularities either before or during labor. Cardiotocography is an electronic monitoring system that is used to measure fetal characteristics such as heart rate and uterine contractions. A physician reads and interprets the cardiotocogram signal to screen for potential abnormalities. Based on the analysis, the physician may then recommend an intervention procedure, such as a Caesarian section or assisted delivery.

However, there are several incentives to automate this process. One reason is that an intelligent system could facilitate remote, around-the-clock maternity care, which would allow for monitoring without the need for clinical visits. From a resource utilization perspective this would be beneficial for hospitals because it would release time for other patients that would otherwise be devoted to maternal care. The problem of observer variability would be addressed as well since the elements of varying judgment and human error would be removed by an objective computing method. This paper will attempt to replicate the diagnoses of a panel of three obstetricians by exploring various machine learning approaches for classification and then comparing and evaluating the performance of the models for this very purpose.

Figure 1: A Cardiotocogram signal showing FHR and uterine contractions



Literature Review

Fetal Abnormalities

A 2011 study by Afors et al. described the various types of fetal abnormalities observed in high-risk pregnancies and the CTG characteristics used to identify them. In the case of fetal hypoxia, decelerations followed by loss of accelerations, subsequent rise in baseline heart rate and gradual loss of variability is a typical pattern. However, the study also points out that a Cochrane review found no evidence to support the use of antepartum CTG for improving perinatal outcomes, with the warning that most of these studies lacked data to compare antenatal CTG testing on fetus' less than 37 weeks compared to fetus' of 37 or more completed weeks. Timing of detection is also very important given that fetal state evolves, so detecting fetal distress near the time of delivery has less potential to improve clinical outcomes compared to an advance warning of say, 1.5 hours.

Cardiotocography and Fetal Well-Being

A 2012 study by Rahman et al. examined the predictive value of the admission cardiotocogram (CTG) in detecting fetal hypoxia at the time of admission in labor as well as the perinatal outcome in high-risk obstetric cases. The results of the study, which looked at FHR monitoring data and outcomes, showed that CTG is an excellent, simple, and non-invasive test that can serve as a screening tool in high-risk obstetric patients. However, other studies such as Warrick et al. (2010) have suggested that continuous monitoring may induce physical and psychological distress in the mothers, thus creating false positives and forming issues where none may previously exist.

Neural Network Implementation

A 2018 study by Yang et al. proposed a classification method of the FHR signal based on neural networks. The researchers suggest that for some patients, the physiological parameters detected in the hospital vs. those detected in a familiar environment may be different, owing to certain physical and psychological factors. This is significant because it furthers the case for automation; however, there are several issues with this study. First, the deep learning architecture proposed severely limits the explainability of the model, which from a healthcare compliance perspective could prove unviable. Second, the researchers claim to have automated the feature acquisition step that removes the "guesswork" of physician interpretation, but the overall methodology is poorly explained. Furthermore the dataset that the researchers used has still not been released, so external validation of these results cannot be conducted.

Random Forest in Medical Applications

A 2019 study by Alam et al. showed that by using feature ranking and only using highly ranked features in a Random Forest implementation, highly accurate predictions could be made. 10 different diseases were chosen for the study, and feature ranking and selection strategies led to successful implementations. For instance, the best models for the breast cancer dataset achieved F1 scores of above 95%. The researchers thus not only conducted successful runs on benchmark datasets, but also presented a general methodology that should perform well for other diseases that show similar characteristic patterns.

Data

The data for this paper was obtained from the UCI Machine Learning Repository, and was originally contributed by researchers from the University of Porto and the Biomedical Engineering Institute of Porto in Portugal. The raw dataset contains 2126 automatically processed instances of fetal cardiotocograms, each with 23 diagnostic measurements. A panel of three physicians classified each instance with respect to morphologic pattern and fetal state. For the purposes of this paper, the latter was used as a schema to facilitate binary classification. This schema does not present a problem in terms of potential outcomes because regardless of whether the classification is "suspect" or "pathologic" the patient would still have to consult a physician in terms of intervention. So a good classifier would minimize the number of false negatives and maximize the number of true positives.

Table 1: Feature Set Description

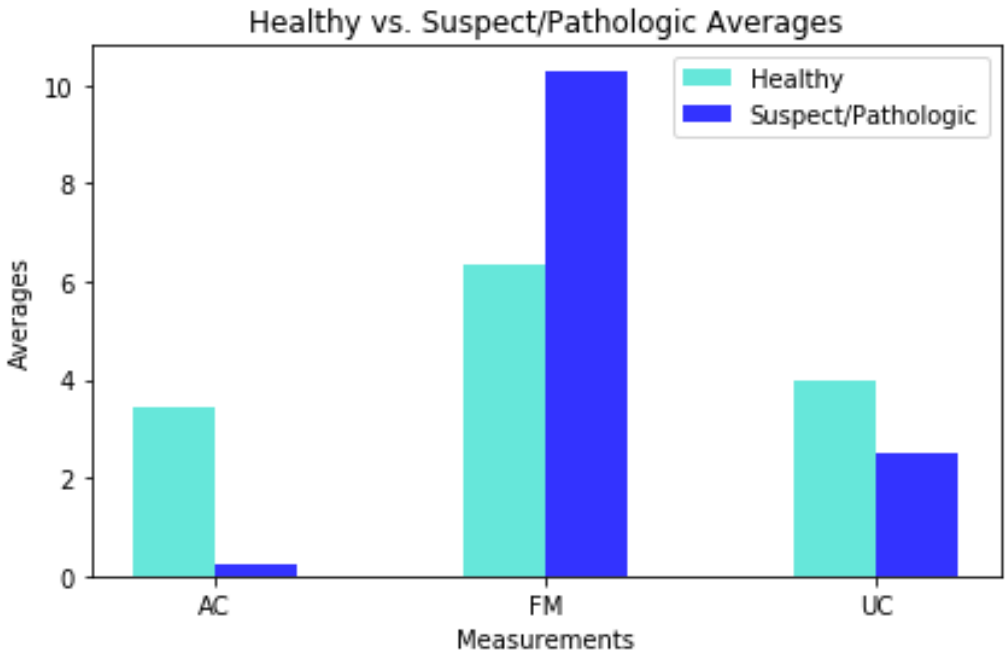
Measurements	**Description**
LB	Baseline fetal heart rate, measured in beats per minute (bpm)
AC, FM, UC	Accelerations, fetal movements, and uterine contractions (per second)
ASTV, mSTV, ALTV, mLTV	Percentage of time with abnormal short term/long-term variability and mean values of those variability
DL, DS, DP, DR	Light, severe, prolonged, and repetitive decelerations
Width, Min, Max, NMax, NZeros	Histogram width, min/max frequency, # of peaks, and # of zeros

Mode, Mean, Median, Variance	Histogram statistics
Tendency	Histogram tendency; -1 = left asymmetric, 0 = symmetric, and 1 = right asymmetric

The 2015 FIGO Intrapartum Fetal Monitoring Guidelines have outlined some of the general characteristics that help classify cases as normal, suspect, or pathologic. Instances with baseline heart rates between 110 and 160 BPM, variability between 5 and 25 BPM, and no repetitive decelerations are considered normal, while instances that lack at least one of these is considered suspect. For pathological cases, baseline heart rate is generally <100 BPM with decelerations greater than 5 minutes in length, and there are repetitive late or prolonged decelerations.

Initial analysis of the dataset confirmed the data quality in conjunction with the FIGO guidelines. For instance, the average number of fetal heart rate accelerations in the subset of the data containing only healthy samples was significantly higher than those of the suspect/pathologic class. However, the baseline heart was not much different (131 BPM for normal vs. 137 BPM for suspect), so it is evident that a collection of factors beyond the basic FIGO guidelines is necessary for classification.

Figure 2: Double Bar Chart: Accelerations, fetal movements, and uterine contractions



Class Imbalance

In the 2126 total instances, 1655 samples were labeled as "healthy" while 295 and 176 were labeled "suspect" and "pathologic" respectively. Even after the suspect and pathologic classes were combined to facilitate binary classification, the number of normal samples still dominated the dataset, accounting for nearly 80% of the observations. This problem was addressed using the oversampling module from Python's imblearn library, which equalized the ratio of the two classes. When the models were re-built and implemented, it was easier to measure performance in terms of false negatives because the classification accuracy metric was no longer misleading.

Feature Selection

The raw data contained two columns that represented redundant attributes, namely LBE and LB. The measurements were taken from different sources (one from the SisPorto measuring device, and one from the expert) but because the values were identical for all instances and would only contribute to multicollinearity, LBE was dropped. The start and end times were also dropped from the final feature set as they were simply time markers and were unnecessary for classification.

The feature set was made up of primarily five components – acceleration and fetal movements counts, FHR variability metrics, decelerations, and histogram statistics – and model performance suffered whenever one of these components was dropped. So in terms of model performance, the rest of the feature set was retained to avoid information loss.

Results

In total, five different algorithms were tested and evaluated. Initial results with Decision Trees, Random Forest, and Support Vector Machine were very encouraging, but owing to the heavy class imbalance these results needed further evaluation. Synthetic Minority Oversampling Technique (SMOTE) was applied to resolve the class imbalance.

Table 2 – Performance of K-Neighbors Implementation

Data	Accuracy	Precision	Recall	F1
Unbalanced	0.9231	0.8878	0.7196	0.7949
SMOTE	0.9122	0.7345	0.9015	0.8095

Table 3 – Performance of Naïve Bayes Implementation

Data	Accuracy	Precision	Recall	F1
Unbalanced	0.8589	0.7326	0.5401	0.6218
SMOTE	0.8793	0.7142	0.7299	0.7220

Table 4 – Performance of Decision Tree Implementation

Data	Accuracy	Precision	Recall	F1
Unbalanced	0.9874	0.9923	0.9485	0.9699
SMOTE	0.9733	0.9473	0.9264	0.9368

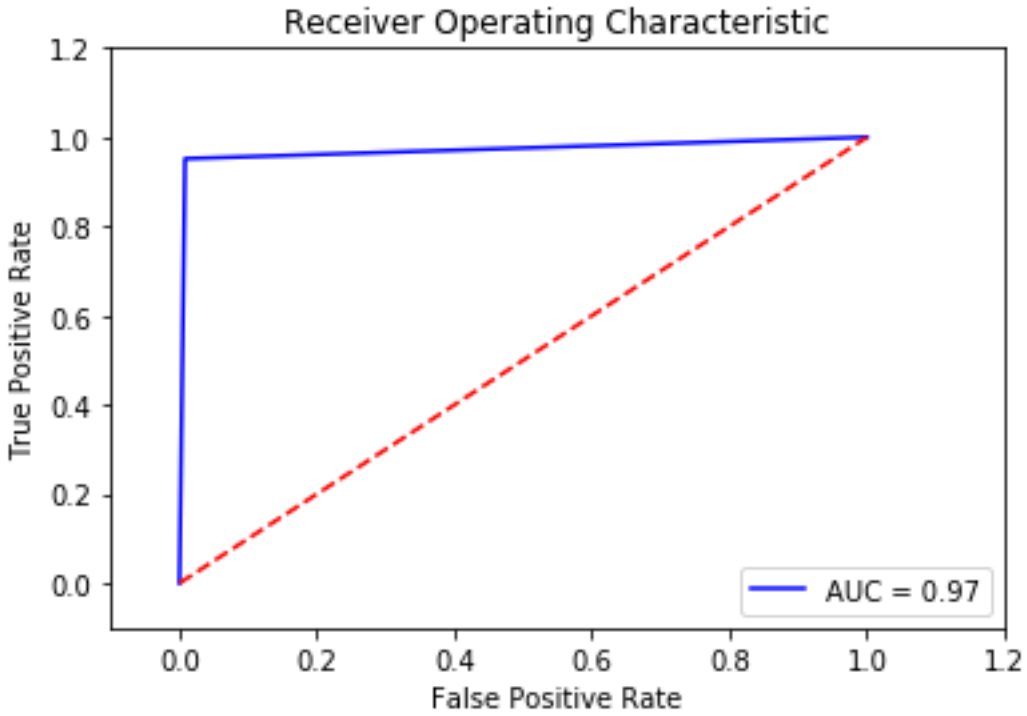
Table 5 – Performance of Support Vector Machine Implementation

Data	Accuracy	Precision	Recall	F1
Unbalanced	0.9890	1.0	0.9469	0.9727
SMOTE	0.9921	1.0	0.9621	0.9810

Table 6 – Performance of Random Forest Implementation

Data	Accuracy	Precision	Recall	F1
Unbalanced	0.9811	0.9784	0.9379	0.9577
SMOTE	0.9827	0.9718	0.9517	0.9616

Figure 3: ROC Curve of SVM Implementation



Discussion

For the targeted classification schema of healthy or suspect/pathologic (0 or 1), the Random Forest and SVM approaches proposed in this paper correctly identified the overwhelming number of samples, with high performance scores across the board and good predictive power in terms of false negatives. This shows that a simpler mathematical approach like SVM or an ensemble method like Random Forest is sufficient to replicate the predictions in the UCI dataset while also proving that a deep learning implementation is unnecessary here. The AUC of 0.97 of the SVM implementation shows that the diagnostic ability of this binary classifier system is excellent.

On the other hand, regardless of class imbalance, K-Neighbors and Naïve Bayes did not perform well. The Naïve Bayes performance is particularly surprising, given that the data appeared to be a good candidate for probability-based modeling. Both N-Neighbors and Naïve Bayes failed to reach 95% accuracy and the corresponding confusion matrices showed several misclassifications in terms of false positives. Though the 95% benchmark is somewhat arbitrary, it is reasonable given that the domain in question concerns potential life-or-death scenarios.

The Decision Tree classifier performed far better than expected, reaching accuracies of over 97% for both balanced and unbalanced versions of the dataset. SVM and Random Forest performed even better, with excellent precision and recall across the board. This indicates that the algorithms achieved high proportions of correctly identified positive classifications as well as actual positives identified correctly.

Conclusion

The overall performance of the models in terms of interpretability is very promising. Specifically, Random Forest and SVM appear to perform the best, with high accuracy metrics across the board and clear differences in performance when key attributes are removed. However, there are some concerns that are not within the scope of this paper and will need to addressed with further research. One of these concerns is the efficacy of CTG diagnosis. In terms of the "big picture" application of this technology, it would be premature to claim that such systems are ready for deployment. The marriage between the clinical research and the machine learning application potential can only succeed if there is overwhelming consensus that CTG data is indeed a trustworthy tool to improve outcomes. But as of today there are still questions surrounding the necessity, such as the triggering of unnecessary interventions.

Another challenge that should be addressed is that of data consolidation. At present, the issues of how, when, and where data is collected does not seem to be standardized. The UCI dataset was a good example of a certain method of collection, but since the neural networks study data was never released there is little to no means of comparison. The Warrick study presents a similar problem, but it is exacerbated by the fact that their dataset only has 264 total instances. There is also no way to definitively say that outcomes could be improved by these modeling techniques, since there is no bridge between good classification and favorable outcome. In other words, physician diagnoses were replicated successfully, but whether or not the patients involved actually saw a return on investment is still unclear.

In conclusion, machine learning methods will likely play an important role in influencing fetal health outcomes in the future, but the path to viability will demand more research and likely involve more personalized approaches to care delivery.

References

Afors, Karolina, and Edwin Chandraharan. "Use of Continuous Electronic Fetal Monitoring in a Preterm Fetus: Clinical Dilemmas and Recommendations for Practice." *Journal of Pregnancy*, Hindawi, 13 Sept. 2011, www.hindawi.com/journals/jp/2011/848794/

Elder, Jennifer Harrison, et al. "Clinical Impact of Early Diagnosis of Autism on the Prognosis and Parent-Child Relationships." *Psychology Research and Behavior Management*, Dove Medical Press, 24 Aug. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5576710/

Galazios, Georgios, et al. "Fetal Distress Evaluation Using and Analyzing the Variables of Antepartum Computerized Cardiotocography." *Archives of Gynecology and Obstetrics*, U.S. National Library of Medicine, Feb. 2010, www.ncbi.nlm.nih.gov/pubmed/19455348.

Rahman, Hafizur, et al. "Admission Cardiotocography: Its Role in Predicting Fetal Outcome in High-Risk Obstetric Patients." *The Australasian Medical Journal*, Australasian Medical Journal, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3494822/.

Saifur, Rahman, M., et al. "A Random Forest Based Predictor for Medical Data Classification Using Feature Ranking." *Informatics in Medicine Unlocked*, Elsevier, 13 Apr. 2019, www.sciencedirect.com/science/article/pii/S235291481930019X.

Tang, et al. "The Design and Implementation of Cardiotocography Signals Classification Algorithm Based on Neural Network." *Computational and Mathematical Methods in Medicine*, Hindawi, 3 Dec. 2018, www.hindawi.com/journals/cmmm/2018/8568617/.

Warrick, Philip A, et. al *A Machine Learning Approach to the Detection of Fetal Hypoxia during Labor and Delivery*. Association for the Advancement of Artificial Intelligence, 2010, https://www.researchgate.net/publication/45365428_Classification_of_Normal_and_Hypoxic_Fetuses_From_Systems_Modeling_of_Intrapartum_Cardiotocography